Universidad Nacional de Misiones. Facultad de Ciencias Exactas, Químicas y Naturales. Secretaría de Investigación y Postgrado. Doctorado en Ciencias Aplicadas

Doctoranda
Lic. Nancy Beatriz Ganz

Detección de factores de fracaso en implantes dentales mediante la aplicación de Ciencia de datos

Tesis de Doctorado presentada para obtener el título de "Doctor en Ciencias Aplicadas"

"Este documento es resultado del financiamiento otorgado por el Estado Nacional, por lo tanto queda sujeto al complimiento de la Ley N° 26.899".

Director
Dr. Horacio Daniel Kuna
Co-Directora
Dra. Alicia Esther Ares

Posadas, Misiones 2021

Universidad Nacional de Misiones
Facultad de Ciencias Exactas, Químicas y Naturales

DETECCIÓN DE FACTORES DE FRACASO EN IMPLANTES DENTALES MEDIANTE LA APLICACIÓN DE CIENCIA DE DATOS

Por Lic. Nancy B. GANZ

Tesis presentada a la Facultad de Ciencias Exactas, Químicas y Naturales de la Universidad Nacional de Misiones para optar al grado académico de

DOCTOR EN CIENCIAS APLICADAS

Posadas, República Argentina

2021

Director
Dr. Horacio D. KUNA

Co-director
Dra. Alicia E. ARES

TRIBUNAL EXAMINADOR (Resolución Consejo Directivo Nº 375/20)

Dr. José Ángel OLIVAS VARELA	Universidad de Castilla La Mancha
Dr. Hernán MERLINO	Universidad de Buenos Aires
Dra. Marina QUIROGA	Universidad Nacional de Misiones

DEFENSA ORAL Y PÚBLICA (Resolución Consejo Directivo Nº 056/21)

Posadas, 5 de Marzo de 2021

DETECCIÓN DE FACTORES DE FRACASO EN IMPLANTES DENTALES MEDIANTE LA APLICACIÓN DE CIENCIA DE DATOS

Nancy B. GANZ

Lugar de desarrollo del trabajo de tesis
INSTITUTO DE MATERIALES DE MISIONES

COMISIÓN DE SUPERVISIÓN (Resolución Consejo Directivo Nº 256/16 y su modificación Nº 404/17)

Dr. Hernán MERLINO	Universidad de Buenos Aires
Dra. Alicia MON	Universidad Nacional de La Matanza
Dr. Oscar ALBANI	Universidad Nacional de Misiones

CARRERA DE DOCTORADO EN CIENCIAS APLICADAS

Proyecto de Carrera Nº 10933/11
Con reconocimiento de la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU) Nº 344/11.

AGRADECIMIENTOS

Al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), por la beca otorgada para realizar este Doctorado.

A la Universidad Nacional de Misiones, especialmente a la Facultad de Ciencias Exactas Químicas y Naturales y al Instituto de Materiales de Misiones por posibilitar mi educación superior de posgrado.

Al Colegio de Odontólogos de la Provincia de Misiones por el asesoramiento y respaldo para realizar esta investigación.

A mis directores de Tesis, el Dr. Horacio D. Kuna y la Dra. Alicia E. Ares, por el apoyo, la confianza, el acompañamiento y la comprensión que me brindaron durante estos años.

A todos los Odontólogos, por su gran colaboración y buena predisposición al permitirme recolocar datos de sus historias clínicas. En especial a los Odont. Diego H. Padula y Gabriel A. Aimone Romero, por su enorme ayuda y orientación.

A mi compañero y amigo Facundo Domínguez, por su colaboración en este trabajo.

A mis amigos José Luis Montejano, Miryan Puchini y Daniel Oneddú, por ser incondicional y estar siempre predispuestos a ayudarme.

A mis compañeros del ProMyF, por las orientaciones, charlas y mates compartidos.

A mi familia, pilar fundamental, ya que sin su apoyo y amor incondicional hubiese sido imposible la realización de este trabajo.

A Dios, por cuidarme, darme la salud y por bendecirme con esta gran oportunidad y con las personas que compartieron conmigo este camino.

Nancy B. Ganz

RESUMEN

El gran volumen de datos existente en el sector de la salud dificulta la toma de decisiones por parte de los especialistas, debido a que no se aplican técnicas que aprovechen al máximo la información disponible, ocasionando la dificultad de reconocer patrones de comportamiento y extraer conocimiento oculto de los datos almacenados. Además, la no predicción del comportamiento, basado en el conocimiento previo, puede acarrear un alto porcentaje de inexactitud, más aún cuando se trata de un campo tan primordial como el de la salud.

Hoy en día, la predicción del éxito o fracaso de un implante dental está determinado a través de una evaluación clínica y radiológica. Por esta razón, las predicciones dependen en gran medida de la experiencia del implantólogo. Es por esto, que es de vital importancia contar con un registro de datos de las condiciones del paciente, las características del implante e información del proceso de colocación, debido a que la rehabilitación oral a través de implantes dentales puede presentar riesgos relacionados con la etapa del proceso de oseointegración (implante / tejido). Estos riesgos pueden estar relacionados con las condiciones de salud del paciente, la técnica quirúrgica empleada por el especialista implantólogo, el tipo de implante, así como el tabaquismo, entre muchas otras.

De aquí, surge la necesidad de aplicar técnicas de Ciencia de Datos, debido a que son capaces de extraer patrones, de predecir comportamientos, regularidades y, de sacar provecho a la información automatizada. En esta tesis se estudió el beneficio de la utilización de múltiples técnicas de Ciencia de Datos, para la predicción de factores de fracasos a partir de un conjunto de datos de historias clínicas sobre implantes dentales. Especialmente, se buscó identificar factores que contribuyan al fracaso de estos implantes y determinar las condiciones óptimas que debe tener el paciente y el implante dental utilizado por el profesional implantólogo.

Este trabajo de tesis permitió lograr la creación de un registro novedoso de historias clínicas de pacientes que se han sometido a procesos quirúrgicos de colocación de implantes dentales en la Provincia de Misiones, Argentina. Además, permitió proponer un modelo de aprendizaje automático para identificar y descartar las características redundantes e irrelevantes del conjunto de datos de estudio. Asimismo, se detectaron los factores que ejercen una mayor influencia en el proceso de osteointegración a través de un procedimiento de clasificación y validación por expertos humanos.

Se concluyó que el procedimiento propuesto permitió conocer las características más relevantes y mejoró la precisión en la clasificación de la clase objetivo (fracaso del implante dental), permitiendo no sesgar la toma de decisión en base a la aplicación y resultados de método individuales.

Palabras claves: Ciencia de datos, aprendizaje automático, clasificación, implantes dentales, predicción de fracasos.

ABSTRACT

The large volume of data in the health sector makes it difficult for specialists to make decisions, due to the fact that techniques that make the most of the available information are not applied, causing the difficulty of recognizing behavior patterns and extracting hidden knowledge of the stored data. In addition, the non-prediction of behavior, based on previous knowledge, can lead to a high percentage of inaccuracy, especially when it comes to such a fundamental field as health.

Nowadays, the prediction of the success or failure of a dental implant is determined through a clinical and radiological evaluation. For this reason, predictions are highly dependent on the experience of the implantologist. This is why it is of vital importance to have a data record of the patient's conditions, the features of the implant and information on the placement process, because oral rehabilitation through dental implants can present risks related to the stage of the osseointegration process (implant / tissue). These risks may be related to the patient's health conditions, the surgical technique used by the implantologist, the type of implant, as well as smoking, among many others.

From here, the need arises to apply Data Science techniques, because they are capable of extracting patterns, predicting behaviors, regularities and, taking advantage of automated information. In this thesis, the benefit of the use of multiple Data Science techniques was studied, for the prediction of failure factors from a set of data from dental implant medical records. In particular, we sought to identify factors that contribute to the failure of these implants and to determine the optimal conditions that the patient and the dental implant used by the implantologist should have.

This thesis work allowed the creation of a novel registry of medical records of patients who have undergone surgical procedures for the placement of dental implant in the Province of Misiones, Argentina. In addition, it allowed proposing an machine learning model to identify and rule out redundant and irrelevant features of the study data set. Likewise, the factors that have the greatest influence on the osseointegration process were detected through a classification and validation procedure by human experts.

It was concluded that the proposed procedure allowed to know the most relevant features and improved the accuracy in the classification of the target class (dental implant failure). Allowing, not to bias the decision making based on the application and individual methods results.

Keywords: Data science, machine learning, classification, dental implants, failure prediction.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

En Argentina, como en otros países del mundo se fabrican biomateriales para diferentes aplicaciones con el objetivo de restaurar las funciones del cuerpo humano, como es el caso de los implantes dentales. Sin embargo, es necesario contar con información suficientemente calificada y accesible sobre: tipos de implantes dentales que se fabrican en el país y cuáles se importan, propiedades de los biomateriales, causas de fallo, características y condiciones de salud de los pacientes que requieren de estos implantes, entre otros. Para esto, es necesario relevar información de diferentes fuentes, como revistas científicas, bases de datos, Internet, empresas nacionales e internacionales, hospitales, sanatorios, médicos, odontólogos, implantólogos, investigadores en biomateriales, pacientes, ministerios de salud, etc.

La carencia de un registro digital provincial o nacional de implantes dentales, que contenga datos sobre enfermedades sistémicas, condiciones del paciente a la hora de la intervención, características del implante utilizado, datos del procedimiento de la fase quirúrgica y datos del seguimiento postoperatorio, hace dificultosa la tarea de análisis y extracción de conocimiento desconocido u oculto sobre patrones que podrían llegar a influir en el proceso de oseointegración de un implante. Un registro digital con estas características, podría aportar conocimiento valioso para los especialistas implantólogos sobre la relación implante dental / condición del paciente.

De aquí, surge la necesidad de crear un registro automatizado que reúna las variables que representan el proceso de colocación de un implante dental, con el objetivo de identificar los factores que contribuyen al fracaso de los implantes dentales colocados en la Provincia de Misiones, Argentina a través de la aplicación de técnicas de Ciencia de Datos, y el diseño de un procedimiento con métodos híbridos.

1.1 PLANTEO DEL PROBLEMA

El campo de la Ciencia de Datos [1] ha tenido muchos avances respecto a la aplicación y el desarrollo de técnicas en el sector de la salud. Estos avances se ven reflejados en la predicción de enfermedades, clasificación de imágenes, identificación y reducción de riesgos, soporte a la toma de decisiones en base al análisis de conjuntos de datos, entre muchos otros.

Existen trabajos que aplicaron métodos para la predicción del éxito de los implantes dentales, como Tamez *et al.* [2] muestran un análisis estadístico para determinar los factores que influyen en el éxito de los implantes dentales colocados en el Posgrado de Prostodoncia e Implantología de la Universidad De La Salle Bajío, México. Domínguez *et al.* [3] realizan un estudio donde determinan si existe relación entre los fracasos de los implantes dentales y las enfermedades sistémicas (concretamente sobre la osteoporosis, hipertensión, diabetes e hipotiroidismo), en una población de pacientes sometidos a cirugía de implantes dentales en el hospital San José de Santiago, Chile. En el trabajo de Oliveira *et al.* [4] presentan un

análisis comparativo de tres técnicas de aprendizaje automático: Máquina de Vector Soporte (SVM) [5], SVM ponderado y una red neuronal [6] con parámetro de selección, para la predicción del éxito de los implantes dentales. Las características consideradas en dicho trabajo fueron: edad del paciente, sexo, tipo de implante, posición del implante, técnica quirúrgica, indicación de que si el paciente era fumador o no, he indicación de que si el paciente tenía una enfermedad previa (diabetes u osteoporosis) o un tratamiento médico (radioterapia). El conjunto de datos que utilizan consta de 157 casos, registrados por un solo cirujano de la Faculdade de Odontologia, Universidade Federal do Rio Grande do Sul, Brasil. Otro trabajo de iguales características es el de Moayeri *et al.* [7], donde presentan un modelo predictivo combinado para evaluar el éxito de los implantes dentales. Los clasificadores utilizados en este modelo son un árbol de decisión (J48) [8], un SVM, una red neuronal, un k vecino más cercano (KNN) [9] y una red bayesiana (Naive Bayes) [10]. Para evaluar la eficacia de los algoritmos propuestos, utilizan 224 casos de pacientes que tenían colocados implantes dentales. Este conjunto de datos pertenece a la Escuela de Odontología de la Universidad de Teherán, Irán y contiene diferentes variables, las cuales son: sexo, edad, fumar, ubicación del implante, tiempo de colocación, protocolo de carga, diámetro y longitud del implante, tipo de conexión del implante, sobre-dentadura y elevación de senos maxilares. En el trabajo de Braga *et al.* [11] proponen un conjunto de modelos logísticos binarios para evaluar la probabilidad de éxito o no éxito en el proceso de rehabilitación oral, teniendo en cuenta algunos factores genéticos, hábitos individuales y factores clínicos y no clínicos. El estudio se realizó en una evaluación retrospectiva y consistió en 155 sujetos sometidos a rehabilitación oral en la región norte de Portugal.

Si bien estos trabajos utilizan conjuntos de datos de implantes dentales, no centran su atención al biomaterial (tratamiento de superficie del implante), es por esto que este trabajo aborda conjuntamente el estudio de las características del implante propiamente dicho, ligado a los rasgos y condiciones de salud de los pacientes, así como a las condiciones del proceso quirúrgico.

Por este motivo, el trabajo estuvo organizado en tres grandes etapas:

1. Creación de un registro con datos reales de historias clínicas de pacientes que se han sometido a procesos quirúrgicos de colocación de implantes dentales en la Provincia de Misiones, Argentina.
2. Propuesta de un procedimiento para la selección de las características más importantes y para la clasificación a través de la aplicación de técnicas de Ciencia de Datos, mediante el diseño de un procedimiento con métodos híbridos, y la aplicación de la metodología CRISP-DM para asentar el proceso.
3. Validación del procedimiento con expertos humanos.

1.2 SOLUCIÓN PROPUESTA

Crear un registro automatizado que reúna las variables que representan el proceso de colocación de un implante dental, y diseñar un procedimiento con métodos híbridos de Ciencia de Datos para identificar los factores que contribuyen al fracaso de los implantes dentales colocados en la provincia de Misiones, Argentina.

1.3 OBJETIVOS

1.3.1 OBJETIVO GENERAL

El objetivo general de este trabajo fue desarrollar y adaptar un procedimiento de Ciencia de Datos a través de una metodología hibrida, para predecir fracasos en implantes dentales de la provincia de Misiones, Argentina.

1.3.2 OBJETIVOS ESPECÍFICOS

Los objetivos específicos del presente trabajo son:

1. Relevar, analizar y evaluar los antecedentes relacionados con la aplicación de la Ciencia de Datos en el análisis de datos médicos.
2. Analizar las principales técnicas utilizadas en la Ciencia de Datos y su aplicabilidad a temas relacionados con el estudio de caso.
3. Determinar las características que representan el proceso quirúrgico de colocación de un implante dental.
4. Relevar, confeccionar y adaptar un conjunto de datos de historias clínicas de implantes dentales para aplicar técnicas de Ciencia de Datos.
5. Analizar la estructura y tipos de datos del conjunto.
6. Preparar los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la aplicación de procedimientos de Ciencia de Datos sobre ellos.
7. Caracterizar los tipos de tratamientos de superficie de los implante dentales.
8. Analizar las particularidades y necesidades de adaptación de los métodos necesarios para aplicar procedimientos de Ciencia de Datos sobre el estudio de caso.
9. Diseñar un procedimiento de Ciencia de Datos que permita:
    a. Seleccionar las características más importantes del conjunto de datos.
    b. Predecir fracasos.
    c. Determinar las condiciones influyentes en el fracaso del proceso de oseointegración (tejido óseo / implante).
10. Realizar pruebas para evaluar la calidad del procedimiento propuesto.
11. Realizar una validación del procedimiento propuesto con expertos humanos (Implantólogos).
12. Analizar los resultados obtenidos y elaborar conclusiones del trabajo realizado.
13. Escribir y defender de la tesis doctoral.

1.4 DESARROLLO DE LA TESIS

La tesis se estructura principalmente en ocho capítulos: “Introducción”, “Marco Teórico”, “Materiales y Métodos”, “Resultados y Discusión”, “Conclusiones”, “Trabajos Futuros”, “Producción Científica” y “Referencias” a los que se agregan tres anexos con información complementaria.

En el capítulo “Introducción” se presenta el contexto y problema de estudio de esta Tesis Doctoral, la solución propuesta, los objetivos planteados y se describe el contenido del documento.

En el capítulo “Marco Teórico” se presenta una revisión de la bibliografía relevante en cuanto a los diferentes aspectos de la problemática abordada. En donde, se introduce conceptos relacionados al descubrimiento de conocimiento, se describen metodologías para procesos de extracción de información más utilizados, se abordan numerosos métodos de Ciencia de Datos y términos necesarios sobre Implantología Dental.

En el capítulo “Materiales y Métodos” se presentan los materiales, las soluciones y las técnicas utilizados en esta Tesis Doctoral para alcanzar los objetivos propuestos.

En el capítulo de “Resultados y Discusión” se presenta los resultados obtenidos y se plantea la discusión de los mismos.

En el capítulo “Conclusiones” se exponen las aportaciones a las que se arriba en la presente Tesis Doctoral.

En el capítulo “Trabajos Futuros” se señalan futuras líneas de investigación relacionadas con aspectos del desarrollo de esta investigación.

En el capítulo “Producción Científica” se presenta la producción obtenida como producto de la presente Tesis Doctoral.

En el capítulo “Referencias” se listan todas las publicaciones consultadas para el desarrollo de esta Tesis Doctoral.

2. MARCO TEÓRICO

2.1 CIENCIA DE DATOS

La Ciencia de Datos, del inglés *Data Science*, es un campo interdisciplinario cuyo objetivo es extraer valor de los datos. Es una disciplina que surge de los campos del análisis estadístico y de la Minería de Datos [12]. Tiene como tarea el desarrollo de estrategias para analizar los datos, mediante la construcción de modelos de aprendizaje, con el propósito de generar información que les sirva a las organizaciones o instituciones para la toma decisiones [13].

Este campo se nutre de tres áreas o núcleos (Figura 1) y engloba habilidades asociadas al procesamiento de datos.

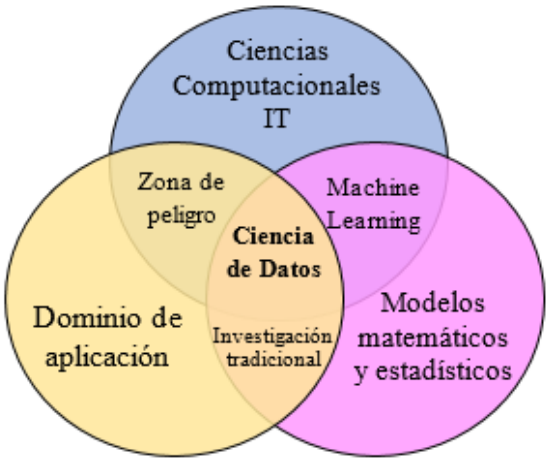


Figura 1. Áreas intervinientes en la Ciencia de Datos. [13]

En proyectos de este tipo es necesario precisar la "pregunta" que se quiere responder. Definir la pregunta es una de las partes más importantes. Cuando se habla de pregunta, lo que se quiere decir es: ¿Qué se quiere predecir?, ¿Cómo se puede garantizar la calidad de los datos?, ¿Qué modelo es el más apropiado? o ¿Qué técnicas se necesitará aplicar? Claro está, que la calidad de los datos (formatos, completitud, disponibilidad, integridad de fuentes, etc.) también es una parte muy importante. Cualquier dato que tiende a ser ruidoso, corrupto, incompleto o inconsistente puede afectar los resultados.

La Ciencia de Datos es ampliamente utilizada y aplicada a una gran variedad de dominios, como en bancos para modelos de riesgo crediticios; en procesos de manufactura para la detección de anomalías o predicción de potenciales problemas; en la asistencia sanitaria para no solo el análisis de imágenes médicas, sino que también para el descubrimiento de medicamentos, modelos predictivos para el diagnóstico de enfermedades, bioinformática y asistencias virtuales; en el trasporte en coches auto-conducidos; en finanzas para la toma de decisiones estratégicas y en muchas otras áreas [1].

En fin, la Ciencia de Datos es un campo de estudio que incluye todo, desde análisis de grandes conjuntos de datos (Big Data) [14], Minería de Datos, modelado predictivo, visualización de datos, matemáticas y estadísticas [13].

Los pasos que generalmente involucra y atraviesan los proyectos de Ciencia de Datos son cinco (Figura 2), abarcando desde las tareas de obtención de las fuentes de datos, la limpieza, preparación, análisis y visualización de los datos, así como la construcción y optimización de modelos para la predicción de datos [13].

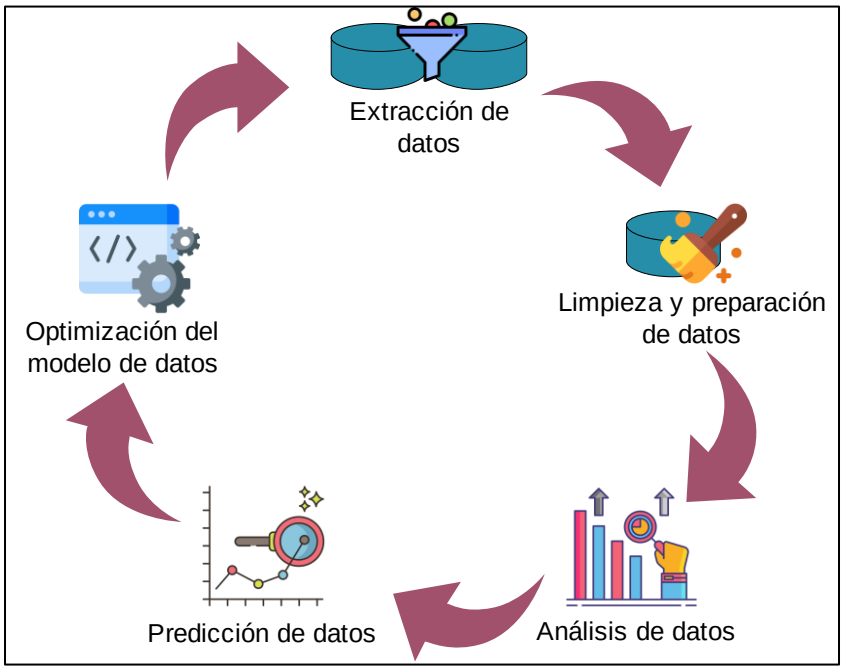


Figura 2. Pasos involucrados en la Ciencia de Datos. [13]

Como se observa en la Figura 2, la Ciencia de Datos comprende las siguientes etapas:

Extracción de datos: como primer paso la Ciencia de Datos inicia definiendo claramente el problema, para luego comenzar con la recuperación de datos. La recuperación incluye el descubrimiento, acceso y adquisición de los datos necesarios para el proyecto. Los datos pueden ser estructurados o no estructurados.

Limpieza y preparación de datos: implica la limpieza de los datos, la transformación y evaluación de datos faltantes. Aquí se examinan los datos para entender su naturaleza, calidad y formato. Este paso es trascendental para el desarrollo del proceso, ya que organiza los datos y los convierte en útiles para el análisis posterior.

Análisis de datos: involucra la selección de técnicas analíticas para encontrar patrones o tendencias en los datos. En esta etapa la visualización de los datos es muy importante, ya que permite a través de la representación gráfica captar, identificar factores o predecir

comportamientos. Conjuntamente, ayuda a la toma de decisiones sobre las técnicas más adecuadas a ser empleadas.

Predicción de datos: consiste en generar modelos que permitan realizar predicciones útiles con los datos, esto mediante la utilización de algoritmos de aprendizaje. Existe una gran variedad de algoritmos utilizados para predicción y clasificación, los cuales permiten trabajar con datos históricos para pronosticar y capturar patrones ocultos en los datos.

Optimización del modelo de datos: esta etapa está destinada a optimizar el modelo de aprendizaje automático desarrollado en el paso anterior, con la finalidad de mejorar su rendimiento y ofrecer resultados precisos.

2.2 KDD

El proceso de extracción de conocimiento en base de datos, del inglés *Knowledge Discovery in Databases* (KDD), tiene como objetivo el descubrimiento de conocimiento e información útil en grandes conjuntos de datos [15]–[17]. Este es un proceso metodológico no automático e iterativo que examina los datos para determinar relaciones. Permite extraer información de calidad y puede usarse para mostrar conclusiones basadas en las relaciones de los datos.[17]

Las primeras etapas del proceso KDD, como lo muestra la Figura 3, comprenden el procesamiento de los datos, luego se aplica un proceso de Minería de Datos [12], para finalmente obtener patrones de comportamiento que permitan la evaluación e interpretación del conocimiento descubierto.

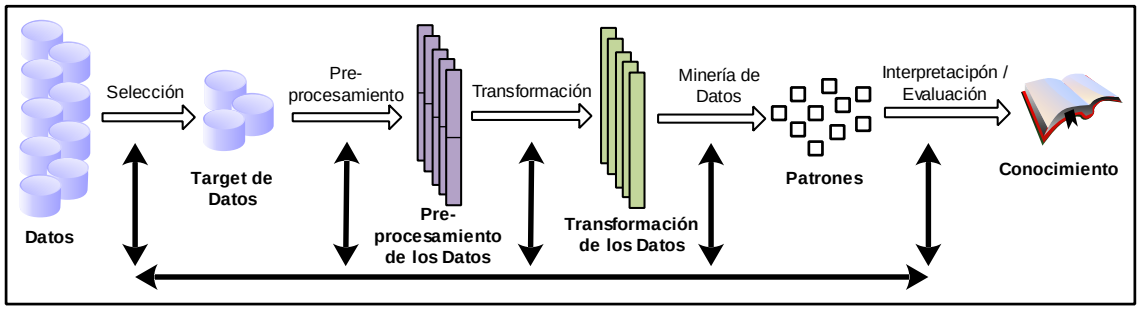


Figura 3. Proceso KDD. [15]–[17]

El proceso KDD se divide en cinco fases (Figura 3) y estas son:

Selección de datos: reside en determinar las fuentes de donde se extraerán los datos y el tipo de información a utilizar. Esta primera etapa permite confeccionar el conjunto de datos, el cual es la colección de información, ya sea cuantitativa o cualitativa y que está compuesto por columnas (variables, características o atributos) que representan las propiedades del problema de estudio, y por filas (casos) que hacen referencia a sucesos que se presentaron en el escenario de estudio.

Pre-procesamiento: consiste en la preparación y limpieza de datos extraídos desde distintas fuentes, en una forma manejable. Se aplican distintos puntos de vistas para manejar datos faltantes, datos inconsistentes, outliers [18] o que están fuera de rango, obteniéndose al final una estructura adecuada para su posterior transformación. Esta fase es de vital importancia en procesos de extracción de conocimiento, ya que la detección de datos considerados anómalos usando técnicas apropiadas, podría detectar previamente grupos de datos que pueden ser de especial interés al estudio.

Transformación: consiste en el tratamiento preliminar de los datos, transformación y generación de las variables a partir de las ya existentes, con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente. Además, en esta etapa se pueden aplicar técnicas para reducir la dimensionalidad del conjunto de datos [19].

Minería de Datos: en esta fase se aplican métodos inteligentes, ya sean de clasificación, regresión o agrupación (la selección de los métodos dependerán de los objetivos planteados), con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles, que están contenidos u "ocultos" en los datos.

Interpretación y Evaluación: se identifican los patrones obtenidos e interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Además de las fases descritas, frecuentemente se incluye una fase previa de análisis de las necesidades y definición del problema, en la que se establecen los objetivos del proceso de extracción de conocimiento. También es usual incluir una etapa final, donde los resultados obtenidos se integran al negocio, para la realización de acciones preventivas o dar soporte a la toma de decisiones [17], [20].

2.2.1 MINERÍA DE DATOS

La Minería de Datos se encuentra relacionada con el proceso de Ciencia de Datos, ya que brinda las técnicas necesarias para la construcción de los modelos predictivos, y de esta manera permite la detección de información o hechos previamente desconocidos o ignorados en los datos.

La Minería de Datos, del inglés *Data Mining* ,se define como *"un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir, de forma automatizada, tendencias o comportamientos y descubrir modelos previamente desconocidos"* [21]. Es decir, la Minería de Datos plantea dos desafíos, por un lado trabajar con grandes bases de datos y por el otro aplicar técnicas que conviertan en forma automática los datos en conocimiento.

Como se observa en la Figura 3 la Minería de Datos es parte del proceso KDD y se aplica a numerosas áreas, como: comercio (segmentación de clientes, previsión de ventas, análisis de riesgo), medicina (diagnóstico de enfermedades, clasificación de tumores, efectividad de tratamientos), seguridad y detección de fraude (reconocimiento facial, identificaciones biométricas), recuperación de información (minería de texto, minería web), astronomía (identificación de estrellas y galaxias), agricultura, pesca y minería (identificación de áreas de uso para cultivos, pesca o explotación minera en bases a datos de imágenes satélites), entre muchas otras [22].

Una característica transcendental de la Minería de Datos es que invierte la dinámica del método científico. Es decir, primero se recopilan los datos y posteriormente se los examina para que de ellos surjan las hipótesis. Por lo tanto, la Minería de Datos presenta un enfoque más bien exploratorio, y no confirmador.

2.3 TIPOS DE MODELOS DE APRENDIZAJE

El proceso de análisis y extracción de conocimiento en conjuntos de datos genera modelos que pueden ser de dos tipos, descriptivos o predictivos [23]–[25].

2.3.1 MODELOS DESCRIPTIVOS

Los modelos descriptivos son aquellos que exploran las particularidades y propiedades de los datos, con el objetivo de generar etiquetas o agrupaciones. Estos modelos siguen un tipo de aprendizaje no supervisado, dado que no se dispone de datos "etiquetados" para el entrenamiento. Por lo tanto, solo permiten resumir o describir la estructura de los datos, para intentar encontrar algún tipo de organización o información oculta en ellos. Por esta razón, tienen un carácter exploratorio [23], [24].

Los modelos descriptivos o de aprendizaje no supervisado, frecuentemente son aplicados a problemas de agrupación (Clustering) [26]–[34], correlación [35], [36] o de asociación [37]–[41].

2.3.2 MODELOS PREDICTIVOS

Los modelos predictivos tienen como objetivo la estimación de valores desconocidos de variables o características de interés. Estos modelos siguen un tipo de aprendizaje supervisado, dado que se dispone de datos "etiquetados" para el entrenamiento. Intenta encontrar una función que, dadas las características de entrada, le asigne la etiqueta de salida adecuada. Por lo que, el modelo (algoritmo) se entrena (aprende) con un histórico de datos y predice la etiqueta de salida para un nuevo caso. El atributo a predecir se lo conoce como

variable dependiente u objetivo, mientras que los atributos utilizados para entrenar se llaman variables independientes [23], [24].

Principalmente, los modelos predictivos o de aprendizaje supervisado se aplican a problemas de clasificación [10], [42], [51], [43]–[50] o de regresión [52], [53].

2.4 TIPOS DE TÉCNICAS DE APRENDIZAJE

2.4.1 TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

2.4.1.1 CLUSTERING

El agrupamiento, del inglés *Clustering* trata de analizar datos para generar etiquetas. Se aplican a problemas en los que se desea agrupar las instancias creando clusters o grupos de similares características. Se utilizan por ejemplo, en la segmentación de clientes de un supermercado, jerarquías por temas, en medicina para el diagnóstico de enfermedades a través de imágenes, en el monitoreo de redes sociales, mercadotecnia, finanzas, entre otros.
Existen dos grandes técnicas de agrupamiento:

1. Clustering jerárquico, que puede ser aglomerativo (se empieza a agrupar desde cada elemento individual) o divisivo (se parte de un único cluster que engloba todos los datos y luego se divide en clusters más pequeños).
    - Ejemplos de algunos algoritmos: DIANA/AGNES [54], BIRCH [28], [55], CURE [29], [56], CHAMELEON [30], [57], ROCK [58], [59], Two-Step Cluster (variante de BIRCH) [60].
2. Clustering no jerárquico, el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su proximidad. La decisión del número de clusters (conocido como K) es uno de los retos de este tipo de técnicas de agrupamiento.
    - Ejemplos de algunos algoritmos por particiones: K-means (es un algoritmo muy conocido y utilizado en clustering) [61], Fuzzy C-Means [31], PAM/CLARA/CLARANS [32], BFR [33], Mapas auto organizados (SOM) o Redes de Kohonen [62].
    - Ejemplos de algunos algoritmos por densidad: DBSCAN [34], [63], OPTICS [64], DenClue [65], SNN/ Jarvis-Patrick [66], KNNCLUST [67].

2.4.1.2 REGLAS DE ASOCIACIÓN

Las reglas de asociación tienen como objetivo encontrar relaciones no explicitas entre atributos. Por lo que buscan patrones frecuentes, asociaciones, correlaciones, o estructuras causales en los conjuntos de datos. Se aplican típicamente en el análisis del contenido de un carrito de compra, para la identificación de artículos que muchos clientes compran

conjuntamente (y de esta manera diseñar estanterías estratégicas). Además, se aplican a búsquedas de secuencias o patrones temporales.

- Ejemplos de algunos algoritmos: Apriori/AprioriTid (el algoritmo Apriori es ampliamente utilizado para problemas de asociación) [68], FP-Growth [69], OPUS [70], DHP/DIC [40], CHARM [71], CLOSET [72], ECLAT [41].

2.4.1.3 CORRELACIÓN

La correlación es similar a las reglas de asociación, solo que se utilizan para determinar el grado de similitud de los valores de dos variables numéricas. Estos algoritmos ayudan a entender mejor cómo varía una variable en función de otra. Es decir, consisten en ver la relación de crecimiento o decrecimiento de una variable observando el crecimiento de la otra. Por ejemplo: la correlación entre altura y el peso.

- Ejemplos de algunos algoritmos: Correlación (coeficiente de Pearson, Spearman, Kendall) [73].

2.4.2 TÉCNICAS DE APRENDIZAJE SUPERVISADO

2.4.2.1 CLASIFICACIÓN

El problema fundamental en la clasificación está directamente relacionado con la separabilidad de las clases. Dado un conjunto de datos, donde cada tupla se encuentra etiquetada con el valor de la clase a la que pertenece, el objetivo de este tipo de técnicas es predecir a que clase pertenece una nueva instancia, considerando que los atributos pueden asumir valores discretos.

Las técnicas de clasificación se diferencian por trabajar con variables objetivos de tipo categóricas. Dentro de estas técnicas existen dos tipos de clasificación, la binaria donde solo se puede asignar dos clases posibles (0 o 1 - por ejemplo, éxito o fracaso) y la clasificación multiclase, la cual permite asignar varias categorías a la clase (por ejemplo el conjunto de datos Iris – setosa, versicolor, virginica).

Este tipo técnicas tiene un sinfín de aplicaciones, como la detección de fraudes o riesgos financieros, predicción de enfermedades, clasificación de imágenes, entre muchas otras.

- Ejemplos de algunos algoritmos: Máquinas de Vector Soporte (SVM) [5], [74], Árboles de Decisión TDIDT (J48, ID3, C4.5) [8], [75]–[77], Árboles Aleatorios (Bagging, Random Forest, AdaBoost) [78], [79], K vecino más cercano (KNN) [9], [80], Redes Bayesianas (Gaussian, Multinomial) [10], [43], [48], Redes Neuronales (Perceptrón, Perceptrón Multicapa - MLP) [6], [81], [82], Regresión Logística [83], [84], Análisis Discriminante [85], Algoritmos Genéticos [24], [86], [87].

2.4.2.2 REGRESIÓN

En este caso, el valor a predecir es numérico. Por lo que estas técnicas se distinguen por trabajar con variables objetivos de tipo numérico. La regresión lineal es el ejemplo más popular de los algoritmos de regresión. Estos algoritmos su utilizan para predecir precios, probabilidad de que los clientes se desvíen de un producto, predicciones meteorológicas, expectativa de vida, de crecimiento, entre otras.

- Ejemplos de algunos algoritmos: Regresión Lineal (GLM) [52], Regresión con Vectores de Soporte (SVR) [53], Regresión del Proceso Gaussiano (GPR) [53], Regresión con Árboles de Decisión (rpart) [88], Regresión con Bosques Aleatorios (Random Forest Regression) [52], Regresión con Redes Neuronales [89]–[92].

2.5 METODOLOGÍAS PARA PROYECTOS DE CIENCIA DE DATOS

Un proyecto de Ciencia de Datos involucra, en general las siguientes fases [93]: comprensión del negocio y del problema que se quiere resolver, determinación, obtención y limpieza de los datos necesarios, creación de modelos matemáticos, ejecución, validación de los algoritmos, comunicación de los resultados obtenidos; e integración de los mismos. La relación entre todas estas fases implica una complejidad que se traduce en una jerarquía de sub fases. Por lo tanto, ante la necesidad de integración y una aproximación sistemática para la implementación de proyectos de este tipo, diversas empresas han especificado un proceso de modelado, diseñado para guiar al usuario a través de una sucesión formal de pasos.

2.5.1 CRISP-DM

En el año 1999 un grupo de empresas europeas, NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), desarrollaron una metodología de libre distribución denominada CRISP-DM (Cross-Industry Standard Process for Data Mining) [94]. Esta metodología estructura el ciclo de vida de un proyecto de Ciencia de Datos en seis fases, que interactúan entre sí de forma iterativa durante el desarrollo del proyecto (Figura 4) [95]. Es de libre distribución, la más utilizada y recomendada para aplicar a procesos de este tipo.

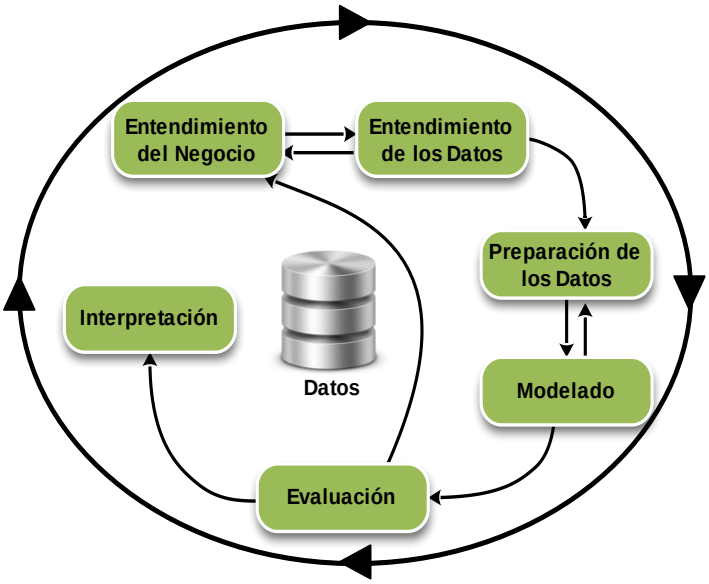


Figura 4. Fases de la metodología CRISP-DM. [94]

Las flechas internas de la Figura 4, indican las relaciones más frecuentes entre las fases, aunque se pueden establecer relaciones entre cualquier fase. Mientras, que el círculo exterior simboliza la naturaleza cíclica del proceso de modelado. Las etapas de esta metodología son:

Entendimiento del negocio: incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

Entendimiento de los datos: comprende la recolección inicial de datos, permite establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

Preparación de los datos: incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado, limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. La fase de preparación de los datos, se encuentra muy relacionada con la fase de modelado, puesto que en función de la técnica de modelado que se emplee, los datos requerirán ser procesados en diferentes formas.

Modelado: implica la selección de técnicas de modelado más apropiadas para el proyecto de Ciencia de Datos. Las técnicas a utilizar en esta fase se seleccionan en función de los siguientes criterios: ser apropiada al problema, disponer de datos adecuados, cumplir con los requerimientos, tiempo necesario para obtener un modelo y conocimiento de la técnica.

Evaluación: es esta fase se evalúa el modelo, no desde el punto de vista de los datos, sino desde el cumplimiento de los objetivos. Se debe revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso, en el que se haya podido cometer errores.

Si el modelo generado es válido en función de los criterios establecidos en la fase inicial, se procede a la aplicación del mismo.

Interpretación: la última fase puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos. Es fundamental documentar todo el proceso y presentar los resultados de manera comprensible.

En la Tabla 1 se detallan las tareas que componen cada una de las seis etapas [77], [96]–[99].

Tabla 1. Tareas de cada fase de la Metodología CRISP-DM.

Entendimiento del negocio	Entendimiento de los datos	Preparación de los datos	Modelado	Evaluación	Interpretación
Determinar los objetivos del negocio Antecedentes Objetivos Criterios de éxito **Evaluar situación** Inventario de recursos, requerimientos, criterios, restricciones, riesgos, contingencias, Terminologías, costos y beneficios **Determinar Metas** Metas de minería de datos y criterios de éxito **Generar plan de proyecto** Plan de proyecto, evaluación inicial de herramientas y técnicas	**Recolección inicial de datos** Reporte inicial de datos **Descripción de datos** Reporte inicial de la descripción de datos **Exploración de datos** Reporte de exploración de datos **Verificación de calidad de datos** Reporte	**Selección de datos** Reporte de Inclusión / exclusión **Limpieza de datos** Reporte **Construcción de datos** Atributos derivados, generación de registros **Integración de datos** Datos fusionados **Formato de datos** Datos reformateados Set de datos y descripciones	**Selección de técnicas** Técnicas de modelado, modelos supuestos. **Diseño de pruebas** Pruebas, Diseño de pruebas **Construcción del modelo** Ajuste de parámetros, modelos, descripciones **Evaluar modelo** Evaluar Revisión de parámetros	**Evaluar resultados** Evaluación de datos, minería y criterios de éxito de negocio. Aprobación del modelo **Revisión de procesos** Revisión de procesos **Determinar próximos pasos** Lista de posibles acciones, Decisión	**Desarrollo del plan** Desarrollar el plan **Plan de monitoreo y mantenimiento** Plan monitoreo y mantenimiento **Generar reporte final** Reporte final, presentación final **Revisión del proyecto** Documentar experiencia

2.5.2 TDSP

El proceso de Ciencia de Datos en equipo, del inglés *Team Data Science Process* (TDSP) es *"una metodología de Ciencia de Datos ágil e iterativa para ofrecer soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente "* [100], de esta manera lo describen en el sitio web de Microsoft.

Esta metodología ayuda a mejorar la colaboración y el aprendizaje en equipos de Ciencia de Datos. TDSP contiene mejores prácticas y estructuras de Microsoft, así como de otras empresas líderes de la industria, para ayudar en la implementación correcta de iniciativas de Ciencia de Datos.

Sus principales componentes son [100], [101]:

1. Definición de un ciclo de vida de Ciencia de Datos.
2. Estructura de proyecto estandarizada.
3. Infraestructura y recursos recomendados para proyectos de Ciencia de Datos.
4. Herramientas y utilidades recomendadas para la ejecución de proyectos.

El ciclo de vida de TDSP es similar a CRISP-DM y su coordinación de procesos utiliza varios elementos de la metodología ágil Scrum [102], incluidos artefactos claramente definidos, backlog, sprints y roles de equipo [101].

El ciclo de vida de la metodología TDSP [101] describe cinco fases principales por las que normalmente pasan proyectos de este tipo, y a menudo de forma iterativa:

Comprensión del problema empresarial: esta etapa consiste en definir objetivos e identificar fuentes de datos. La definición de objetivos está orientado a identificar las variables comerciales o de negocio que se necesita predecir. Además, implica definir los objetivos del proyecto a través de preguntas precisas, tales como:

- ¿Cuánto? Si el problema a resolver consiste en estimar la relación entre variables numéricas, seguramente será necesario analizar la aplicación de técnicas de regresión.
- ¿Qué categoría? Si el problema a resolver consiste en estimar o predecir una categoría o etiqueta, se necesitará aplicar técnicas de clasificación.
- ¿Qué grupo? Si los datos a tratar no cuenta con una etiqueta, se tendrá que evaluar y definir técnicas de agrupamiento para extraer información oculta en los datos, mediante el análisis de grupos que cuenten con características o particularidades similares.
- ¿Es este caso raro? Ante la situación de eventuales casos, será necesario definir técnicas para la detección de anomalía.
- ¿Qué opción debería tomarse? Si el problema a abordar necesita proporcionar alguna recomendación, entonces será necesario definir técnicas para implementar sistemas de recomendación.

Conjuntamente, en este paso se define el equipo interviniente, especificando roles y responsabilidades de los miembros y las métricas de éxito con las que se medirá los resultados del proyecto.

El equipo se define mediante cuatro funciones (no necesariamente excluyentes entre sí:

1. Administrador de grupo: supervisa toda la unidad de Ciencia de Datos.
2. Líder de equipo: administra el equipo de Ciencia de Datos.
3. Líder de proyecto: administra las actividades diarias en el proyecto especificado.
4. Colaborador individual del proyecto: miembro del equipo de desarrollo (incluye científico de datos, analista de negocios, ingeniero o arquitectos de datos, especialistas, etc.).

En cuanto a las fuentes de datos, estas se identifican en función de los objetivos, las necesidades y las características de interés del dominio del problema.

Adquisición y comprensión de los datos: el objetivo de esta etapa es producir un conjunto de datos limpio y de calidad. Además, desarrollar una arquitectura de solución a través de tres tareas: ingesta de datos, exploración de datos y una configuración para la canalización de los datos.

Modelado: la meta de esta fase es identificar las características y crear un modelo de aprendizaje automático que sea óptimo, preciso y adecuado para la producción. Esta fase consta de tres tareas: ingeniería de características, entrenamiento de modelos y evaluación de los modelos.

Despliegue: el único propósito de esta fase es la de implantar modelos en un entorno de producción o similar al de producción para la aceptación final del cliente, exponiendo los modelos a través de una interfaz abierta.

Aceptación del cliente: la fase final permite verificar que se cumplan las necesidades de los clientes. Las dos tareas principales que incluye esta fase son: validar el sistema y transferir el proyecto. Finalmente, para el cierre y documentación del proyecto se emite un informe de salida para el cliente. Este informe es técnico y contiene todos los detalles del proyecto que son útiles para aprender a operar el sistema. TSDP proporciona plantillas de informes, los cuales pueden ser personalizados en funciones de las necesidades específicas del cliente.

La Figura 5 es una representación visual del ciclo de vida de la metodología TDSP propuesta por Microsoft [100].

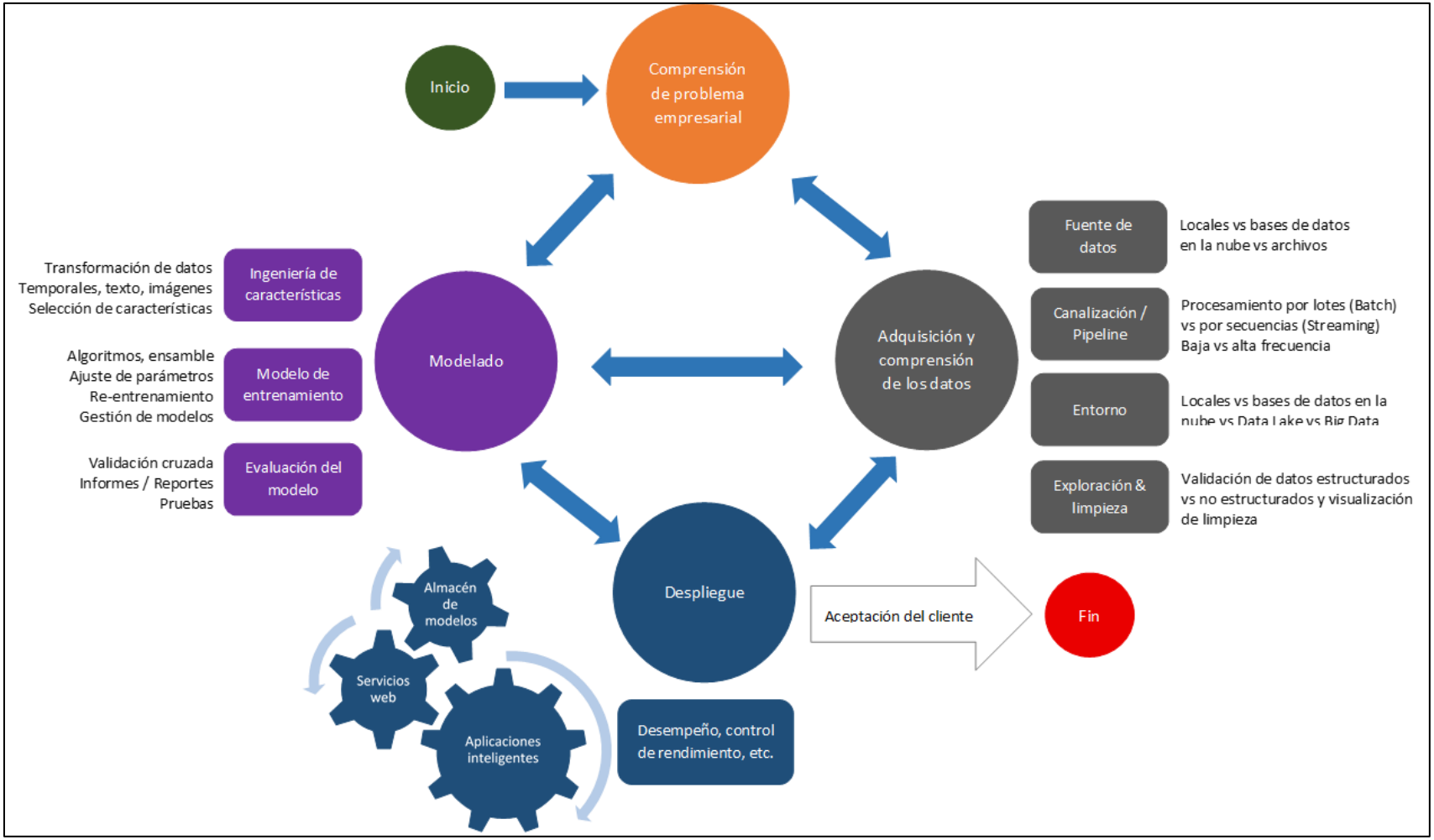


Figura 5. Ciclo de vida de TDSP. [100]

2.5.3 ASUM-DM

La metodología ASUM-DM, del inglés *Analytics Solutions Unified Method for Data Mining* (ASUM-DM), es un proceso iterativo para implementar proyectos de Minería de Datos basado en la metodología CRISP-DM. Es un proceso propuesto por IBM y hace énfasis en las nuevas prácticas de la Ciencia de Datos, como el uso de Big Data, análisis de texto, modelado predictivo y automatización de procesos [103].

Fue creado para acelerar el tiempo de generación de valor y disminución del riesgo mediante el establecimiento de enfoques y procesos coherentes que aumentan la eficiencia de su implementación. Contiene pasos estructurados, actividades de desarrollo, roles y responsabilidades, plantillas y directrices.

IBM define a ASUM como una guía de pasos para efectuar la implementación del ciclo de vida de soluciones de analítica de datos. Está compuesto por las siguientes seis fases [103]:

Analizar: en esta fase se determinan los requisitos y necesidades del usuario, objetivos del proyecto, los entregables y se prevé de forma inicial las primeras soluciones posibles.

Diseñar: incluye los procesos que tienden al desarrollo de la minería y análisis de datos. Conjuntamente, se identifica los recursos y se instala el entorno de desarrollo.

Configurar y construir: configurar, construir e integrar componentes basados en un enfoque iterativo e incremental. En esta fase se define un plan de pruebas y validación en múltiples entornos.

Despliegue: esta fase establece el flujo de trabajo para implementar los resultados dentro de la empresa de forma que no afecte la actividad productiva normal de la misma. Debe incluir programas de apoyo.

Operar y optimizar: esta fase cumple el rol de controlador, analiza el proyecto implementado (el uso de la solución de IBM Analytics), determina la necesidad de aumentar el alcance del proyecto o de realizar pequeños ajustes o mantenimientos.

Gestión de proyecto: adicionalmente, ASUM-DM incorpora una quinta fase paralela (Figura 6), encargada de que las distintas fases del proyecto fluyan sin inconvenientes, esta labor es similar a la realizada por un Scrum Master, en la metodología Scrum. Esta fase ayuda a la gestión y el monitoreo del progreso y mantenimiento del proyecto.

Una característica relevante de esta metodología es la carencia de detalle en los procesos que conllevan a la minería y el análisis de datos. Además, incorpora una definición de los interesados del proyecto con las características de cada perfil y tareas típicas (como una matriz de asignación de tareas). Cabe resaltar que la guía para la implementación de ASUM-DM solo es accesible mediante un registro en el sitio web de IBM y mediante la instalación de archivos ejecutables que dan acceso a las guías en formato HTML, los cuales son compatibles solo con sistemas operativos Windows [103].

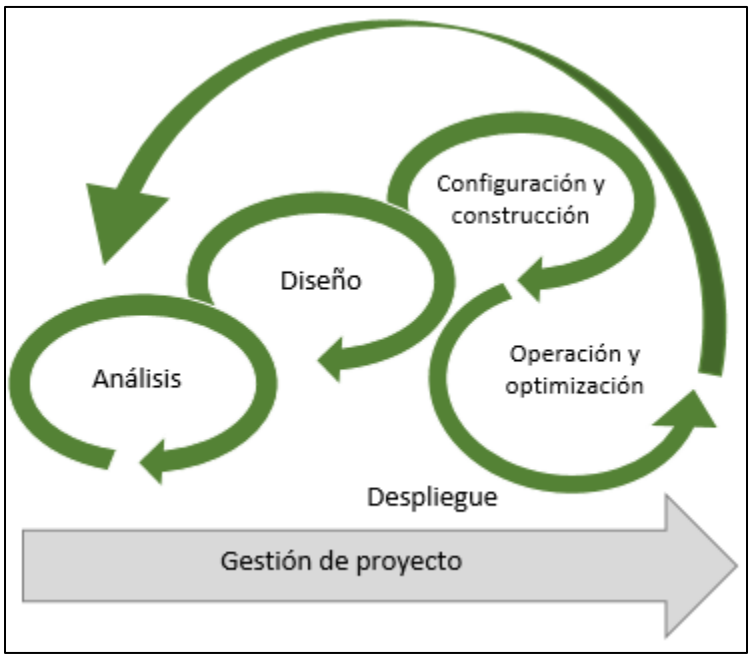


Figura 6. Ciclo de trabajo propuesto por la metodología ASUM-DM de IBM. [103]

2.5.4 SCRUM-DM

Scrum es una metodología de gestión de proyectos iterativa e incremental que involucra buenas prácticas, para controlar el riesgo y optimizar la gestión de un proyecto. En esta metodología los miembros del equipo deben trabajar juntos para entregar valor incremental (productos) [102].

En entornos de análisis de datos, Scrum requiere adaptaciones para funcionar con las características de proyectos de este tipo. Como por ejemplo, los equipos deben definir sprints (iteraciones de duración fija). Así como las metodologías TDSP y ASUM-DM se basan en algunos conceptos de Scrum y CRISP-DM, existen otras propuestas donde se integran metodologías, una de ellas es la metodología Scrum integrada con CRISP-DM (Scrum-DM).

La metodología Scrum-DM [104] utiliza elementos de la metodología ágil Scrum para la gestión del trabajo y CRISP-DM como estrategia para seguir el desarrollo del proyecto de Minería de Datos (Figura 7). Esta metodología se inicia a partir de la fase de comprensión o entendimiento del negocio de CRISP-DM donde se realiza el análisis de los objetivos y del problema y finaliza en la fase de despliegue donde se realiza la integración de los resultados de la Minería de Datos. El desarrollo se realiza en una fase intermedia, llamada Sprint, donde se concentran las demás fases de CRISP-DM.

Esta metodología también se basa en la recomendación de Scrum, donde los equipos deben estar formados por 3 a 9 miembros de desarrollo y puntualiza los tres roles [105]:

1. Propietario del producto (Product Owner): establece la visión del producto y define los incrementos potenciales del producto (también conocidos como historias de usuario o características).
2. Maestro de Scrum (Scrum Master): facilita el proceso Scrum (eliminando impedimentos) como líder del proyecto.
3. Equipo de desarrollo (Development Team): conformado por profesionales que entregan incrementos de producto. En el contexto de equipos de análisis de datos, serían: científicos de datos, ingenieros de datos, analistas de datos, analistas de sistemas e ingenieros de software, entre otros.

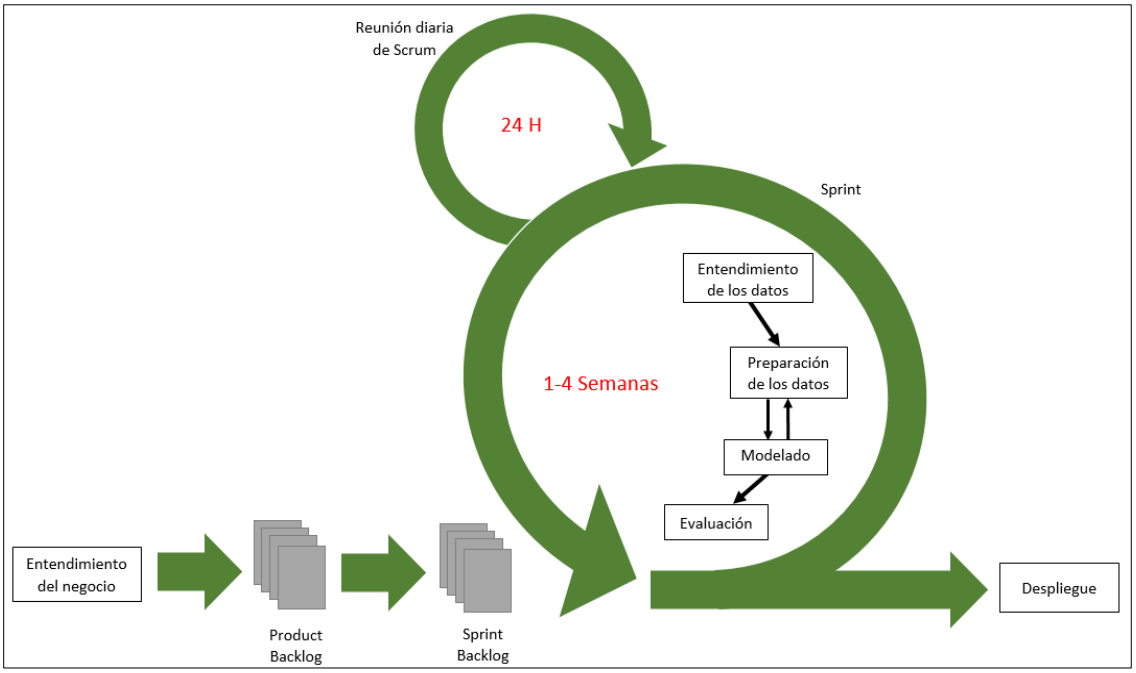


Figura 7. Pasos de la metodología Scrum-DM.

Como se aprecia en la Figura 7, Scrum-DM divide el proyecto de Minería de Datos en pequeños proyectos, cada uno de una duración constante y fija que va de una semana a un mes (como la metodología Scrum). Cada ciclo de esos pequeños proyecto, llamado Sprint, comienzan con una reunión (planificación de sprint) donde el propietario del producto (Product Owner) define y explica las principales prioridades. El equipo de desarrollo define qué incrementos pueden ofrecer al final del sprint (Product Backlog), es decir que tareas de Minería de Datos y otros requerimientos van a desarrollar, y luego elabora un plan de sprint para desarrollar estos incrementos (Sprint Backlog), en base a las tareas de la metodología CRISP-DM. Durante el sprint, se coordinan y desarrollan planes en las reuniones diarias (por ejemplo tipo de técnica a utilizar, métodos de visualización, etc.). Al final del sprint, el equipo demuestra los incrementos a las partes interesadas y solicita comentarios durante la revisión del sprint. Para cerrar un sprint, el equipo se examina a sí mismo y planifica cómo puede mejorar en el siguiente sprint durante la retrospectiva del sprint.

2.5.5 SEMMA

A esta metodología se la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones desconocidos. Esta metodología es propuesta por la empresa SAS, especialmente para trabajar con su propio software de minería de datos. El acrónimo SEMMA se corresponde a sus cinco fases (Figura 8).

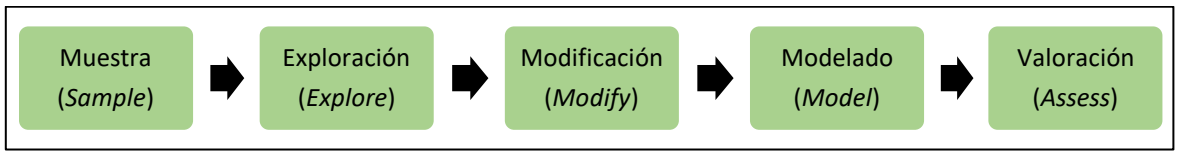


Figura 8. Fases de la metodología SEMA. [106], [107]

Las etapas que componen a esta metodología son:

Muestra (*Sample*): la metodología inicia con la extracción de una muestra de la población sobre la que se va a trabajar. El objetivo de esta fase consiste en seleccionar una muestra representativa del problema en estudio. La representatividad de la muestra es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan de ser admisibles. La forma más común de obtener una muestra es la selección al azar, además para cada muestra a considerar se debe asociar un nivel de confianza de la misma.

Exploración (*Explore*): una vez determinada la muestra o conjunto de muestras representativas de la población, la metodología SEMMA indica que se debe proceder a una exploración de la información disponible, con el fin de simplificar en lo posible el problema para optimizar la eficiencia del modelo. Para cumplir este objetivo propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo.

Modificación (*Modify*): consiste en la manipulación y formateo de los datos. En esta etapa se procede a la limpieza de valores anómalos, se realiza tratamientos para datos faltantes, y se seleccionan, crean o modifican las variables con las que se trabajará.

Modelado (*Model*): consiste en la creación del modelo que permitirá predecir las variables de respuesta a partir de las variables explicativas. Incluye la utilización de métodos estadísticos tradicionales, así como técnicas de redes neuronales, lógica difusa, árboles de decisión, entre otras.

Valoración (*Assess*): en esta etapa se evalúa la utilidad y la exactitud del modelo obtenido. [106], [107].

En la Figura 9 se puede apreciar un esquema de la dinámica general de la metodología.

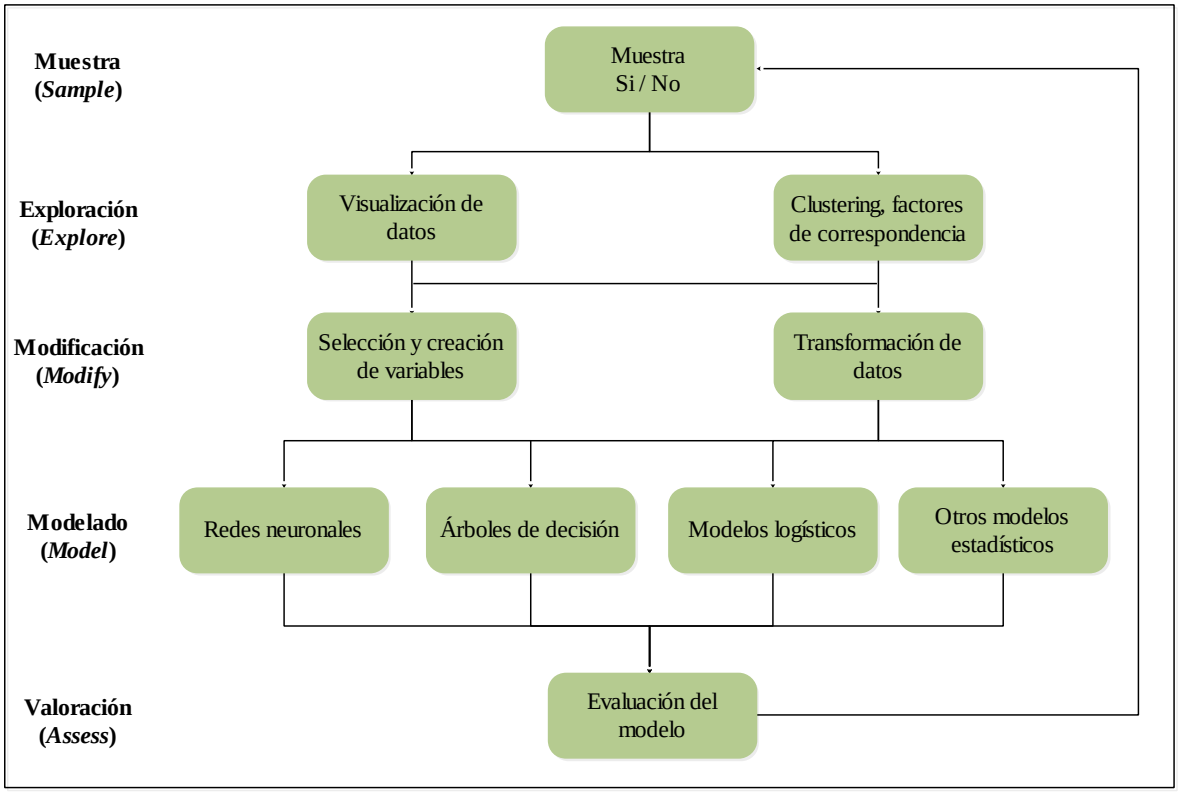


Figura 9. Dinámica de la Metodología SEMMA. [106], [107]

2.5.6 P^{3}TQ

La metodología Catalyst, comúnmente conocida como P^{3}TQ [108] (Product, Place, Price, Time, Quantity), fue propuesta por Dorian Pyle en el año 2003. Está compuesta por la formulación de dos modelos, el Modelo de Negocio y el Modelo de Extracción de Información (Figura 10). El Modelo de Negocio proporciona una guía de pasos para el desarrollo y la construcción de un modelo que permita identificar un problema de negocio. El Modelo de Extracción de Información proporciona una guía de pasos para la preparación de los datos, selección de las herramientas y modelos iniciales, ejecución, evaluación y comunicación de los resultados.

Pyle propone cinco situaciones o puntos de partida diferentes para un proyecto (Figura 10):

Dato: explorar los datos en búsqueda de relaciones útiles e interesantes.

Problema / Oportunidad: dado un problema u oportunidad de negocio, ver cómo el análisis de los datos puede colaborar con la misma.

Prospectiva: proyecto diseñado para descubrir dónde el análisis de los datos puede aportar valor en la organización.

Modelo definido: utilizar métodos o técnicas de Minería de Datos para construir un modelo específico para una situación determinada.

Estrategia: dada una situación estratégica, analizar si procesos de este tipo pueden ser útiles para explicar la situación actual y descubrir cuáles son las opciones para resolverla.

Tomando algunos de estos puntos o escenarios de partida, la metodología propone una serie de pasos y herramientas para llegar a descubrir el problema y los requerimientos organizacionales que abordará el proyecto, así como los datos necesarios para efectuar el análisis. La Figura 10 resume las etapas de cada modelo. En la parte superior (Modelo de Negocio) se encuentran los posibles escenarios de partida, en el centro las herramientas principales que se pueden utilizar, y en la salida la definición de los datos necesarios y los requerimientos para el proyecto. En la parte inferior (Modelo de Explotación de Información) se explicitan los pasos de esta metodología para el descubrimiento de patrones o relaciones, en base al problema de negocio identificado.

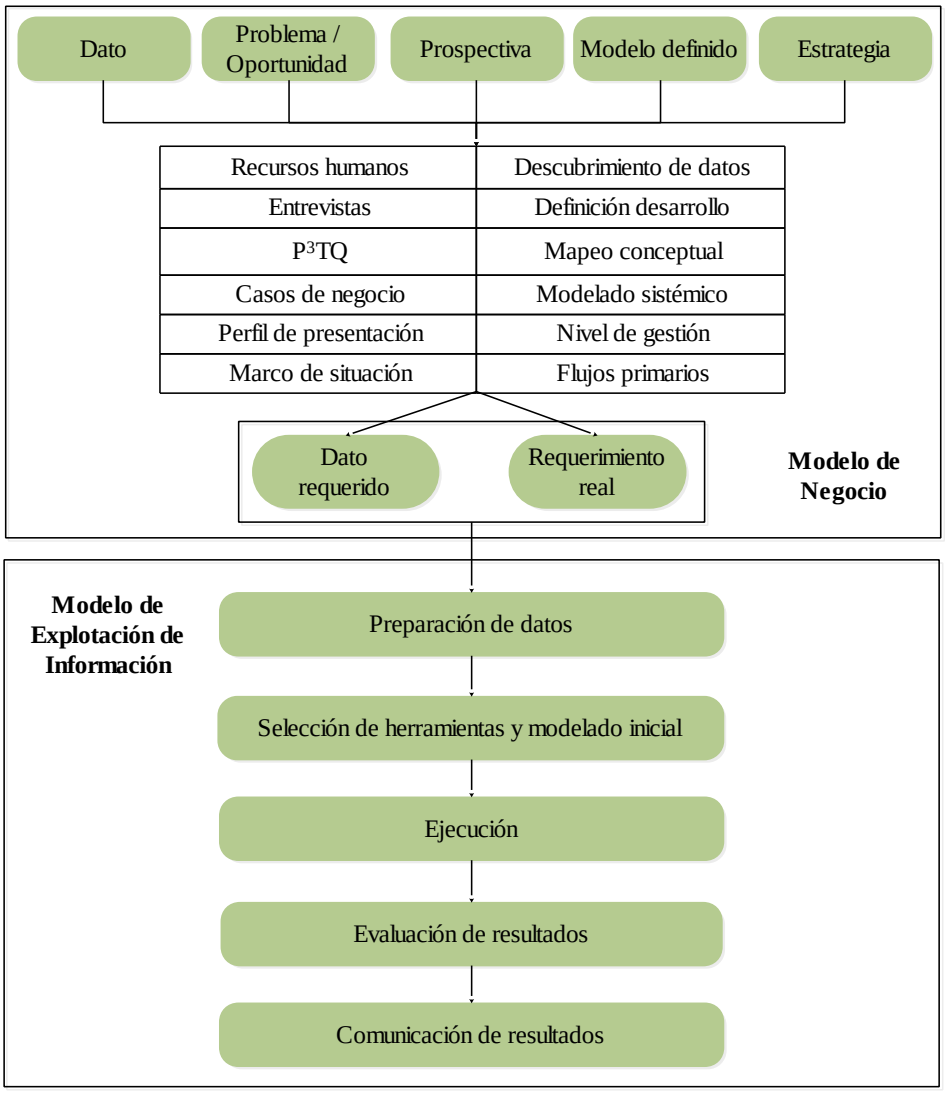


Figura 10. Fases de la metodología P^{3}TQ. [108]

2.5.7 COMPARACIÓN DE METODOLOGÍAS

Las metodologías CRISP-DM, TDSP, ASUM-DM, Scrum-DM, SEMMA y P^{3}TQ están compuestas por diversas fases relacionadas entre sí, con la finalidad de guiar el desarrollo de proyectos de Ciencia de Datos. A continuación, se resumen algunos puntos importantes de estas metodologías abordadas:

- El flujo de trabajo en CRISP-DM, TDSP, ASUM-DM, Scrum-DM y P^{3}TQ es iterativo, exceptuando SEMMA.
- ASUM-DM se origina como una extensión del estándar abierto CRISP-DM, TDSP no detalla su origen.
- ASUM-DM si bien es iterativa, establece un orden secuencial de sus fases (Figura 6), mientras que TDSP dispone estados en paralelo con múltiples posibilidades de inter-comunicaciones (Figura 5). Como se expuso anteriormente, ambas metodologías persiguen la modalidad ágil, a pesar de sus diferentes formas de aplicación [100], [103], [109].
- ASUM-DM es más estructurada y detallista, mientras que TDSP es más flexible y abierta [100], [103], [109].
- ASUM-DM identifica los roles según las responsabilidades distribuidas a los fines del proyecto, mientras que TDSP organiza los roles en función de la gestión del equipo de trabajo en Ciencia de Datos, a través de distintos proyectos [100], [103], [109].
- En ASUM-DM la figura del cliente tiene mayor presencia, ya que se le atribuyen un conjunto de roles que TDSP no reconoce dentro de la metodología [100], [103], [109].
- Scrum-DM integra conceptos de la metodología Scrum y CRISP-DM [104].
- CRISP-DM, SEMMA y P^{3}TQ son metodologías en cascada o tradicionales.
- La metodología SEMMA no proporciona una guía de actividades específicas a realizar en cada una de sus etapas y además, inicia su primera fase con el muestreo de datos, mientras que CRISP-DM y P^{3}TQ se centran en el análisis de los requerimientos y el entendimiento del negocio.
- CRISP-DM, SEMMA y P^{3}TQ contemplan la selección y preparación de los datos. Pero SEMMA propone trabajar con una muestra de los datos originales (en caso de tener un gran volumen de datos).
- SEMMA interpreta y evalúa los resultados en base al desempeño del modelo, mientras que P^{3}TQ realiza una validación en función de los objetivos del proyecto. Para el caso de CRISP-DM los resultados se evalúan en función del desempeño del modelo y el cumplimiento de los requerimientos iniciales del proyecto.
- Otro aspecto a considerar sobre CRISP-DM, SEMMA y P^{3}TQ es el uso de herramientas que pueden emplear, por ejemplo, SEMMA está relacionada con productos comerciales de la empresa SAS, como Enterprise Miner y Text Miner. Lo que conlleva que el análisis de los datos tenga que ajustarse a las técnicas y

herramientas de la misma. Sin embargo, CRISP-DM y P^{3}TQ fueron diseñadas como metodologías de libre distribución, por lo que pueden adaptarse a cualquier herramienta ya sea libre o comercial.

- TDSP es una buena opción para equipos de Ciencia de Datos que aspiran a entregar productos de Ciencia de Datos a nivel de producción. Puede que no sea apropiado para científicos de datos de un solo equipo o para proyectos sin un objetivo de producción [109].
- ASUM-DM carece de detalles en los procesos que conllevan a la minería y el análisis de datos. La guía para la implementación de ASUM-DM solo es accesible mediante un registro en el sitio web de IBM y mediante la instalación de archivos ejecutables que dan acceso a las guías en formato HTML, los cuales son compatibles solo con sistemas operativos Windows [103].

2.5.8 METODOLOGÍA SELECCIONADA

Las metodologías agiles son extremadamente flexibles y están pensadas para adaptarse al cambio, por lo que son más adecuadas para proyectos complejos. Mientras que para proyectos más sencillos o de menor escala, los enfoques tradicionales son los más apropiados. Además, los objetivos y la forma en que se lleva a cabo un proyecto en las metodologías en cascada son bien definidos y detallados.

En el apartado *2.5 METODOLOGÍAS PARA PROYECTOS DE CIENCIA DE DATOS* se abordaron algunas metodologías basadas en conceptos ágiles (TDSP, ASUM-DM y Scrum-DM), las cuales tienen la ventaja de dar rápidas respuestas a cambios. Pero para asentar el proceso de requerimiento de negocio, limpieza y preparación de los datos, modelado y evaluación de esta tesis, se emplea la metodología CRISP-DM. Esta metodología tiene las características de ser iterativa, de libre distribución, la más utilizada y recomendada para aplicar a procesos de este tipo. Además, CRISP-DM se caracteriza por hacer énfasis en los detalles de cada una de sus fases, es decir, cada etapa se divide en diferentes tareas y actividades. Por lo que, varios autores [110]–[116] utilizan, recomiendan e indican que esta metodología hace que los proyectos, grandes o pequeños, de Ciencia de Datos sean rápidos de desarrollar, fiables y manejables.

2.6 SELECCIÓN DE CARACTERÍSTICAS

La selección de características es una técnica de pre-procesamiento, que tiene por objeto identificar un subconjunto de características tan buenas o mejores que el conjunto de datos original y descartar las características redundantes e irrelevantes. Esto permite agrupar la información más relevante del conjunto de datos, lograr una predicción más precisa, reducir la dimensionalidad del conjunto de datos y mejorar el rendimiento y la complejidad computacional de los algoritmos de aprendizaje [117], [118].

Las características se definen generalmente como relevantes, irrelevantes o redundantes. Las características relevantes son las que influyen en el resultado y ninguna otra característica puede desempeñar el mismo papel, las irrelevantes son las que no influyen en el resultado, pueden detectarse mediante el método conocido como ganancia de información, utilizando la entropía o Shannon [119] y sus valores se generan aleatoriamente para cada ejemplo. Por último, las características redundantes son las que pueden desempeñar el papel de otra característica [19], [120], [121].

Para dar comienzo a una explicación matemática de estos métodos es necesario definir el concepto de entropía (ecuación 1). La entropía es utilizada muy comúnmente en la teoría de la información [122]. Esta permite caracterizar la pureza de una colección arbitraria de ejemplos. La entropía de Y es:

$$H(Y) = -\sum_{y \epsilon Y} p(y) log_2(p(y)) \qquad 1$$

donde $p(y)$ es la función de probabilidad para la variable aleatoria Y. Si los valores observados de Y en el conjunto de datos de entrenamiento S se dividen de acuerdo con los valores de una segunda característica X, y la entropía de Y con respecto a las particiones inducidas por X es menor que la entropía de Y antes de la partición, entonces hay una relación entre las características Y y X (ecuación 2). La entropía de Y después de observar X es entonces:

$$H(Y|X) = -\sum_{x \epsilon X} p(x) \sum_{y \epsilon Y} p(y|x) log_2(p(y|x)) \qquad 2$$

donde $p(y|x)$ es la probabilidad condicional de y dado x [119].

2.6.1 Tipos de Métodos de Selección de Características

En general, los métodos de selección de características pueden dividirse en tres categorías: Filter, Wrapper y Embedded [123]–[125].

2.6.1.1 Filter

Los métodos de filtrado, del inglés *Filter* [124] usan una escala aproximada para calificar un subconjunto de características y son considerablemente rápidos. Estos métodos utilizan una función de evaluación que se basa únicamente en las propiedades de los datos, por lo que son independientes de cualquier algoritmo específico. Su propósito es medir la relevancia de las características por su correlación con la variable dependiente. Por lo que utilizan métodos estadísticos para la evaluación de las características y un umbral para eliminar las características que estén por debajo de dicho valor.

– Como ejemplos de este tipo de métodos encontramos: Mutual Information [126], Coeficiente de correlación [19], Gain Ratio [127], Information Gain [119], Symmetrical Uncertainty [128], Relief [129], ReliefF [130], Chi-Square [131].

En la Figura 11, se aprecia el funcionamiento de estos tipos de métodos, donde primero evalúa cada característica individual, elige el subconjunto (en base al umbral definido) y luego los pone a prueba con un algoritmo de aprendizaje.

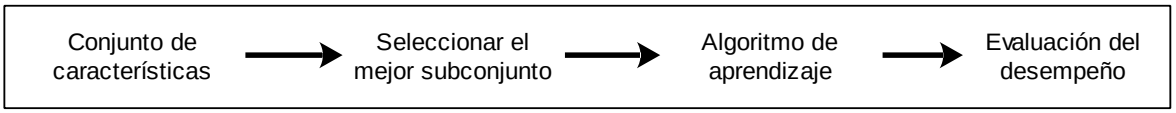


Figura 11. Funcionamiento del método Filter. [118], [132], [133]

2.6.1.2 WRAPPER

Los métodos de envoltura, del inglés *Wrapper* [134] primero emplean un algoritmo de optimización en el que se agregan o eliminan varias características para formar diferentes subconjuntos. Estos métodos miden la ganancia de las características para optimizar el rendimiento del clasificador. Lo que conlleva a ser computacionalmente más costosos en comparación con los métodos de filtrado, debido a los reiterados pasos de aprendizaje y validación cruzada. Durante el procedimiento de búsqueda (Figura 12) se generan y evalúan diversos subconjuntos de características. La evaluación de un subconjunto de características se basa en el entrenamiento y prueba del predictor (algoritmo de aprendizaje).

– Como ejemplos de este tipo de método encontramos: selección secuencial (Sequential Forward Selection), selección inversa (Sequential Backward Selection), búsqueda bidireccional (Bidirectional Search), relevancia en el contexto (Relevance in Context) [135].

Los métodos de envoltura buscan medir la utilidad de un subconjunto de características al entrenar un modelo. Además, utilizan validación cruzada para la evaluación de los subconjuntos.

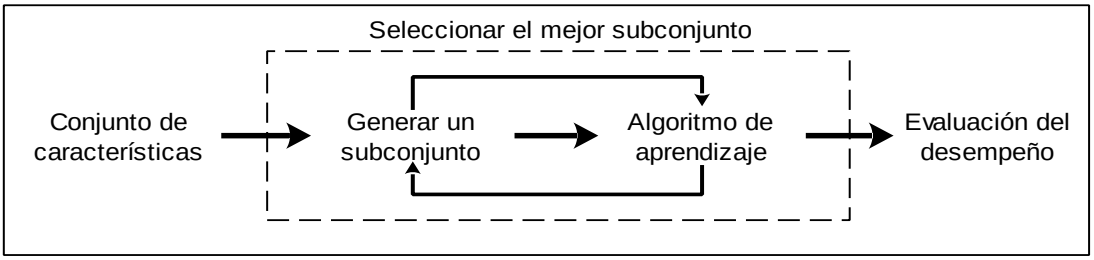


Figura 12. Funcionamiento del método Wrapper. [118], [132], [133]

2.6.1.3 EMBEDDED

Los métodos integrados, del inglés *Embedded* [123], [124] son bastante similares a los métodos de envoltura, ya que también se utilizan para optimizar la función objetivo o el rendimiento de un algoritmo / modelo de aprendizaje. Lo diferencia la incorporación de la selección de características como parte del proceso de entrenamiento.

Estos métodos consideran la selección de un subconjunto de características como un problema de búsqueda, en donde las diferentes combinaciones son evaluadas y comparadas. Para esto se utiliza un modelo predictivo y luego se asigna una puntuación a cada combinación (subconjunto de características) basada en la precisión del modelo. Finalmente, se selecciona el subconjunto que permitió lograr la mejor precisión (Figura 13).

- Como ejemplos de este tipo de método encontramos: regulación L1 (LASSO) [19], arboles de decisión (como CART [136], Random Forest importance [125], [137]).

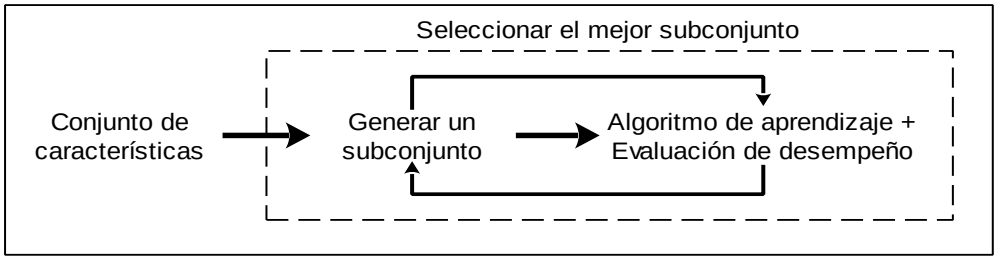


Figura 13. Funcionamiento del método Embedded. [118], [132], [133]

2.7 ANTECEDENTES EN LA UTILIZACIÓN DE MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS

Muchos investigadores han evaluado y comparado diferentes métodos de selección de características, para elegir, clasificar y eliminar las características irrelevantes. Con el propósito de mejorar los resultados en la etapa de clasificación. A continuación, se presentan y examinan algunos trabajos previos sobre este tema.

Kou *et al.* [138] evaluaron varios métodos de selección de características, entre ellos Infromation Gain, Gain Ratio, Gini Index [137], [139], Chi-Squared, Mutual Information [126], entre otros, para la clasificación de texto, estos fueron elegidos por sus variaciones de rendimiento. Chaudhary *et al.* [140], en cambio, evaluaron el rendimiento de tres métodos de selección de características: Correlación [123], [124], Gain Ratio y Information Gain, con Naive Bayes [141]–[143] optimizado en un dispositivo móvil. Basado en las siguientes medidas de rendimiento: accuracy, tasa de verdaderos positivos (tnr) y recall [144], [145], y concluyeron que Gain Ratio tuvo un rendimiento comparativamente mejor que los otros dos métodos.

Karimi *et al.* [146] combinaron los métodos de selección de características Information Gain y Symmetrical Uncertainty para seleccionar, clasificar y eliminar las características irrelevantes en el proceso de detección de intrusos. Para la clasificación, utilizaron el algoritmo de Naive Bayes. Llegaron a la conclusión de que esta combinación mejoraba la precisión y reducía los costos. Gao *et al.* [45] utilizaron el clasificador Support Vector Machine (SVM) [75], [147], [148] apoyado por el método Information Gain para filtrar los genes irrelevantes y redundantes. Posteriormente, evaluaron cinco conjuntos de datos de expresión genética del cáncer y seleccionaron algunos genes. Los genes seleccionados sirvieron de base para el clasificador. Los resultados demostraron que, en comparación con otro método de selección de características, la combinación propuesta lograba la mejor precisión de clasificación.

Novaković *et al.* [149] compararon los métodos Information Gain, Gain Ratio, Symmetrical Uncertainty, Relief y Chi-Squared utilizando dos conjuntos de datos reales. La clasificación se realizó mediante cuatro algoritmos de aprendizaje: K Nearest Neighbors [141]–[143], Naive Bayes, árbol de decisión C4.5 [75] y una red neuronal. Llegaron a la conclusión de que, para seleccionar un subconjunto de características que ofrezca la máxima precisión, se recomienda utilizar varios métodos con diferentes percepciones.

Phyu y Oo [150] propusieron un algoritmo de selección de características basado en la perspectiva de la Información Mutua condicional [126]. Estos autores evaluaron la eficacia del algoritmo propuesto comparándolo con otros algoritmos de selección de características como: Information Gain, Symmetrical Uncertainty y ReliefF y utilizaron conjuntos de datos estándares de UC Iravine y Weka. Seguidamente, evaluaron el rendimiento del algoritmo propuesto mediante la precisión en la clasificación de los clasificadores Naive Bayes y J48 [75], así como por el número de características seleccionadas. Los autores llegaron a la conclusión de que, aunque algunos algoritmos pueden reducir aún más el número de características, su precisión en la clasificación no era muy buena. Además, afirmaron que su algoritmo selecciona el menor número posible de características y es más preciso en la clasificación de varios de los conjuntos de datos utilizados.

Dag *et al.* [151] cotejaron tres métodos: Information Gain, Gain Ratio y el algoritmo basado en la correlación aplicado a conjuntos de datos médicos con los clasificadores J48 y C4.5.

Peker *et al.* [152] utilizaron los métodos de selección de características: Redundancia mínima Relevancia máxima [142] y Relief para seleccionar el conjunto de características que mejor se adapte a las necesidades, utilizando los clasificadores Random Forest [78], [148], C4.5, SVM, Naive Bayes y dos tipos de red neuronal. Los autores descubrieron que los mejores resultados se obtenían cuando se utilizaba el subconjunto de características obtenidas del algoritmo Relief con el clasificador de Random Forest.

Finalmente, Parimala y Nallaswamy [153] usaron diferentes métodos de selección de características: Correlation-based Feature Selection [123], [124], Chi-Squared, Information

Gain, Gain Ratio, Mutual Information, Symmetrical Uncertainty y Relief con los clasificadores: árbol de decisión, análisis discriminante lineal [142], Naive Bayes y SVM. Evaluaron las diferentes combinaciones entre estos métodos de selección de características y los clasificadores propuestos. Sus resultados demostraron que la combinación de los métodos utilizados permitía a los clasificadores alcanzar la mayor precisión en la clasificación.

El abordaje de estos trabajos permite aprecia la visión de comparación y combinación de varios métodos de selección de característica, los cuales se basan en criterios diferentes. No solo para mejorar la precisión en la clasificación, sino que también, para asegurar que las características seleccionadas son las que mayor ganancia de información aportan al problema.

2.7.1 Métodos de Selección de Características Seleccionados

Luego del relevamiento de los antecedentes en la utilización de métodos para la selección de características más importantes, se decidió aplicar diferentes métodos con la finalidad de no sesgar la decisión de selección según los resultados de una sola técnica. Por esta razón, se han utilizado métodos basados en diferentes conceptos, los cuales utilizan distintas medidas de evaluación y distintos procedimientos de búsqueda de subconjuntos. De esta manera, es posible obtener varios puntos de vistas y evaluar diversas posibilidades de rangos de importancia para las mismas características.

Los métodos elegidos son de tipo Filter y Embedded ya que, en base a la revisión bibliográfica, dichas aproximaciones son las más adecuadas y utilizadas para el abordaje de problemas similares [132].

2.7.1.1 Information Gain

Information Gain (IG) es un método ampliamente utilizado en la teoría de la información y la Ciencia de Datos. Dicho método mide la información obtenida sobre el atributo clase para cada característica del conjunto de datos.

Dada la entropía [122] de un criterio de impureza en un conjunto de entrenamiento S, se puede definir una medida que refleja la información adicional sobre Y proporcionada por X, la cual representa la cantidad que disminuye la entropía de Y (ecuación 3). Esta medida está dada por:

$$IG(X) = H(Y) - H(Y|X) = H(X) - H(X|Y) \qquad 3$$

dónde $H(Y)$, $H(Y|X)$ y $H(X|Y)$ son la entropía de Y, la entropía condicional de Y dada X y la entropía condicional de X dada Y respectivamente.

IG es una medida simétrica (ver ecuación 3). Es decir, la información obtenida sobre Y después de observar X es igual a la información obtenida alrededor de X después de observar Y. La desventaja de este método es que está sesgado a favor de características con más categorías o valores, incluso cuando no aportan más ganancia [119].

2.7.1.2 GAIN RATIO

Gain Ratio (GR) es una medida asimétrica que se introduce para compensar el sesgo del método IG [127], [151]. GR está dada por:

$$GR = \frac{IG}{H(X)} \qquad 4$$

Como presenta la ecuación 4, a la hora de predecir la variable Y, normalizamos IG dividido por la entropía de X, y viceversa. Debido a esta normalización, GR siempre se encuentran en el rango [0, 1]. Un valor de GR = 1 indica que el conocimiento de X predice completamente Y, mientras que GR = 0 significa que no hay relación entre Y y X. A diferencia de IG, GR favorece a las variables que tienen menos categorías o valores [36], [127], [154].

2.7.1.3 RANDOM FOREST IMPORTANCE

El método Random Forest importance (RFI) encuentra pesos para las características empleando el algoritmo Random Forest. El conjunto de árboles aleatorios [78], evalúa la importancia de una variable X_m para predecir Y sumando las disminuciones ponderadas de impurezas $\mathrm{p(t)}\Delta i(\mathrm{s_t}, \mathrm{t})$ para todos los nodos t donde se usa X_m, promediadas sobre todos los árboles N_T en el bosque:

$$RFI(x_m) = \frac{1}{N_T}\sum_{T}\sum_{t\epsilon T : v(S_t)=X_m} \mathrm{p(t)}\Delta i(\mathrm{s_t}, \mathrm{t}) \qquad 5$$

donde $\mathrm{p(t)}$ es la proporción N_T/N de muestras que alcanzan t y $v(S_t)$ es la variable utilizada en la división $\mathrm{s_t}$. Se puede definir cualquier medida de impureza $i(t)$. La ecuación 5 toma como función de impureza la importancia de la disminución media de impureza (Mean Decrease Impurity - MDI) o más comúnmente conocida como importancia de Gini [137], [139].

2.7.1.4 RELIEF

El algoritmo Relief (R) encuentra el peso de una característica muestreando repetidamente una instancia y considerando el valor de la característica dada para la instancia más cercana de la misma y diferente clase. Luego, evalúa las ponderaciones de los atributos.

$$R_f = P\left(\frac{\text{different value of f}}{\text{class different}}\right) - P\left(\frac{\text{different value of f}}{\text{same class}}\right) \qquad 6$$

Cómo se aprecia en la ecuación 6, el cálculo del peso se basa en la probabilidad de los vecinos más cercanos de dos clases diferentes, con valores diferentes para un elemento y la probabilidad de dos de los más cercanos vecinos de la misma clase, que tengan el mismo valor de la característica. Cuanto mayor sea la diferencia entre estas dos probabilidades, más significativa será la característica. [129], [130].

2.7.1.5 CHI SQUARED

La prueba Chi Squared (ChiS) o también referenciado como X^2, es una prueba estadística que se aplica a grupos de características categóricas para evaluar la probabilidad de correlación o asociación entre ellas utilizando su distribución de frecuencias. El método evalúa el valor de una característica calculando el valor de la estadística ChiS con respecto a la clase (ecuación 7). La hipótesis inicial H_0 es la suposición de que las dos características no están relacionadas, y se prueba mediante la siguiente fórmula cuadrada:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad 7$$

donde O_{ij} es la frecuencia observada y E_{ij} es la frecuencia esperada (teórica), afirmada por la hipótesis nula. Cuanto mayor sea el valor de X^2, mayor será la evidencia en contra de la hipótesis H_0 [138], [155]. Este método es ampliamente utilizado para la selección de características categóricas [138], [149], [153], [155]–[158].

2.8 ANTECEDENTES EN EL ENSAMBLE DE CLASIFICADORES

En la toma de decisiones, la combinación de modelos de clasificación puede ser fundamental, ya que dicha combinación tiene por objeto obtener una solución apropiada para un problema concreto. Individualmente, los métodos de clasificación se basan en conceptos o procedimientos de estimación diferentes. Es lógico intentar aunar las mejores propiedades de cada uno de ellos combinándolos de alguna manera. Por lo tanto, es posible combinar decisiones obtenidas con el mismo o distintos clasificadores de base [159]. Los métodos de combinación son aquellos en los que, dado un conjunto de clasificadores ya entrenados, los resultados se combinan de diferentes maneras para obtener un valor más preciso que el de los clasificadores individuales [160]. Esta integración suele ser más precisa, porque los datos de entrenamiento pueden no proporcionar suficiente información para elegir un mejor clasificador y, en esta situación, la combinación es la mejor opción. Por lo tanto, la combinación puede ser equivalente a árboles de decisión muy complejos [143].

Varios estudios han evaluado la combinación o integración de clasificadores para mejorar el porcentaje de éxito o incluso para no sesgar la decisión sobre los resultados de un solo clasificador [161].

Miao *et al.* [162], por ejemplo, proponen un procedimiento para mejorar la precisión en la identificación de los genes mediante la integración de los clasificadores SVM, Random Forest y Extreme Learning Machines [163], aplicando ReliefF [118] para seleccionar las características más relevantes del conjunto de datos. Luego del entrenamiento y predicción con los tres clasificadores, los autores combinaron los resultados mediante el método de votación mayoritaria [161]. La integración de las predicciones les permitió obtener una mayor precisión que con los clasificadores individuales. De igual manera, Catal y Nengir [47] presentan un modelo para la clasificación de sentimientos mediante la combinación de los clasificadores Naive Bayes y variantes de SVM. Para la integración de las predicciones, los autores utilizaron el método de votación mayoritaria y demostraron que los sistemas de clasificación múltiple mejoran la precisión. Otro trabajo de similares características es el de Pandey y Taruna [164], donde proponen un clasificador integrado utilizando un árbol de decisión J48, un KNN y Aggregating One-Dependence Estimators (AODE) [165], sobre un conjunto de datos de rendimiento académico de estudiantes de ingeniería. En este modelo, cada clasificador individual genera su valor de predicción, los cuales se integran a través del producto de las probabilidades, donde la etiqueta de clase final se encuentra representada por el máximo de una probabilidad posterior. Así también, Yan *et al.* [166] también plantean la integración de los clasificadores Naive Bayes, árbol de decisión ID3 [8], [76] y Maximum Entropy [167], [168] con el esquema de votación mayoritaria para el análisis de dependencia semántica en chino. Cada uno de los tres clasificadores fue entrenado con los mismos datos de entrenamiento. El enfoque propuesto logró una precisión del 86% en la experimentación, lo que es prometedor para el análisis de dependencia semántica en chino según los autores. Ruano-Ordás *et al.* [169] exponen un modelo para determinar automáticamente la actividad biológica de moléculas basadas en 2048 subestructuras químicas (codificadas usando valores binarios) y 84 propiedades fisicoquímicas (codificadas usando valores discretos y continuos). Los autores realizan el proceso en tres etapas: agrupamiento de características, construcción y optimización de híper parámetros de cada clasificador y clasificación. Asimismo, emplean el uso de SVM con núcleo Radial Basis Function (RBF) [170], AdaBag [79] y rpart [88]. Combinaron los resultados individuales de cada clasificación en un único resultado utilizando el método de votación por mayoría. Además, Oliveira *et al.* [171] abordan el problema de la detección de peatones mediante la utilización de un MLP y un SVM. Para combinar las salidas de los clasificadores, utilizaron dos tipos de métodos de fusión: voto mayoritario e integral difusa [172]. Los autores demuestran que la integración permite mejorar el porcentaje de acierto en la clasificación. Nweke *et al.* [148] presentan un relevamiento del uso de sistemas de clasificadores múltiples en el reconocimiento de la actividad humana y monitoreo de la salud. Estos autores también trataron de reducir la incertidumbre y la ambigüedad fusionando los resultados generados por diferentes modelos

de clasificación. Para ello, abordaron diferentes enfoques de diseño y fusión con clasificadores múltiples, como: SVM, árboles de decisión (ID3, J48, C4.5) [8], [76], KNN, red neuronal artificial, Naive Bayes, Random Forest, entre otros.

Inspirados en las ideas anteriores, se propone la utilización de múltiples clasificadores para el estudio de caso.

2.8.1 TÉCNICAS DE APRENDIZAJE SELECCIONADAS

Un paso importante en la predicción es la búsqueda de los mejores clasificadores individuales para el estudio de caso. Se decidió seleccionar varios tipos de clasificadores de aprendizaje supervisado, basados en conceptos de inferencia o estimación diferentes (máquina de vector soporte, árbol de decisión, k vecino más cercano, Naive Bayes y red neuronal), con el propósito de analizar los resultados desde diversas perspectivas y asegurar una clasificación más precisa que de forma individual.

Ahora bien, porque estos tipos de clasificadores y no otros. Pues bueno, una de las razones por la que se ha seleccionado estos clasificadores es porque son ampliamente utilizados para problemas de similares características al que se aborda en esta tesis, permiten trabajar con representaciones de variables categóricas, soportan problemas binarios, toleran conjuntos desbalanceados y son óptimos para conjuntos de datos pequeños [45], [51], [162], [173]–[179].

Así mismo, existe una gran variedad de alternativas y tipos de clasificadores. Como por ejemplo los algoritmos genéticos, pero estos son más óptimos para problemas complejos, tienen un alto coste computacional, dependen de muchos híper parámetros y son muy dependientes del problema. Además, son utilizados para optimizar espacios de búsqueda muy grandes, como por ejemplo: optimización de rutas, de costos, espacios, entre otros.

Seguidamente, se puntualiza el fundamento teórico y algunas particularidades de cada uno de los clasificadores seleccionados.

2.8.1.1 SVM

Support Vector Machine (SVM) es un tipo de máquina de vectores soporte, puede clasificar conjuntos de datos lineales o no lineales con la ayuda de un kernel [5], [180]. Puede usarse para clasificación o regresión [181], [182], su funcionamiento consiste en construir un conjunto de híper planos en un espacio dimensional alto. La separación se mide como la distancia entre los híper planos y se denomina margen funcional [45], [47], [183]. Cuanto mayor es el margen, menor es el error de generalización del clasificador. Como ejemplo de algunos kernel [170] (ecuación 8) encontramos:

$$linear: (x, x')$$
$$polynomial: (\gamma(x, x') + r)^d \qquad 8$$
$$rbf: exp(-\gamma\|x, x'\|^2)$$

La clasificación que busca el algoritmo basado en un núcleo, son funciones en el espacio de características: $f(X) = W^T\emptyset(X)$ para algún vector de peso $w \in F$ [75], [181]. Los métodos de aprendizaje basados en el kernel utilizan un mapeo implícito de los datos de entrada en un espacio de características de alta dimensión, definido por una función del kernel. El vector de entrenamiento x_i se mapea en un espacio de característica dimensional superior y luego el aprendizaje tiene lugar en el espacio de característica.

La principal razón por la que SVM es ampliamente utilizado [42], [45], [47], [155], [181], [184], [185], es porque soporta problemas no lineales. Además, funciona muy bien en espacios de muchas dimensiones. Aunque, es un poco ineficiente para entrenar cuando se tiene muchos ejemplos de entrenamiento.

2.8.1.2 RANDOM FOREST

Random Forest (RF) fue introducido por Leo Breiman [78], es un algoritmo de aprendizaje cada vez más popular basado en árboles de decisión, permite un rápido entrenamiento, un excelente rendimiento y una gran flexibilidad para manejar todo tipo de datos [186], [187]. Entre las principales reglas utilizadas para dividir los datos binarios se encuentra el índice de Gini (ecuación 9):

$$\mu = \sum_{a=1}^{A} p_a(1 - p_a) \qquad 9$$

donde A es la clase objetivo y p_a la proporción de la muestra de la clase. Este índice mide la impureza del nodo y es el más utilizada [173], [186]–[189]. Un pequeño valor de a indica que el nodo contiene predominantemente observaciones de una sola clase, es decir, es un nodo de pureza con buena separación entre las clases [188].

Entre sus ventajas se encuentran el bajo costo computacional, los resultados pueden ser fácilmente interpretados y puede manejar características irrelevantes. Entre sus desventajas podemos mencionar que es propenso al sobreajuste o sobre-entrenamiento, no maneja valores atípicos y suele tener inconvenientes cuando existen demasiadas ramificaciones [152], [173], [187], [189], [190].

2.8.1.3 KNN

K Nearest Neighbors (KNN) es un tipo de aprendizaje basado en instancias o aprendizaje no generalizado [191], [192]. Este método busca en un conjunto D, los k vecinos q más cercanos al objeto p a clasificar en D, y asigna la etiqueta de clase en función a la mayoría de sus vecinos (ecuación 10), con $dist\ (p,q) \leq dist(p,o)$, es decir:

$$KNN_k\ (p) = \{q | \forall q \in D, dist(p,q) \leq dist(p,o)\} \qquad 10$$

Donde $dist(p,o)$ es la distancia entre p y el objeto k-ésimo o. Para contribuir realmente al ajuste, tanto la elección óptima del valor k como la distancia a utilizar para los vecinos más cercanos dependen en gran medida de los datos. Este algoritmo soporta valores numéricos y nominales [141].

KNN es un clasificador muy sencillo. Suele utilizarse como punto de referencia entre clasificadores más complejos como SVM y las redes neuronales. A pesar de su simplicidad, KNN puede superar a los clasificadores más potentes y se usa en una variedad de aplicaciones tales como reconocimientos de patrones, pronósticos económicos, extracción y compresión de datos, detección de intrusos, entre otros. KNN no hace suposiciones explícitas sobre la forma funcional de los datos. Es un algoritmo simple para comprender e interpretar. Cuenta con una alta precisión (aunque puede verse afectado por el ruido o características irrelevantes). Además, es insensible a valores atípicos. Sin embargo, tiene un alto costo computacional y necesita gran cantidad de memoria. [9], [80], [142], [174], [175].

2.8.1.4 NAIVE BAYES

El algoritmo Naive Bayes (NB) se basa en el principio del teorema de Bayes, el cual asume que las características de entrada son independientes entre sí, denominado independencia condicional (ecuación 11). Está dado por:

$$f_i(X) = \prod_{j=1}^{N} P\left(x_j \middle| c_i\right) P(c_i) \qquad 11$$

donde $x_j = (x_1, x_2, \dots, x_N)$ es el vector de característica y c_i, con $i = 1, 2, \dots, N$, indica posibles etiquetas de clase. La fase de entrenamiento consiste en estimar las probabilidades condicionales $P\left(x_j \middle| c_i\right)$ y las probabilidades previas $P(c_i)$ [43]. En esta tesis se aplica una variante denominado *Multinomial Naive Bayes (MNB)*, el cual soporta datos categóricos y es principalmente utilizado para la clasificación de documentos y textos [10], [48], [176], [193], [194] debido a su simplicidad, eficiencia y eficacia. MNB es un modelo muy sencillo, rápido y fácil de implementar.

2.8.1.5 MULTI-LAYER PERCEPTRON

Multi-layer Perceptron (MLP) es ampliamente utilizado debido a su capacidad para usar aplicaciones tanto lineales como no lineales [44], [81], [82], [177], [178], [195]. Consta de una capa de entrada, una o más capas ocultas y una de salida. El número de neuronas en la capa de entrada se corresponde con el número de características y el número de neuronas en la capa de salida es el número de salidas. La conexión entre las neuronas de las distintas capas se calcula empleando pesos (ecuación 12). Su propósito de entrenamiento es encontrar valores adecuados para los pesos de los enlaces entre las neuronas. La función de salida general y la de error están dado por:

$$y_i = f\left(\sum_{i=1}^{N} w_{ji} x_i\right) \qquad 12$$

$$E = \frac{1}{2}\sum_{i} (d_i - y_i)^2$$

donde x_i son los datos de entrada, w_{ji} los valores de peso, $f(\cdot)$ la función de activación, y_i la i-th salida de la red, y d_i la i-th salida esperada [6].

MLP es utilizado comúnmente para resolver problemas de asociación de patrones, segmentación de imágenes, compresión de datos, entre otros [44], [82], [177]–[179], [195].

2.9 EVALUACIÓN DEL DESEMPEÑO DE MODELOS DE APRENDIZAJE AUTOMÁTICO

Una de las etapas principales en todo proceso de análisis de datos es la evaluación de la calidad del modelo, cuantificando mediante técnicas y métricas de rendimiento.

2.9.1 TÉCNICAS DE EVALUACIÓN PARA CLASIFICADORES

En los métodos supervisados, la clasificación depende del entrenamiento que se realice sobre el clasificador, por lo que su efectividad dependerá en gran medida de qué datos del conjunto se utilicen para entrenamiento y cuáles para prueba. El criterio de selección de estos conjuntos de datos y la cantidad de veces que se evalúe el clasificador utilizando distintos criterios, definirá la confiabilidad de la evaluación. Los métodos más utilizados son: Holdout y validación cruzada.

2.9.1.1 HOLDOUT

El método de retención, del inglés *Holdout* [52] es el más sencillo y consiste en dividir la totalidad de casos de un conjunto de datos de manera aleatoria en dos partes: entrenamiento y validación. En ocasiones se necesita un tercer conjunto de datos para ajustar y evaluar determinados aspectos del modelo, denominado prueba. En la opción a) de la Figura 14 se

aprecia el conjunto de datos dividido en dos partes, en este caso el proceso consiste en aprender con el conjunto de entrenamiento y evaluar el rendimiento con el conjunto de prueba. Generalmente se utilizan dos tercios para entrenamiento y un tercio para prueba. La opción b) presenta la división del conjunto en tres partes, creando una partición para prueba del modelo y estimaciones. Por lo general, se suele dividir en 50% para entrenamiento, 30% para validación y 20% para prueba.

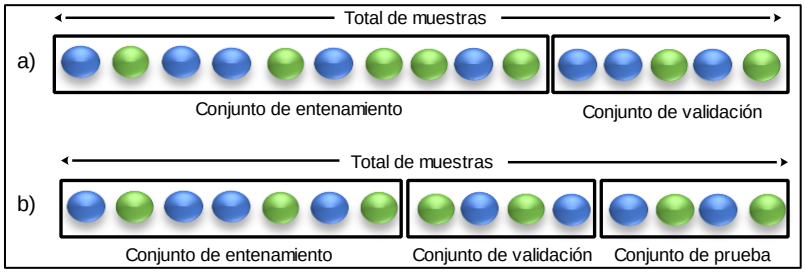


Figura 14. Método Holdout. a) dos particiones, b) tres particiones. [52]

La división del conjunto de datos va a depender en gran medida del número total de muestras y del modelo a entrenar [171], [179], [185], [196], [197]. Para el estudio de caso se dividió los datos de forma aleatoria para preservar la distribución de ambas clases en: 70 % para entrenamiento y 30 % para validación [44], [49], [173], [178], [194], [195], [198], [199], es decir como la opción a) de la Figura 14. Garantizando que todos los casos se encuentren representados en ambos conjuntos.

2.9.1.2 VALIDACIÓN CRUZADA

La validación cruzada o *cross-validation* (por su correspondiente en inglés) es una técnica muy utilizada para evaluar los resultados en la clasificación y garantizar que sean independientes de la partición entre datos de entrenamiento y prueba. Esta técnica consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza cuando el objetivo principal es la predicción y se quiere estimar la precisión de un modelo [200]–[202].

En la Figura 15 se aprecia la aplicación de una validación cruzada con *k* pliegues igual a *n*.

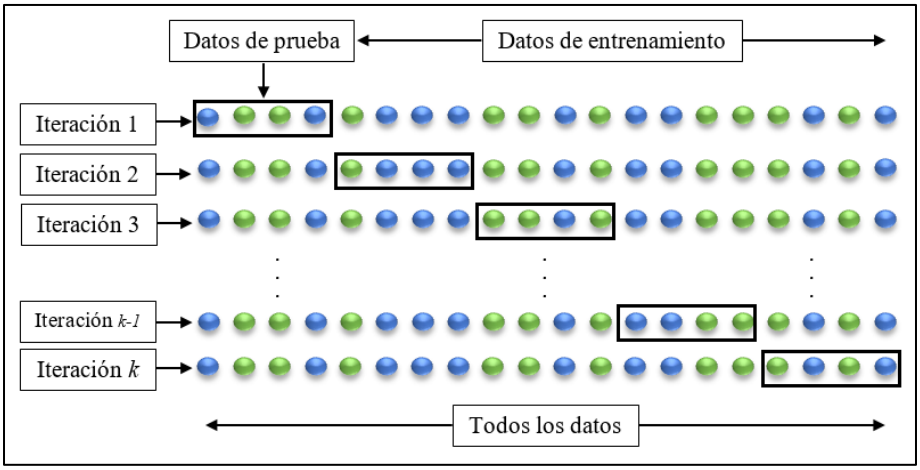


Figura 15. Esquema de validación cruzada con $k=n$. [200]–[202]

Para evaluar los clasificadores, se empleó una validación cruzada de 10 pliegues. Este número de iteraciones es adecuado cuando se trabaja con conjuntos de datos pequeños, está demostrado en los trabajos citados en los apartados *2.7 Antecedentes en la utilización de Métodos de Selección de Características* y *2.8 Antecedentes en el Ensamble de Clasificadores*, entre ellos [45], [114], [140], [146], [149], [150], [152], [153], [162].

2.9.2 Métricas

Para evaluar el rendimiento de clasificadores utilizados en procesos de aprendizaje automático es necesario recurrir a métricas, las cuales permiten medir la precisión de casos correctamente e incorrectamente clasificados.

Las métricas comúnmente utilizadas para evaluar el desempeño de un clasificador son: verdaderos positivos (TP) y negativos (TN), falsos positivos (FP) y negativos (FN), sensibilidad, especificidad, accuracy (acc), error, precisión equilibrada (bac) y área bajo la curva (auc) [144], [203].

TP es el porcentaje de observaciones correctamente clasificadas de la clase objetivo, FN es el porcentaje de observaciones clasificadas erróneamente para esta clase. TN es el porcentaje de observaciones correctamente clasificadas de la clase no objetivo y FP hace referencia a las observaciones erróneamente clasificadas de la clase no objetivo.

Conjuntamente, en base a las medidas TP, FN, FP y TN se puede confeccionar la matriz de confusión para una clasificación binaria. Las etiquetas de clase en el conjunto de entrenamiento pueden tomar solo dos valores posibles, que son llamados positivos o negativos [144]. Esta matriz refleja el número de instancias que se incluyen en cada una de las cuatro categorías [145]. Esto se aprecia en la ecuación 13.

$$Confusion\ Matrix = \begin{bmatrix} TP\ FN \\ FP\ TN \end{bmatrix} \qquad 13$$

Sensibilidad, del inglés *sensitivity* (ecuación 14) es la capacidad del modelo para clasificar correctamente las muestras objetivo. Mientras que la especificidad, del inglés *specificity* (ecuación 15) es la fracción de muestras no objetivo clasificadas como muestras no objetivo por el modelo. La métrica accuracy (ecuación 16) es la proporción total de instancias correctamente clasificadas para ambas clases. El error (ecuación 17) permite medir la proporción total de instancias incorrectamente clasificadas para ambas clases.

Exactitud equilibrada (bac) es la media con respecto a TP y TN (ecuación 18), mientras que el área bajo la curva (auc) es el gráfico que resulta de comparar TP y FP para muchos umbrales diferentes, cuanto más cerca esté esta curva de la esquina superior izquierda, mejor será el rendimiento del clasificador, es decir maximiza los TP mientras se minimiza los FP (ecuación 19).

$$Sensibilidad = \frac{TP}{TP+FN} \qquad 14$$

$$Especificidad = \frac{TN}{TN+FP} \qquad 15$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad 16$$

$$Error = \frac{FP+FN}{TP+TN+FP+FN} \qquad 17$$

$$bac = \frac{TP+TN}{2} \qquad 18$$

$$auc = 1 - Specificity = \frac{FP}{TN+FP} \qquad 19$$

Debido a que la experimentación del procedimiento es realizado sobre un conjunto de datos desbalanceado, se buscó validar específicamente las métricas de rendimiento **TN** y **accuracy**, ya que son medidas de referencia para tratar conjunto de datos con esta característica [41,51-53]. Conjuntamente, se tuvo en cuenta las métricas: sensibilidad, especificidad, error, precisión equilibrada y área bajo la curva.

2.9.3 CONJUNTOS DE VALIDACIÓN

Otra alternativa a la hora de evaluar el rendimiento de un procedimiento de aprendizaje automático, es utilizar conjuntos de datos artificiales u obtenidos de repositorios.

2.9.3.1 CONJUNTOS ARTIFICIALES

Para generar conjuntos de datos artificiales, podemos recurrir al algoritmo SMOTE [204], el cual genera nuevas tuplas artificiales para equilibrar la muestra de datos basado en la regla del vecino más cercano. En el que, para clasificar una nueva instancia, se realiza un cálculo (ecuación 20) de la distancia entre cada atributo de la nueva instancia y el resto de instancias del conjunto de datos y se asocia a la clase de la instancia más cercana. Por lo tanto, dado x_i, $\bar{x} \in N_{min}$ este algoritmo puede ser descrito como:

$$x_{syn} = x_i + (\bar{x} - x_i) \; . \; \times rand(0,1) \qquad 20$$

Aquí, x_i es la muestra de clase minoritaria que se debe sobremuestrear, $\bar{x}$ es otra muestra minoritaria que generalmente es seleccionada de N_{min} cerca de x_i, la expresión . × representa la multiplicación por elemento, y $rand(0,1)$ indica un número aleatorio en el intervalo (0,1).

Este método es ampliamente utilizado para balancear conjuntos de datos [27], [205]–[208], tiene la ventaja de no perder información pero puede repetir muestras con ruido. Es necesario proporcionar como entrada del método: número de muestras de la clase minoritaria (*T)*; cantidad de ejemplos SMOTE a generar (*N*%); número de vecinos más cercanos a considerar (*k)*, y como salida devuelve (*N*/100) * *T* muestras sintéticas de la clase minoritaria [204].

2.9.3.2 REPOSITORIOS DE DATOS

Existen varios repositorios disponibles de acceso abierto para descargar conjuntos de datos en función de la tarea objetivo (clasificación, regresión, clustering u otra), el tipo de los atributos (categóricos y/o numéricos), la naturaleza de los datos (multivariados, series temporales, textuales, etc.), el área de conocimiento y los aspectos relativos al propio conjunto de datos, como su tamaño (número de elementos y dimensionalidad) y su formato (tabular, texto, imagines u otro).

Podemos mencionar a:

- *Kaggle*[1]: es una plataforma de Ciencia de Datos y de aprendizaje automático que cuenta con más de 56 mil conjuntos de datos de todos los tipos y tamaños diferentes y de una gran amplitud de dominios.

[1] Kaggle. Disponible en https://www.kaggle.com/datasets. (Consultado el 09/10/2020).

- *OpenML*[2]: cuenta con más de 20 mil conjuntos de datos de varios dominios. El proyecto Open Machine Learning es un movimiento inclusivo para construir un ecosistema en línea abierto y organizado para el aprendizaje automático.
- *UCI Machine Learning Repository*[3]: es un espacio con más de 550 conjuntos de datos. El repositorio fue creado en 1987 por David Aha y otros estudiantes graduados de la UC Irvine. Desde entonces, ha sido ampliamente utilizado por estudiantes, docentes e investigadores de todo el mundo como fuente primaria de conjuntos de datos de aprendizaje automático, para evaluar y comparar algoritmos.

Para validar el procedimiento propuesto de aprendizaje supervisado, se emplea un conjunto artificial generado con el algoritmo SMOTE y otros dos conjuntos de datos obtenidos de los repositorios *Kaggle* y *OpenML*.

2.9.4 Rendimiento de clasificación por Expertos Humanos

El rendimiento a nivel humano [209]–[213] permite estimar una tasa de error óptima y corroborar el funcionamiento de un sistema de clasificación. Para evaluar el rendimiento del enfoque propuesto en esta tesis sobre el dominio de conocimiento de los implantes dentales, se realizó una comparación de los resultados obtenidos con la opinión y clasificación de expertos humanos. Estos fueron seleccionados del *"Registro de Profesionales que practican Cirugía Buco maxilofacial, Implantología, Periodoncia y Manipulación de Tejidos"* del Colegio de Odontólogos de la provincia de Misiones, Argentina.

2.10 Herramientas Software para Proyectos de Ciencia de Datos

Los lenguajes de programación más utilizados para implementar procesos de Ciencia de Datos son R y Python [214].

2.10.1 R

R es un lenguaje de programación de código abierto, creado en 1995 por Ross Ihaka y Robert Gentleman para mejorar las prestaciones en visualización y análisis de datos. Hoy en día muchos de los estudiantes, profesionales o investigadores que utilizan R provienen del área de la estadística. Es una herramienta muy potente para la comprensión de información

[2] OpenML. Disponible en https://www.openml.org/search?type=data. (Consultado el 09/10/2020).
[3] UCI Machine Learning Repository. Disponible en https://archive.ics.uci.edu/ml/datasets.php. (Consultado el 09/10/2020).

compleja, cuenta con una excelente gama de paquetes de código abierto y de alta calidad. La visualización de datos es una fortaleza clave con el uso de bibliotecas como ggplot2[4].

R también cuenta con una librería para realiza aprendizaje automático, llamado mlr[5]. Esta librería dispone de una gran cantidad de técnicas para clasificación y regresión. También, incluye una extensión para el análisis de supervivencia, agrupación y remuestreo (incluido la validación cruzada, el bootstrap y el sub-muestreo). Conjuntamente, permite el ajuste de híper parámetros con técnicas de optimización para problemas de una o varias clases y contempla distintos tipos de métodos para la selección de características.

La implementación del procedimiento de selección de características propuesto en esta tesis, se realiza con la herramienta R junto al paquete mlr.

2.10.2 PYTHON

Python[6] es un lenguaje de programación muy popular, potente, interactivo, orientado a objetos y de propósito general. Fue diseñado alrededor del 1990 por Guido Van Rossum. Python es gratuito y puede ser utilizado con fines comerciales. Incluye una biblioteca de extensiones estándares para realizar operaciones que van desde manipulaciones de cadenas y expresiones regulares, hasta gráficos de interfaz de usuario. Los paquetes pandas[7], scikit-learn[8] y Tensorflow[9] hacen de Python una opción sólida para aplicaciones de aprendizaje automático.

A la hora de desarrollar procesos de aprendizaje en proyectos de Ciencia de Datos, Python dispone del módulo scikit-learn, el cual integra una amplia gama de algoritmos de aprendizaje automático para resolver problemas de aprendizaje supervisado y no supervisado [215]. Es fácil de utilizar, rápido y posee excelente documentación. Cuenta con dependencias mínimas y se distribuye bajo la licencia de Berkeley Software Distribution, para favorecer su

[4] ggplot2. Disponible en https://ggplot2.tidyverse.org/. (Consultado el 09/10/2020).
[5] mlr. Disponible en https://mlr.mlr-org.com/. (Consultado el 09/10/2020).
[6] Python. Disponible en http://www.python.org. (Consultado el 09/10/2020).
[7] Pandas. Es una herramienta de código abierto para el análisis y manipulación de datos. Es rápida, potente, flexible y fácil de usar, diseñada sobre el lenguaje de programación Python. Disponible en https://pandas.pydata.org/. (Consultado el 09/10/2020).
[8] Scikit-learn. Disponible en https://scikit-learn.org/stable/. (Consultado el 09/10/2020).
[9] Tensorflow. Es una plataforma de código abierto para el aprendizaje automático. Cuenta con herramientas, bibliotecas y recursos para impulsar un proceso de aprendizaje automático innovador. Disponible en https://www.tensorflow.org/. (Consultado el 09/10/2020).

utilización, tanto en el ámbito académico como en lo comercial. Sickit-learn está diseñada para interactuar con las bibliotecas numéricas y científicas NumPy[10] y SciPy[11].

La implementación del procedimiento de clasificación propuesto en esta tesis, se realiza con la herramienta Pyhton y la librería sickit-learn.

2.11 BIOMATERIALES

Debido a que en esta tesis se trabaja sobre un conjunto de datos relacionado a implantes dentales, a continuación se abordan conceptos necesarios y relacionados a este tema.

Los biomateriales se pueden definir como materiales biológicos comunes tales como piel, madera, o cualquier elemento que remplace la función de los tejidos o de los órganos vivos [216]. En otros términos, un biomaterial es una sustancia farmacológicamente inerte diseñada para ser implantada o incorporada dentro del sistema vivo [217].

Los biomateriales se implantan con el objeto de remplazar y/o restaurar tejidos vivientes y sus funciones, lo que implica que están expuestos de modo temporal o permanente a fluidos del cuerpo, aunque en realidad pueden estar localizados fuera del propio cuerpo, incluyéndose en esta categoría a la mayor parte de los materiales dentales que tradicionalmente han sido tratados por separado [218]–[220].

Debido a que los biomateriales restauran funciones de tejidos vivos y órganos en el cuerpo, es esencial entender las relaciones existentes entre las propiedades, funciones y estructuras de los materiales biológicos, por lo que son estudiados bajo tres aspectos fundamentales: materiales biológicos, materiales de implante y la interacción existente entre ellos dentro del cuerpo. Dispositivos como miembros artificiales, amplificadores de sonido para oído y prótesis faciales externas, no son considerados como implantes [220]–[222].

Los requisitos que debe cumplir un biomaterial son:

- Ser biocompatible;
- No ser tóxico, ni carcinógeno;
- Ser químicamente estable e inerte;
- Tener una resistencia mecánica adecuada;
- Tener un tiempo de fatiga adecuado;
- Tener densidad y peso adecuados;
- Tener un diseño de ingeniería perfecto;

[10] NumPy. Es un paquete de código abierto para la computación científica y la manipulación de datos en Python. Permite trabajar con matrices y matrices multidimensionales. Disponible en https://numpy.org/. (Consultado el 09/10/2020).

[11] SciPy. Es un ecosistema de software de código abierto basado en Python para matemáticas, ciencias e ingeniería. Disponible en https://www.scipy.org/. (Consultado el 09/10/2020).

- Ser relativamente barato y reproducible.

Los usos quirúrgicos de los biomateriales son múltiples, por ejemplo, para implantes permanentes: a) en el sistema esquelético muscular, para uniones en las extremidades superiores e inferiores (hombros, dedos, rodillas, caderas, etc.) o como miembros artificiales permanentes; b) en el sistema cardiovascular, corazón (válvula, pared, marcapasos, corazón entero), arterias y venas; c) en el sistema respiratorio, en laringe, tráquea y bronquios, diafragma, pulmones y caja torácica; d) en sistema digestivo: esófago, conductos biliares e hígado; e) en sistema genitourinario, en riñones, uréter, uretra, vejiga; f) en sistema nervioso, en marcapasos; g) en los sentidos: lentes y prótesis de córneas, oídos y marcapasos caróticos; h) otras aplicaciones se encuentran por ejemplo en hernias, tendones y adhesión visceral; i) implantes cosméticos maxilofaciales (nariz, oreja, maxilar, mandíbula, dientes), pechos, testículos, penes, etc. [216]–[224].

2.11.1 OSEOINTEGRACIÓN

Según la bibliografía [95,101,103–114], el éxito de un biomaterial o de un implante depende de tres factores principales:

1. Características o propiedades del implante;
2. Condiciones de salud del receptor;
3. Habilidad del cirujano.

Este último factor es difícil de cuantificar, pero en esta tesis buscamos abordarlo a través de una variable que lo represente en el proceso de extracción de conocimiento.

La oseointegración es el proceso de integración de los tejidos (duros y blandos) con la superficie del implante. También se lo puede definir como, una conexión íntima, directa, funcional y mantenida en el tiempo, entre el hueso y el implante sometido o no, a carga. Este proceso suele durar entre cuatro y seis meses posteriores a la intervención. Una vez trascurrido el período de oseointegración, se procede a la confección de las prótesis o coronas de porcelana definitivas, que cumplen una función estética y funcional a los dientes naturales. [235]–[239].

El cumplimiento de los tres factores mencionados en conjunto con el estado de salud y cuidado por parte del paciente, llevan a que el proceso de oseointegración sea exitoso. En algunos casos puede ocurrir que un implante dental no se integre en un primer intento, con la suficiente fuerza como para resistir la masticación.

2.11.2 IMPLANTES DENTALES

Un implante dental es un tornillo metálico que se adhiere directamente al hueso de la mandíbula o del maxilar, sobre el que se fija una corona que reemplaza la pieza perdida. Están fabricado con materiales biocompatibles, como el Titanio (Ti) o el Circonio (Zr). El Ti es el más utilizado en Argentina, así como en otras partes del mundo, debido a su biocompatibilidad, resistencia y a que el óxido de Ti forma una capa inerte y altamente estable sobre la superficie del implante [236], [239].

En la Figura 16 se aprecia una pequeña comparación entre una pieza dentaria natural y un implante dental.

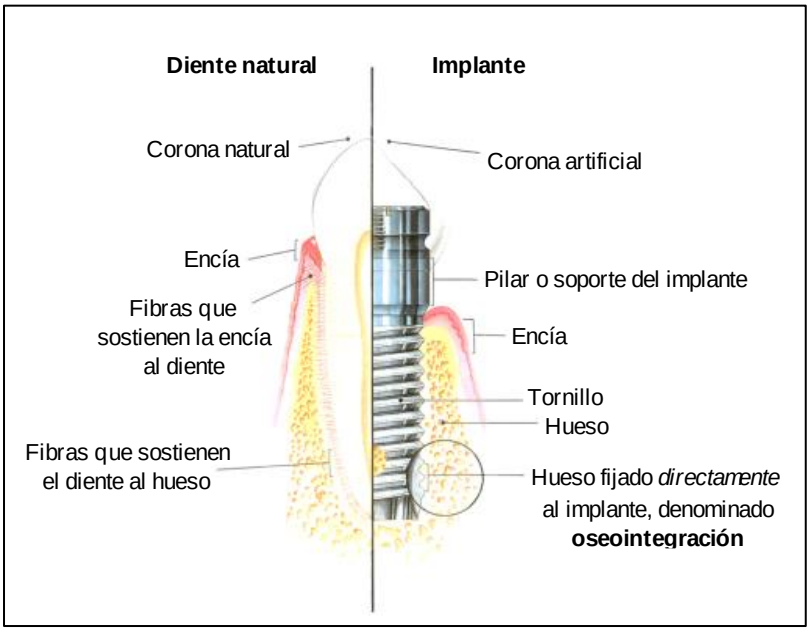


Figura 16. Diente natural vs implante dental. [240]

Como se mencionó el Ti es el material más utilizado tradicionalmente para los implantes dentales. Se trata de un metal ligero, fácilmente moldeable y de una extrema dureza. Su alta biocompatibilidad hace que el organismo difícilmente lo rechace.

En función de la pureza del Ti y de sus aleaciones, el Ti se clasifica en grados. El Ti Comercialmente Puro (CpTi) abarca los grados del 1 al 4 y contienen cantidades de este material superiores al 98%. Estos grados también se denominan grados sin aleación o grados de Ti puro. El Ti con aleación abarca los grados 5 y superiores. El CpTi se utiliza para aplicaciones que requieran una alta resistencia a la corrosión y las aleaciones se utilizan en la fabricación de productos que necesiten una alta resistencia mecánica, resistencia a la corrosión y a la temperatura. Comúnmente, el Ti grado 4 es utilizado para la fabricación de implantes dentales. Los grados de Ti puro ascienden a medida que aumenta la concentración de oxígeno. En la Tabla 2, podemos ver cómo al aumentar la concentración de oxígeno se

obtienen mejoras significativas en la tensión de rotura. La tensión es media en MPa (Megapascales) [225], [231], [241].

Tabla 2. Grados de Ti comercialmente puro.

Grado	Hierro (*Fe*)	Oxígeno (*O*)	Hidrógeno (*H*)	Carbono (*C*)	Nitrógeno (*N*)	Tensión de Rotura
Grado 1	0,2%	0,18%	0,015%	0,08%	0,03%	240MPa
Grado 2	0,3%	0,25%	0,015%	0,08%	0,03%	345MPa
Grado 3	0,3%	0,35%	0,015%	0,08%	0,05%	450MPa
Grado 4	0,3%	0,40%	0,015%	0,08%	0,05%	550MPa

El Zr evita la formación de placa bacteriana a su alrededor, resiste muy bien la corrosión de los ácidos y no provoca problemas por cambios de temperatura. Estos implantes se utilizan en pacientes alérgicos al Ti o que quieren una extremada estética. Su color blanco es muy similar al de las piezas dentales originales y tiene una extraordinaria durabilidad siempre que se mantengan unas condiciones de higiene óptimas, hasta el punto de que muchos dentistas ofrecen una garantía de por vida con estos implantes. Pero tiene la desventaja de que es mucho más costoso que los implantes de Ti [242]–[248].

Por lo tanto, a la hora de evaluar un implante dental se debe tener en cuenta su caracterización superficial, es decir la morfología o topografía, rugosidad, material, así como sus propiedades mecánicas [3], [225], [237]–[239].

2.11.2.1 TIPOS DE IMPLANTES DENTALES

Los implantes dentales pueden ser tipificados de acuerdo a los siguientes criterios [225], [227], [230], [232], [236], [237], [239], [245]:

Según el sitio de localización:

- Endo-óseos u osteointegrados: este tipo de implante es el más utilizado. Se colocan quirúrgicamente en el hueso maxilar o de la mandíbula. Pueden tener forma cilíndrica o cónica. Se utiliza generalmente como una alternativa para los pacientes con puentes o prótesis dentales que son extraíbles.
- Subperiósticos o yuxta-óseos: este tipo de implante consisten en un marco de metal que se coloca en el hueso de la mandíbula justo por debajo del tejido de las encías. Se utilizan para pacientes que no pueden usar las dentaduras convencionales y que tienen una altura ósea mínima.

Según el tipo de conexión (Figura 17):

- Externa: significa que el pilar se conecta con el implante externamente a través de un tornillo. El problema que presentan este tipo de implante es que en algunas ocasiones el tornillo se afloja debido a la masticación o apretamiento dental que se producen (conocido como fuerzas oclusales), incluso puede llegar a deformarse o romperse.
- Interna: implica que el pilar penetra en el interior del implante, aumentando así la resistencia a las fuerzas oclusales. De esta manera se consigue disminuir la posibilidad de que el tornillo de conexión llegue a aflojarse, resultando prácticamente imposible su ruptura o deformidad.

(a) (b)

Figura 17. Tipo de conexión de los implantes dentales. (a) Conexión externa -He-, (b) conexión interna -Hi-.

Según su composición o recubrimiento:

- Cerámicos (fibra de vidrio, alúmina, aluminio cálcico y fosfato tricálcico).
- Carbono (puede ser pirolítico o vítreo).
- Poliméricos (incluye Polimetilmetacrilato, politetrafluoretileno y fibras de carbono).
- Metales (entre los más comunes se encuentran el Ti y sus aleaciones).

Según su diseño (Figura 18):

- Cilindro: también denominados de paredes paralelas o rectas. Este tipo de implante es utilizado mayormente en pacientes con buena calidad ósea, debido a su capacidad para adherirse.
- Cónico: también denominados radiculares o anatómicos, son los más utilizados actualmente. Este tipo de implante dental tiene una base con forma de cono, incrementando su grosor gradualmente en dirección a la corona. Son utilizados cuando la extracción o pérdida de una pieza dental es muy reciente, especialmente en huesos poco densos, gracias a la adherencia que su forma permite con la base.

(a) (b)

Figura 18. Diseño de los implantes dentales. (a) Implante cilíndrico, (b) Implante cónico.

2.11.3 CALIDAD ÓSEA

Otro criterio muy importante a la hora de elegir un tipo de implante, es la calidad ósea del paciente. Existen múltiples clasificaciones de la calidad y cantidad de hueso remanente en zonas del maxilar y de la mandíbula, la más utilizada es la de Misch [249].

En la Figura 19 se muestran los cuatro tipos de densidades óseas localizadas en las áreas del maxilar y de la mandíbula según Misch. El hueso Tipo I es el hueso cortical principalmente denso. El hueso Tipo II presenta hueso cortical denso a poroso, espeso en la cresta y trabecular denso. El hueso Tipo III tiene una cortical fina porosa y un hueso trabecular fino. Por último, el hueso Tipo IV casi no presenta hueso cortical, la mayor parte de volumen óseo se compone de hueso trabecular poroso.

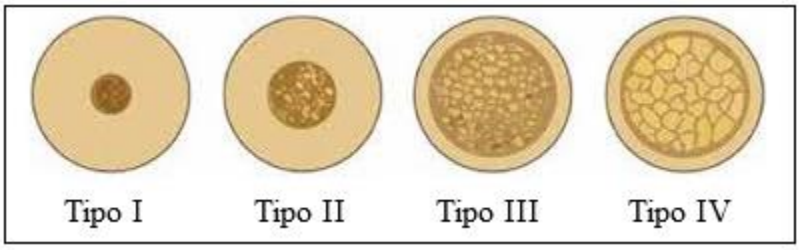


Figura 19. Tipos de hueso de acuerdo a la clasificación de Mish. [249]

3. MATERIALES Y MÉTODOS

A continuación, se describe el planteo realizado para cada una de las fases y tareas que comprende la implementación de la metodología CRISP-DM. Como se mencionó en el apartado *2.5.1 CRISP-DM*, esta metodología estructura el ciclo de vida de un proyecto de Ciencia de Datos en seis fases (Figura 4) y cada fase comprende una serie de tareas (Tabla 1). Estas fases pueden interactuar entre sí de forma iterativa durante todo el desarrollo del proyecto.

3.1 COMPRENSIÓN DEL PROBLEMA

3.1.1 DETERMINACIÓN DE LOS OBJETIVOS

Existen trabajos [2]–[4], [7], [11] que aplican métodos estadísticos o técnicas de aprendizaje automático a conjuntos de datos de implantes dentales, pero no centran su atención al biomaterial, es por esto que en esta tesis se busca abordar conjuntamente el estudio de las características de los implantes dentales utilizados por los especialistas implantólogos.

La carencia de un registro digital provincial o nacional de implantes dentales, que contenga datos sobre enfermedades sistémicas, condiciones del paciente a la hora de la intervención, características del implante utilizado, datos del procedimiento de la fase quirúrgica y datos del seguimiento postoperatorio, hace dificultosa la tarea de análisis y extracción de conocimiento desconocido u oculto sobre patrones que podrían llegar a influir en el proceso de oseointegración de un implante. Un registro digital con estas características, podría aportar conocimiento valioso para los especialistas implantólogos sobre la relación implante dental / condición del paciente.

De aquí, surge la necesidad de crear un registro automatizado que reúna las variables que representan el proceso de colocación de un implante dental, con el objetivo de identificar los factores que contribuyen al fracaso de los implantes dentales colocados en la Provincia de Misiones, Argentina a través de la aplicación de técnicas de Ciencia de Datos, y el diseño de un procedimiento con métodos híbridos.

Por lo tanto, el objetivo es obtener conocimiento del conjunto de datos confeccionado a partir de los datos obtenidos de pacientes que se han sometido a la colocación de implantes dentales en la Provincia de Misiones, especialmente identificar factores que contribuyan al fracaso de estos implantes dentales colocados y determinar las condiciones óptimas que debe tener el paciente y el implante dental utilizado por el profesional implantólogo. Complementariamente, estudiar las propiedades mecánicas, químicas y físicas de los biomateriales utilizados en la implantología dental, para una mejor comprensión de la funcionalidad y resistencia de los mismos.

3.1.2 Evaluación de la situación

Luego, de un primer acercamiento y evaluación del estudio de caso, se determinó que no se disponía de todos los recursos necesarios para la ejecución del proyecto de Ciencia de Datos. Debido a que en la provincia de Misiones no existía un registro digital y estandarizado sobre casos de implantes dentales colocados. Por lo que se planifico la construcción del conjunto de datos necesario, para llevar a cabo el trabajo de investigación. Así mismo, se evalúo las herramientas softwares y hardware necesarias para el diseño e implementación de los proceso necesarios.

3.1.3 Determinación de los objetivos de Ciencia de Datos

Identificar factores que contribuyen al fracaso de los implantes dentales colocados en la provincia de Misiones, Argentina mediante un procedimiento para la selección de las características más importantes y para la clasificación a través de la aplicación de técnicas híbridas de Ciencia de Datos y la validación con expertos humanos (Implantólogos).

3.1.4 Plan del trabajo

Este trabajo se dividió en las siguientes etapas para facilitar su organización:

1. Relevamiento bibliográfico e investigación de la situación actual de la implantología en la provincia.
2. Definición de los expertos en el área.
3. Recolección de datos a través de formularios confeccionados, entrevistas y encuestas.
4. Análisis de la estructura de los datos y de la información del conjunto de datos.
5. Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la aplicación de técnicas de Ciencia de Datos sobre ellos.
6. Elección de las técnicas de modelado y ejecución de las mismas sobre los datos.
7. Análisis de los resultados obtenidos, si fuera necesario repetir el paso 6.
8. Producción de informes con los resultados obtenidos en función de los objetivos.
9. Presentación de los resultados finales.

3.2 Comprensión de los datos

3.2.1 Recolección de datos iniciales

Este trabajo de investigación permitió la creación de un registro digital con datos reales de historias clínicas de pacientes que se han sometido al proceso quirúrgico de colocación de implantes dentales en distintas localidad de la provincia de Misiones, como: Posadas, Apóstoles, Concepción de la Sierra, San Ignacio, San Javier, Eldorado, Iguazú, Montecarlo,

Puerto Rico, San Antonio, El Soberbio, Oberá, Leandro N Alem, Aristóbulo del Valle, Aurora, Jardín América, 25 de Mayo y 2 de Mayo.

Se trabajó en conjunto con varios expertos de diversas áreas de conocimiento (ver Tabla 3), para asegurar la calidad de los datos durante el proceso de recolección de los mismos.

Tabla 3. Expertos

Área de Conocimiento	Experto
Ciencia de Datos	Dr. Horacio Daniel Kuna
Biomateriales	Dra. Alicia Esther Ares
Farmacología	Enf. Máximo Salazar
Gestión	Mgter. Claudia Martínez

Área de Conocimiento	Matrícula[12]	Experto
Patologías bucales	529	Esp. Odont. Padula, Diego Hernán
Rehabilitación compleja en implantología oral	284 1045	Esp. Odont. Acosta, Omar Orlando Esp. Odont. Aimone Romeo, Gabriel A.
Odontólogos	430	Odont. Boychuk, Alejandro
	1047	Odont. Caceres, Guadalupe
	664	Odont. Carrizo, Lucas Miguel
	1181	Odont. Ciszevski, Matías
	288	Odont. Cuba Villalba, Andrés Ignacio
	1092	Odont. Da Silva, Carlos Aníbal
	561	Odont. Dutra, Rafael Ernesto
	923	Odont. Fernández, Gisela Noemí
	142	Odont. Fernández, Néstor Pastor
	1054	Odont. Galmarini, Gabriela María
	1173	Odont. Garrido, German Milton
	724	Odont. Laratro, Federico Ariel
	578	Odont. Leal Chaparro, Gustavo
	674	Odont. Lujan, Federico Fernando
	2229	Odont. Minaberrigaray, María Cecilia
	1036	Odont. Nobs, Federico Ariel
	901	Odont. Padros, Beatriz Marta
	1128	Odont. Quagliozzi, Rubén Oscar
	856	Odont. Schweizer, Eduardo Ariel
	654	Odont. Viñas, Jose Javier
	733	Odont. Von Steiger, Gerardo Andrés

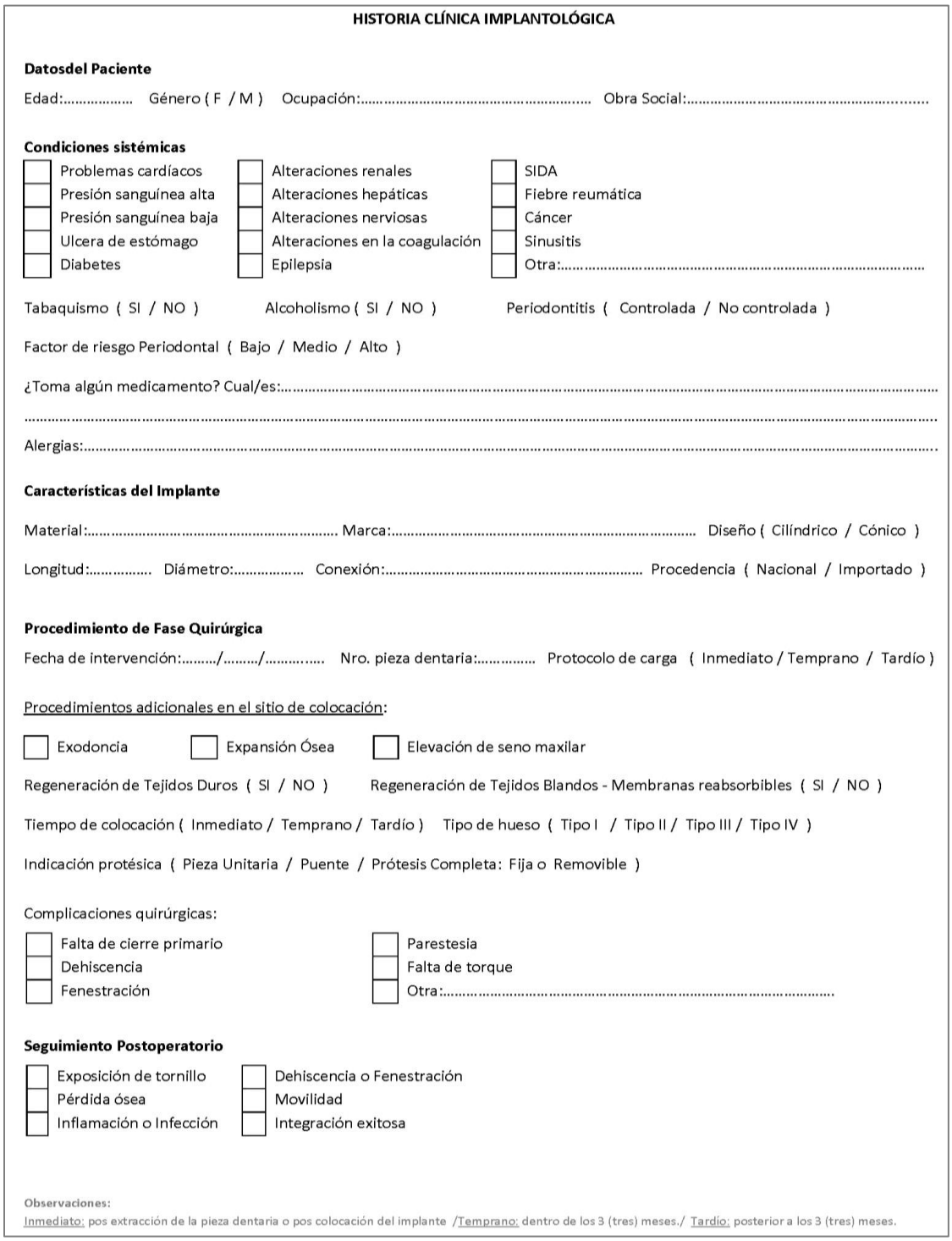
[12] Número de matricula en base al Registro del Colegio de Odontólogos de la Provincia de Misiones, Argentina. Disponible en https://colodmis.org/wp/wp-content/uploads/2020/07/padron-julio-2020.pdf. (Consultado el 20/10/2020).

Primeramente, fue necesario definir junto a los expertos implantólogos cuáles eran los datos de importancia a la hora de considerar la colocación de un implante dental (ver Figura 20). Para ello, se realizaron sucesivas entrevistas y charlas con los especialistas en implantología oral de toda la provincia de Misiones, así como con las autoridades de la mesa directiva del Colegio de Odontólogos de la provincia.

Figura 20. Características relevantes del proceso quirúrgico de colocación de un implante dental.

Para comenzar con el proceso de extracción de conocimiento, los datos fueron obtenidos manualmente a través del diseño y corroboración de un formulario (Figura 21) por parte de los expertos en el área. Cabe remarcar que la recolección de los datos fue realizada entre el año 2016 y 2018.

Dada la gran cantidad de datos a recabar para la realización de este trabajo, se hizo imposible solicitar a cada paciente la autorización para el uso de sus datos de las historias clínicas. Como se puede apreciar en la Figura 20 no se solicitaron datos que identifiquen al paciente ni tampoco se explicita la relación de los datos de cada paciente con el especialista odontólogo. Por esta razón no fue necesario solicitar el consentimiento informado a cada paciente, cumpliendo con el Artículo 5º de la Ley Nacional 25.326 de Protección de Datos Personales, que establece que no será necesario el consentimiento cuando: *"Se trate de listados cuyos datos se limiten a nombre, documento nacional de identidad, identificación tributaria o previsional, ocupación, fecha de nacimiento y domicilio"* [250].



HISTORIA CLÍNICA IMPLANTOLÓGICA

Datosdel Paciente

Edad:..... Género (F / M) Ocupación:..... Obra Social:.....

Condiciones sistémicas

- ☐ Problemas cardíacos
- ☐ Presión sanguínea alta
- ☐ Presión sanguínea baja
- ☐ Ulcera de estómago
- ☐ Diabetes
- ☐ Alteraciones renales
- ☐ Alteraciones hepáticas
- ☐ Alteraciones nerviosas
- ☐ Alteraciones en la coagulación
- ☐ Epilepsia
- ☐ SIDA
- ☐ Fiebre reumática
- ☐ Cáncer
- ☐ Sinusitis
- ☐ Otra:.....

Tabaquismo (SI / NO) Alcoholismo (SI / NO) Periodontitis (Controlada / No controlada)

Factor de riesgo Periodontal (Bajo / Medio / Alto)

¿Toma algún medicamento? Cual/es:.....

.....

Alergias:.....

Características del Implante

Material:..... Marca:..... Diseño (Cilíndrico / Cónico)

Longitud:..... Diámetro:..... Conexión:..... Procedencia (Nacional / Importado)

Procedimiento de Fase Quirúrgica

Fecha de intervención:...../...../..... Nro. pieza dentaria:..... Protocolo de carga (Inmediato / Temprano / Tardío)

Procedimientos adicionales en el sitio de colocación:

☐ Exodoncia ☐ Expansión Ósea ☐ Elevación de seno maxilar

Regeneración de Tejidos Duros (SI / NO) Regeneración de Tejidos Blandos - Membranas reabsorbibles (SI / NO)

Tiempo de colocación (Inmediato / Temprano / Tardío) Tipo de hueso (Tipo I / Tipo II / Tipo III / Tipo IV)

Indicación protésica (Pieza Unitaria / Puente / Prótesis Completa: Fija o Removible)

Complicaciones quirúrgicas:

- ☐ Falta de cierre primario
- ☐ Dehiscencia
- ☐ Fenestración
- ☐ Parestesia
- ☐ Falta de torque
- ☐ Otra:.....

Seguimiento Postoperatorio

- ☐ Exposición de tornillo
- ☐ Pérdida ósea
- ☐ Inflamación o Infección
- ☐ Dehiscencia o Fenestración
- ☐ Movilidad
- ☐ Integración exitosa

Observaciones:

Inmediato: pos extracción de la pieza dentaria o pos colocación del implante /Temprano: dentro de los 3 (tres) meses./ Tardío: posterior a los 3 (tres) meses.

Figura 21. Formulario de recolección de datos.

3.2.2 DESCRIPCIÓN DE LOS DATOS

Los datos del conjunto de datos se agrupan en cuatro dimensiones [251]:

- **Datos del Paciente:** características referidas a los antecedentes y condiciones médicas de los pacientes a la hora de la intervención.
- **Datos del Implante:** características del implante utilizado por el especialista implantólogo.
- **Datos de la Fase Quirúrgica:** características que representan el procesamiento de intervención quirúrgica y mejoramiento del lecho óseo del paciente.
- **Datos del Seguimiento Postoperatorio:** particularidades del resultado del proceso de colocación del implante, es decir si el proceso de oseointegración tejido/implante tuvo éxito o fracasó.

Estas 4 dimensiones se encuentran albergadas en un solo archivo de datos, denominado *"Implantes Dentales.csv*" el cual es el conjunto de trabajo de esta investigación. Está compuesto por un total de 34 características y un atributo clase binario, es decir con dos valores posibles (éxito o fracaso) y 1.165 tuplas o filas, las cuales representan casos de implantes colocados en la provincia de Misiones. Este conjunto, tiene la particularidad de ser un conjunto desbalanceado, lo cual implica que el atributo de clase cuenta con 1.009 casos etiquetados como éxito y 156 como fracaso. Agrupa variables cuantitativas discretas (Edad), cualitativas categóricas (Genero, Tabaquismo, entre otras) y cualitativas ordinales (Riesgo periodontal, Tipo hueso).

En la Tabla 4 se especifican las columnas que componen cada dimensión del conjunto de datos *"Implantes Dentales.csv"*:

Tabla 4. Descripción de los datos.

Dimensión	Columnas
Datos del Paciente	EDAD, GENERO, OCUPACION, O_S, ENFERMEDAD, FUMA, ALCOHOLISMO, PERIODONTITIS, RIESGO_PERIODONTAL, DESDENTADO, INGEST_MED, ALERGIA.
Datos del Implante	MATERIAL, MARCA, DISENIO, LONGITUD, DIAMETRO, CONEXION, PROCEDENCIA.
Datos de la Fase Quirúrgica	FECHA_INTER, LUGAR_PROC, PROF, NRO_DENTAL, PROT_CARGA, EXODONCIA, EXPAN_OSEA, ELEV_SENO_MAXILAR, REG_TEJ_DUROS, REG_TEJ_BLANDOS, TIEM_COLOC, TIPO_HUESO, INDIC_PROTESICA, COMP_QUIRURGICA.
Datos de Seguimiento Postoperatorio	SEGUI_POSTOP, OBSERVACION.

3.2.3 EXPLORACIÓN DE LOS DATOS

En la Tabla 5 se detallan los tipos de datos y valores de cada columna por dimensión:

Tabla 5. Exploración de los datos.

Dimensión	Columna	Tipo de dato	Valores
Datos del Paciente	EDAD	Entero	Entre 17 y 97.
	GENERO	Carácter	F, M.
	OCUPACION	Cadena	Denominación de la ocupación del paciente.
	O_S	Cadena	Razón social de la obra social o "Ninguna".
	ENFERMEDAD	Cadena	Nombre de la enfermedad que padece el paciente o "Ninguna".
	FUMA	Cadena	Si, No.
	ALCOHOLISMO	Cadena	Si, No.
	PERIODONTITIS	Cadena	Si, No.
	RIESGO_PERIODONTAL	Cadena	Bajo, Medio, Alto.
	DESDENTADO	Cadena	Si, No.
	INGEST_MED	Cadena	Nombre del medicamento o "No".
	ALERGIA	Cadena	Penicilina, polen, aspirineta, urobiotic inyectable, no.
Datos del Implante	MATERIAL	Cadena	Titanio
	MARCA	Cadena	B&W, ML, FIA, ODONTIT, BIOMET 3I, NEODENT, TREE-OSS, BIOUNITE, BIOHORIZONS, SMILETECH, STRAUMANN, Q-IMPLANT, NOBEL BIOCARE, ROSTERDENT, BIOCOM, ALPHA BIO, FEDERA, MICROFIT.
	DISENIO	Cadena	Cónico, Cilíndrico.
	LONGITUD	Decimal	Entre 4 mm y 18 mm.
	DIAMETRO	Decimal	Entre 2,5 mm y 5 mm.
	CONEXION	Cadena	Interno, Externo.
	PROCEDENCIA	Cadena	Nacional, Importado.

Dimensión	Columna	Tipo de dato	Valores
Datos de la Fase Quirúrgica	FECHA_INTER	Fecha	Entre 29/11/2003 y 26/10/2017
	LUGAR_PROC	Cadena	Nombre de la localidad de procedencia del paciente.
	PROF	Cadena	Nombre del profesionales que coloco el implante.
	NRO_DENTAL	Cadena	11, 12, 13, 14, 15, 16, 17, 18, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 33, 34, 35, 36, 37, 38, 41, 42, 43, 44, 45, 46, 47, 48.
	PROT_CARGA	Cadena	Inmediato, Temprano, Tardío.
	EXODONCIA	Cadena	Si, No.
	EXPAN_OSEA	Cadena	Si, No.
	ELEV_SENO_MAXILAR	Cadena	Si, No.
	REG_TEJ_DUROS	Cadena	Si, No.
	REG_TEJ_BLANDOS	Cadena	Si, No.
	TIEM_COLOC	Cadena	Inmediato, Temprano, Tardío.
	TIPO_HUESO	Cadena	Tipo I, Tipo II, Tipo III, Tipo IV.
	INDIC_PROTESICA	Cadena	P_U, P, P_C_F, P_C_R.
	COMP_QUIRURGICA	Cadena	Falta de cierre primario, falta de torque, fenestración, parestesia, retiro de resto radicular, ninguna.
Datos de Seguimiento Postoperatorio	SEGUI_POSTOP	Cadena	Éxito, Fracaso.
	OBSERVACIONES	Cadena	Pérdida ósea, movilidad, periimplantitis, exposición de tronillo, mucositis periimplantaria, intoxicación por tabaquismo, ninguna.

3.2.4 VERIFICACIÓN DE LA CALIDAD DE LOS DATOS

Posteriormente de haber realizado la exploración inicial de los datos se puede afirmar que estos son completos. Asimismo, los datos cubren los casos requeridos para la obtención de

los resultados necesarios, y de esta manera poder cumplir con los objetivos del trabajo de investigación.

Los datos fueron verificados y validados por los expertos en el área de implantología. Donde, conjuntamente con el análisis descriptivo se percató de la necesidad de normalizar los datos de las variables, ya que se observaron errores de tipeo. Además, la verificación permitió controlar valores fuera de rango, por lo que no hay riesgo de ruido en el proceso de extracción de conocimiento. En cuanto a valores nulos, no se han encontrado campos con este escenario.

A continuación, se detalla todo el proceso realizado para asegurar la calidad de los datos.

3.3 PREPARACIÓN DE LOS DATOS

3.3.1 SELECCIÓN DE DATOS

En la Tabla 6 se muestra los atributos que se omitieron en las actividades de preparación de datos, debido a que su contenido no requirió modificaciones:

Tabla 6. Selección de datos.

Dimensión	Columnas
Datos del Paciente	GENERO, FUMA, ALCOHOLIMO, DESDENTADO, ALERGIA.
Datos del Implante	DISENIO, CONEXION, PROCEDENCIA.
Datos de la Fase Quirúrgica	PROT_CARGA, EXODONCIA, TIEM_COLOC, TIPO_HUESO, INDIC_PROTESICA, COMP_QUIRURGICA.
Datos de Seguimiento Postoperatorio	SEGUI_POSTOP, OBSERVACIONES.

3.3.2 LIMPIEZA DE LOS DATOS

Sobre el conjunto de datos se aplicaron filtros, para verificar la correcta correspondencia de los valores de cada variable. Conjuntamente, se realizó la búsqueda de campos nulos o vacíos:

```
Datos <- read.table('Implantes Dentales.csv', header = TRUE, sep = ";",
dec = ',')

any(!complete.cases(Datos)) # Detección de filas incompletas

## [1] FALSE
```

Como se mencionó, el análisis descriptivo permitió descubrir la necesidad de normalizar los datos de determinadas variables, como por ejemplo la variable género. Se aprecia en la Figura 22 que aparecen tres posibles valores o tipos de géneros: F, M y f, donde para la percepción del ojo humano F (mayúscula) y f (minúscula) son lo mismo, pero los métodos computacionales lo detectan como un valor posible más de la variable.

Este análisis se ejecutó sobre cada una las variables categóricas. Los resultados arrojados por los filtros se evaluaron junto a los expertos, y se definió conjuntamente si se descarta o simplemente fue un error de tipeo.

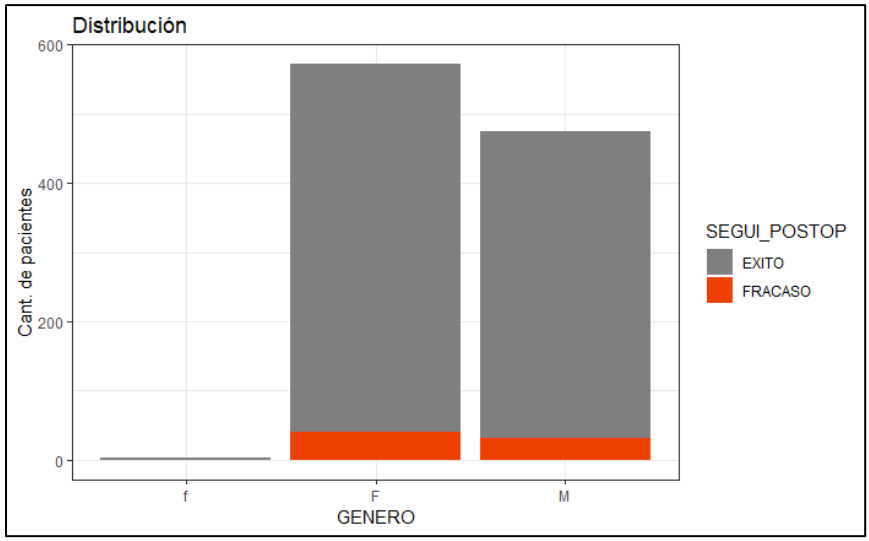


Figura 22. Normalización de la variable GÉNERO.

3.3.3 CONSTRUCCIÓN, INTEGRACIÓN Y FORMATEO DE DATOS

En función de los requerimientos y de los objetivos plateados, se transformaron las siguientes variables:

EDAD en **RANGO_EDAD:** la distribución de la edad (Figura 23) de los pacientes parece ser muy similar entre el grupo de éxitos y fracasos, con dos excepciones: en el rango de edad aproximado de 35 a 40 y de 50 a 60 años, el porcentaje de fracaso es mucho mayor, mientras que, en el extremo opuesto, a partir de los 60 años, la tendencia se iguala, aunque en determinados caso prevalece el fracaso. Mientras, que para las edades inferiores a los 35, el éxito es contundente.

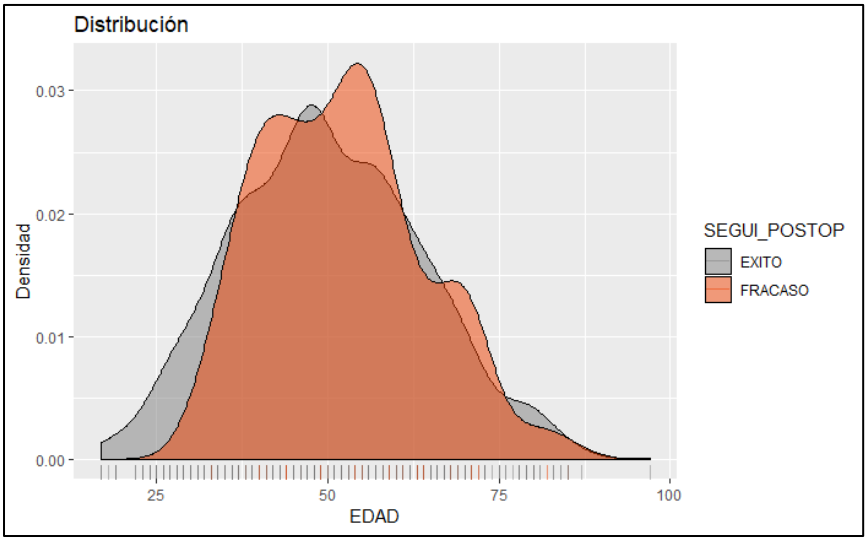


Figura 23. Distribución de la variable EDAD.

Luego de analizar la distribución de esta variable se determinó agrupar las edades en rangos, para que la representación de los datos de la misma sea más significativa, dando origen a la variable RANGO_EDAD. Para reclasificar la variable EDAD, los rangos fueron calculados de forma automática usando un árbol de CART, el cual se entrenó para clasificar el éxito o fracaso usando solo EDAD como variable. De esta manera se lograrón grupos más representativos (ver Figura 24 y Figura 25).

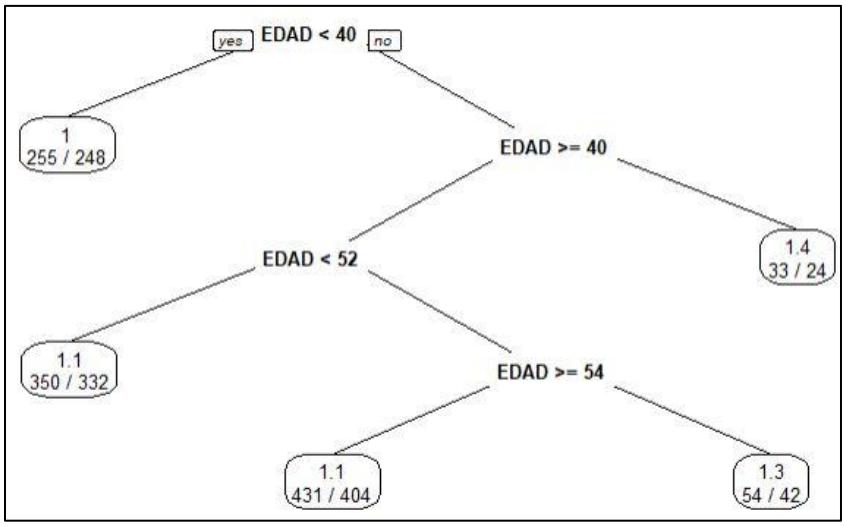


Figura 24. Árbol CART empleando la variable objetivo SEGUI_POSTOP y EDAD, con el método de distancia de poisson.

En base a los resultados arrojados por el árbol de distribución, los rangos de edades más representativos son los que se aprecian en la Figura 25. De esta manera se puede asegurar una adecuada reclasificación del campo EDAD.

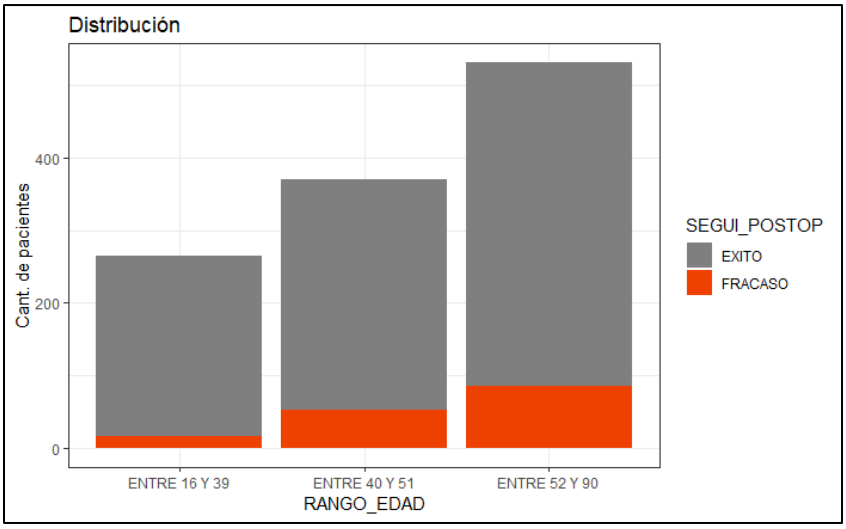


Figura 25. Rango de edades.

Para corroborar esta reclasificación, se procedió a cotejar ambas variables para medir el nivel de significancia o ganancia de información que aporta cada una respecto a la variable clase u objetivo (ver Figura 26).

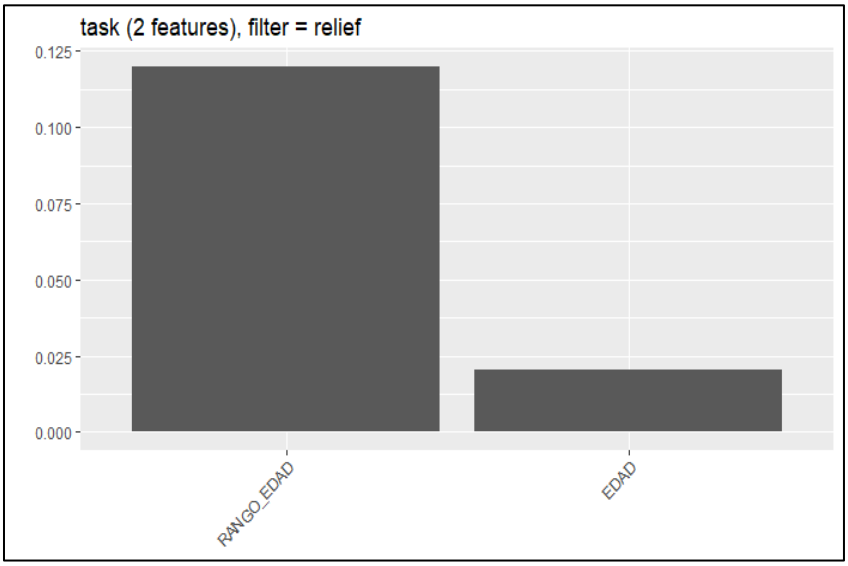


Figura 26. Ganancia de información de las variables EDAD y RANGO_EDAD empleando el método de filtrado Relief.

PROFESION en **OCUPACION:** se procedió a unificar los datos de esta variable en las siguientes categorías:

- DESOCUPADO (desocupado, ama de casa, jubilado).
- ADMINISTRACION (administración).

- PROFESIONAL (medico, abogado, contador, aduanero, escribano, ingeniero, kinesiólogo, veterinario, nutricionista, peluquero, psicólogo, farmacéutico, diseñador gráfico, esteticista, fonoaudióloga, gastronómico, ingeniero civil, locutor, odontólogo, periodista, arquitecto, técnico informático).
- ACADEMICO (docente, preceptora, profesor de educación física, investigador, estudiante)
- TÉCNICO (albañil, electricista, mecánico, agricultor, carpintero, técnico, trabajo rural, gomero).
- EMPLEADO (moza, empleado del banco nación, empleado del poder judicial, camionero, ministro, colectivero, empleado público, empleado de comercio, gerente, gerente de marketing, empleado).
- INDEPENDIENTE (comerciante, empresario, autónomo).

ENFERMEDAD en **ANTECEDENTES:** se procedió a la generación de una nueva variable denominada ANTECEDENTES, esta nueva variable reagrupa y unifica los datos de la variable ENFERMEDAD en siete categorías, estás son:

- PROB_CARD: unifica a todos los pacientes que tienen problemas cardiacos, ya sea de presión alta o presión baja.
- DIABETES: representa a pacientes con problemas en el nivel de azúcar, todos los grados.
- TIROIDES: congrega a pacientes en su mayoría mujeres con trastornos de la glándula tiroides, ya sea hipotiroidismo (trastorno en el cual la glándula tiroides no produce la cantidad suficiente de ciertas hormonas importantes) o hipertiroidismo (trastorno en el cual la glándula tiroides produce demasiada cantidad de hormona tiroxina).
- ARTRITIS: agrupa a los pacientes que sufren de artritis, que es la inflamación de una o más articulaciones que provoca dolor y rigidez, y puede empeorar con la edad.
- COLESTEROL: personifica a pacientes con problemas de colesterol alto o hipercolesterolemia.
- OTROS: representa a todos los pacientes que tienen algún tipo de alteración o inconveniente en cuestiones como: alteraciones nerviosas, ulcera de estómago, asma, gastritis, alteraciones hepáticas, alteraciones renales, anemia e insomnio.
- NO: unifica a los pacientes que no tienen ningún tipo de enfermedad o inconveniente de salud a la hora de la cirugía de colocación del implante dental.

Se integró la variable **PERIODONTITIS** y **RIESGO_PERIODONTAL** en una sola variable. Debido a que ambas representan la misma información. Quedando la variable **PERIO** con los valores:

- BAJO: nivel bajo de afección que tiene el tejido o hueso alrededor de una o varias piezas dentarias del paciente.
- MEDIO: nivel medio de afección que tiene el tejido o hueso alrededor de una o varias piezas dentarias del paciente.
- ALTO: nivel alto de afección que tiene el tejido o hueso alrededor de una o varias piezas dentarias del paciente.
- NO: el paciente no sufre de periodontitis.

INGEST_MED: los datos de la variable INGEST_MED se agrupó en 8 categorías, las cuales se describen a continuación:

- ANTIHIPERTENSIVO: reúne los medicamentos utilizados para el tratamiento de la hipertensión, es decir para normalizar la presión arterial irregular.
- ANTIDIABETICO: agrupa los medicamentos usados para reducir los niveles de glucosa en sangre.
- HORMONAS: representa los medicamentos empleados para restaurar o regularizar la función de la glándula tiroidea, así también son utilizados como anticonceptivos.
- AAA: reúne los medicamentos Analgésicos, Antiinflamatorios y Antipiréticos.
- SUPLEMENTO: concentra a todos aquellos suplementos nutricionales (vitaminas, minerales, etc.) indicados para la prevención o el tratamiento de enfermedades como: anemia, osteoporosis, entre otras.
- ANTIULCEROSO: agrupa a los medicamentos que facilitan la cicatrización de una úlcera.
- OTROS: congrega a los medicamentos ansiolíticos, antibióticos, antiartrósicos, hipolipemiantes, antihistamínicos, hipnóticos y antineoplásicos.
- NO: unifica a los pacientes que no ingieren ningún medicamento a la hora de la intervención quirúrgica.

Para una descripción más detallada de cada categoría o tipo de medicamento ver el apartado *10.1 ANEXO I – INGEST_MED*.

TRAT_SUP: en base a las variables que representan las características del implante (material, marca, longitud, diámetro, diseño, conexión y procedencia), se generó el campo **TRAT_SUP**. Para lograr este campo se realizó una investigación exhaustiva de cada marca (FIA, TREE-OSS, BIOMET 3I, B&W, OSTEOFIT, ML, BIOCOM, ODONTIT, STRAUMANN, BIOUNITE, NEODENT, ROSTERDENT, BIOHORIZONS, Q-IMPLANT, FEDERA, ALPHA-BIO, SMILETECH y NOBEL BIOCARE) con sus características para determinar el tratamiento de superficie del mismo, donde se logró la taxonomía que se aprecia en la Tabla 7.

Tabla 7. Tipificación de los Tratamientos de Superficie de los Implantes Dentales.

Marca del Implante	Nombre de Tratamiento	Descripción de la Técnica	Tratamiento de Superficie						Material de Contacto			Código para la BD
			Blasting	*Simple Tratamiento Ácido*	*Doble Tratamiento Ácido*	*Tratamiento Sol-Gel*	*Tratamiento Electroquímico*	*Tratamiento Térmico*	TiO_x	*Ca*	*P*	
FIA	SLA	Blasting + doble grabado ácido.	x		x				x			T1
ML	SLA	Blasting + doble grabado ácido.	x		x				x			T1
STRAUMANN	SLA	Blasting + doble grabado ácido.	x		x				x			T1
ROSTERDENT	SLA	Blasting + doble grabado ácido.	x		x				x			T1
Q-IMPLANT	SLA	Blasting + doble grabado ácido.	x		x				x			T1
ALPHA-BIO	NANOTEC	Blasting + doble grabado ácido.	x		x				x			T1
ODONTIT	-	Blasting + doble grabado ácido.	x		x				x			T1
TREE-OSS	OXALIFE	Blasting + grabado ácido y tratamiento térmico para una capa aumentada de óxido de titanio.	x	x				x	x			T2
BIOCOM	OXACID	Blasting + oxidación acida y oxidación térmica.	x	x				x	x			T2
BIOMET 3I	OSSEOTITE	Blasting + doble grabado ácido y depósitos de nano cristales de fosfato cálcico.	x		x	x				x	x	T3
B&W	-	Grabado bi-ácido.			x				x			T4
MICROFIT	-	Blasting + grabado ácido.	x	x					x			T5
NEODENT	NEOPOROS	Blasting + grabado ácido.	x	x					x			T5
FEDERA	OSEOMIMETICA	Blasting + grabado ácido.	x	x					x			T5
BIOHORIZONS	RBT	Blasting con fosfato tricálcico + grabado ácido.	x	x					x	x	x	T6
SMILETECH	BIO-CAP	Tratamiento anódico de electro deposición que genera una matriz de Óxido de Titanio enriquecida con Calcio y Fósforo.					x		x	x	x	T7

Marca del Implante	Nombre de Tratamiento	Descripción de la Técnica	Tratamiento de Superficie						Material de Contacto			Código para la BD
			Blasting	*Simple Tratamiento Ácido*	*Doble Tratamiento Ácido*	*Tratamiento Sol-Gel*	*Tratamiento Electroquímico*	*Tratamiento Térmico*	TiO_x	*Ca*	*P*	
BIOUNITE	BIO-CAP	Tratamiento de electrodeposición que genera una matriz superficial de óxido de Titanio enriquecida en calcio y fósforo.					x		x	x	x	T7
NOBEL BIOCARE	TiUnite	Superficie tratada con oxidación anódica, la capa de óxido posee TiO_2 enriquecida con calcio y fósforo por deposición electrolítica.					x		x	x	x	T7

Para una descripción detallada de las categorías o tipificación de los tratamientos de superficie de los implantes dentales ver el apartado *10.2 ANEXO II – TRAT-SUP*.

En la Figura 27 se aprecia una microscopía electrónica de barrido de implantes dentales fabricados mediante distintos métodos: anodización (imágenes a y b) y chorreado con grabado ácido (imágenes c y d) [237]. Debido a que cada implante cuenta con una superficie particular, lo que conlleva a beneficios variados, se decidió tipificar a los mismos por la técnica de fabricación.

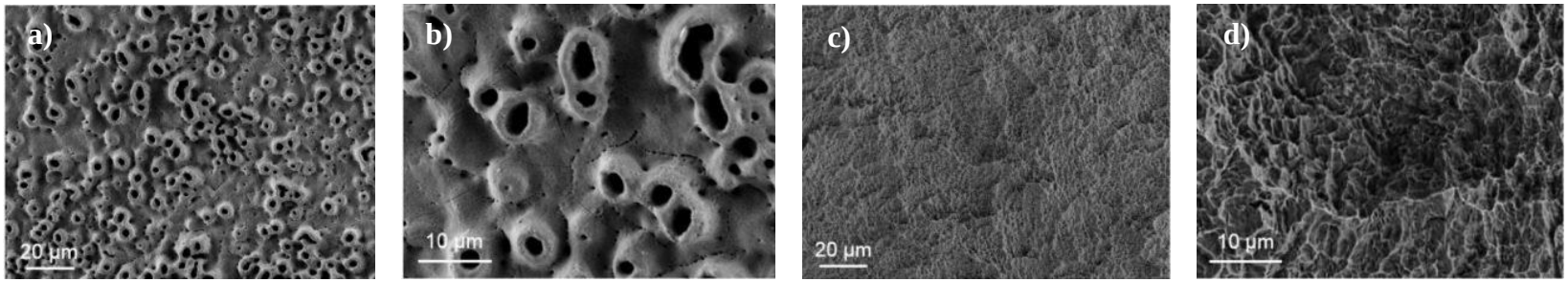


Figura 27. Microscopía electrónica de barrido de implantes dentales. [237]

LONGITUD: se determinó agrupar las medidas de longitud en rangos, para que la representación de los datos de ésta variable sea más significativa. Dando lugar a los siguientes rangos: 4 A 6 MM, 7 A 9 MM, 10 A 12 MM Y 13 A 15 MM.

La Figura 28 muestra la distribución de la nueva variable longitud de los implantes utilizados. Para esta variable prevalece el éxito en la totalidad de los grupos.

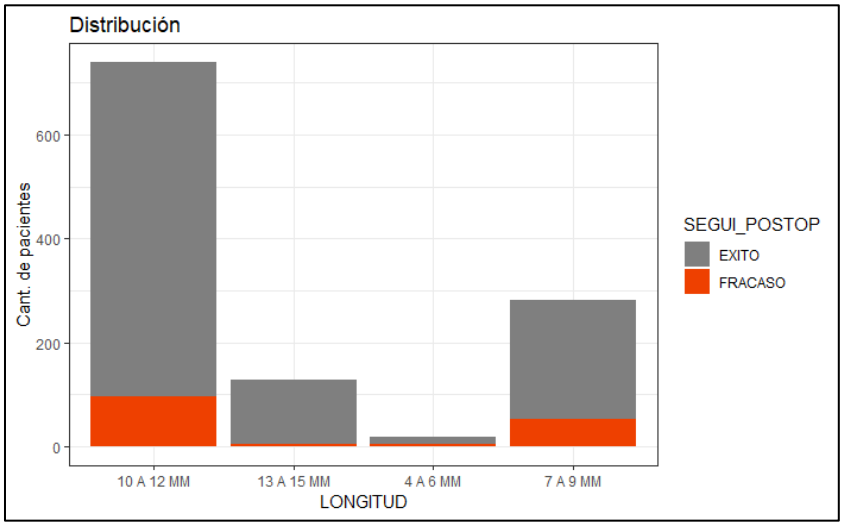


Figura 28. Distribución de variable LONGITUD.

DIAMETRO: se determinó agrupar los valores de diámetros en rangos, para que la representación de los datos de ésta variable sea más significativa. Dando lugar a los siguientes rangos: 2,5 A 3 MM, 3 A 3,5 MM, 3,5 A 4 MM, 4 A 4,5 MM y 4,5 A 5 MM.

La Figura 29 muestra la distribución de la variable diámetro de los implantes utilizados, esta nueva variable se comporta de manera similar a la distribución de las longitudes.

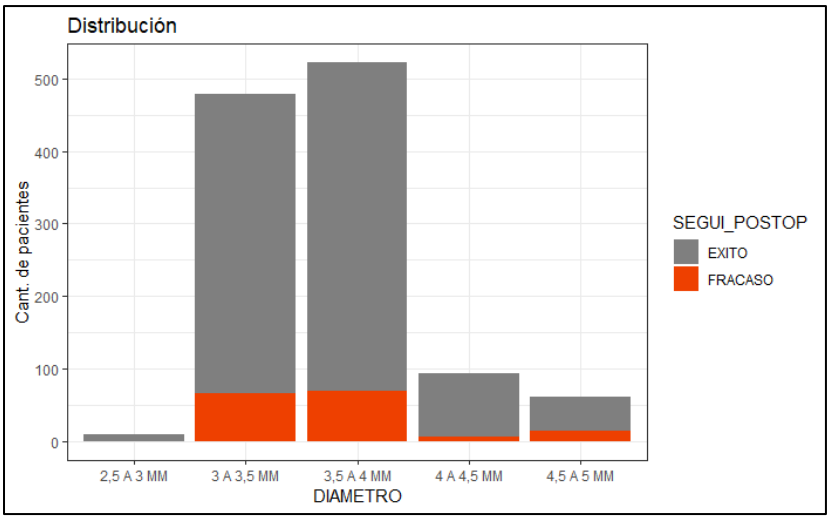


Figura 29. Distribución de variable DIAMETRO.

FECHA_INTER en **ESTACION_ANIO:** se decidió transformar las fechas a estación del año, es decir VERANO (fechas comprendidas entre 21 diciembre y 20 marzo), OTOÑO (fechas comprendidas entre 21 marzo y 20 junio), INVIERNO (fechas comprendidas entre 21 junio hasta 20 septiembre) y PRIMAVERA (fechas comprendidas entre 21 septiembre y 20 diciembre). Esta decisión se tomó junto a los expertos para analizar la posible variación del resultado del proceso de oseointegración en distintas épocas del año.

LUGAR_PROC en **ZONA_PACIENTE:** se agrupó las distintas localidades por las tres zonas divisoras prevalente de la provincia de Misiones, Argentina. Siendo:

- ZONA NORTE: Eldorado, Iguazú, Montecarlo, Puerto Rico, San Antonio, El Soberbio.
- ZONA CENTRO: Oberá, Leandro N Alem, Aristóbulo del Valle, Aurora, Jardín América, 25 de Mayo y 2 de Mayo.
- ZONA SUR: Posadas, Apóstoles, Concepción de la Sierra, San Ignacio, San Javier.

PROF en **REGISTRO:** se transformó la variable PROF en REGISTRO, ya que el nombre del profesional de donde se extrajo el registro de historia clínica implantológica no aporta información, y además para resguardar la identidad del mismo. La nueva variable permite determinar si el odontólogo especialista en implantología dental pertenece al *"Registro de Profesionales que practican Cirugía Buco maxilofacial, Implantología, Periodoncia y Manipulación de Tejidos"*. Definiéndose SI (se encuentra matriculado en el registro) y NO (no se encuentra matriculado en el registro). Este registro permite determinar si el especialista posee dos años de experiencia y 200 horas mínimas en cursos de especialización en implantología dental.

NRO_DENTAL en **PIEZA_DENTARIA:** se convirtió la variable NRO_DENTAL en una nueva variable donde se representa cada número de las piezas dentales por cuadrantes, más específicamente por zona de cada cuadrante (ver Tabla 8). Quedando: maxilar anterior y posterior, mandíbula anterior y posterior.

Tabla 8. Cuadrantes y números de pieza dental.

CUADRANTE 1								**CUADRANTE 2**							
Maxilar Posterior					*Maxilar Anterior*			*Maxilar Anterior*			*Maxilar Posterior*				
18	17	16	15	14	13	12	11	21	22	23	24	25	26	27	28
48	47	46	45	44	43	42	41	31	32	33	34	35	36	37	38
Mandíbula Posterior					*Mandíbula Anterior*			*Mandíbula Anterior*			*Mandíbula Posterior*				
CUADRANTE 4								**CUADRANTE 3**							

PROC_ADIC: se agrupo las variables EXPAN_OSEA, ELEV_SENO_MAXILAR, REG_TEJ_DUROS y REG_TEJ_BLANDOS en una sola variable, para observar y estudiar el comportamiento. Esta nueva variable determina si se realizó alguna técnica o procedimiento para:

- Incrementar la anchura del hueso mandibular (EXPAN_OSEA);
- Incrementar la altura ósea en el maxilar superior al nivel de los molares y premolares (ELEV_SENO_MAXILAR);
- Regeneración de tejidos duros (hueso) en la zona de intervención (REG_TEJ_DUROS); o
- Regeneración de tejidos blandos (encía) en la zona de intervención (REG_TEJ_BLANDOS).

Todas estas integraciones y formateo de los datos de las variables se hicieron en función de las recomendaciones de los expertos.

Finalmente, se procedió a confeccionar un diccionario de datos. En la Tabla 9 se especifican las variables que contiene cada dimensión del conjunto de datos final, junto a una descripción de cada una. Con este conjunto se procede a la próxima etapa, que se corresponde con la etapa de modelado.

Tabla 9. Descripción de las variables finales de cada dimensión.

Dimensión	Variable	Descripción
Datos del Paciente	RANGO_EDAD	Rango de edad en años cumplidos por el paciente al momento de la colocación del implante dental.
	GENERO	Sexo del paciente (femenino o masculino).
	OCUPACION	Ocupación o campo en el que se desempeña el paciente.
	O_S	Este campo permite determinar si el paciente efectúa la intervención implantología por seguro médico o de manera particular.
	ANTECEDENTE	Esta característica se refiere a la historia de las enfermedades sufridas o preexistentes en el paciente.
	FUMA	Este campo determina si el paciente es fumador.
	ALCOHOLISMO	Define si el paciente es alcohólico.

Dimensión	Variable	Descripción
Datos del Paciente	PERIO	Determina el grado de afección (periodontitis) que tiene el tejido o hueso alrededor de una o varias piezas dentarias del paciente.
	DESDENTADO	Este campo permite determinar si el paciente carece de dientes o cuenta con muy pocas piezas dentarias (menos del 80 % de la totalidad).
	INGEST_MED	Esta característica se refiere a si el paciente ingiere algún tipo de medicamento a momento de la cirugía.
	ALERGIA	Establece si el paciente tiene alguna alergia, ya sea a un medicamento o alguna sustancia.
Datos del Implante	TRAT_SUP	Este campo brinda información sobre el tratamiento de superficie del implante utilizado en la intervención quirúrgica por el odontólogo. Considerando: T1: Blasting + doble grabado ácido; T2: Blasting + grabado ácido + tratamiento térmico; T3: Blasting + doble grabado ácido + depósitos de nano cristales de fosfato cálcico; T4: Grabado bi-ácido; T5: Blasting + grabado ácido; T6: Blasting con fosfato tricálcico + grabado ácido o T7: Tratamiento Electroquímico.
	DISENIO	Define la forma o macrogeometría del implante.
	LONGITUD	Concepto métrico en milímetros que define el largo o extensión del implante.
	DIAMETRO	Concepto métrico en milímetros que define el grosor o espesor del implante.
	CONEXION	Representa la forma de conexión entre el pilar y el implante, definiendo la existencia o ausencia de una figura geométrica que se extiende por sobre la superficie de la corana del implante.
	PROCEDENCIA	Determina el origen del implante. El cual puede ser nacional (implantes fabricados en el territorio argentino) o importado (implantes fabricados en territorios de países extranjeros, como: Suiza, EEUU, Israel, Brasil y Chile).

Dimensión	Variable	Descripción
Datos de la Fase Quirúrgica	ESTACION_ANIO	Indica la estación del año en el que el paciente se realizó la intervención quirúrgica, considerando las cuatro estaciones.
	ZONA_PACIENTE	Este campo determina la procedencia del paciente, indicando la zona en la cual reside, correspondiente a la provincia de Misiones, Argentina.
	REGISTRO	Permite comprobar si el odontólogo encargado de la intervención quirúrgica pertenece al *"Registro Provincial de Profesionales que practican Cirugía Buco Maxilofacial, Implantología, Periodoncia y Manipulación de Tejidos"*. Este registro permite precisar si el especialista posee dos años de experiencia y 200 horas mínimas en cursos de especialización.
	PIEZA_DENTARIA	Este campo identifica la localización o el cuadrante del implante en la boca del paciente.
	PROT_CARGA	Hace referencia al tiempo transcurrido entre la extracción de la pieza dentaria y la colocación del implante dental. Puede ser inmediato (24-48 horas), temprano (posterior a las 48 horas y dentro del mes) o tardío (posterior al mes).
	EXODONCIA	Establece si se produjo un acto quirúrgico en el que se extrajo el diente o una parte remanente del mismo que estaba alojado en el alveolo.
	EXPAN_OSEA	Determina si se realizó alguna técnica para incrementar la anchura del hueso mandibular antes de la colocación del implante dental.
	ELEV_SENO_MAXILAR	Determina si se realizó alguna técnica para incrementar la altura ósea en el maxilar superior (al nivel de los molares y premolares) para la colocación del implante dental.
	REG_TEJ_DUROS	Determina si se regeneró tejidos duros (hueso) en la zona donde se colocó el implante.
	REG_TEJ_BLANDOS	Determina si se regenero tejidos blandos (encía) en la zona donde se colocó el implante.

Dimensión	Variable	Descripción
Datos de la Fase Quirúrgica	PROC_ADICIONAL	Establece si el paciente pasó por una instancia de tratamiento previo de preparación del lecho óseo.
	TIEM_COLOC	Hace referencia al tiempo transcurrido entre la colocación del implante y la carga de la corona. Puede ser inmediato (contiguo a la colocación del implante), temprano (dentro de los 3 meses) o tardío (posterior a los 3 meses).
	TIPO_HUESO	Define la calidad o tipo de hueso del paciente en la zona a ser intervenida. Puede ser Tipo I (hueso muy denso), Tipo II (hueso trabecular denso), Tipo III (hueso trabecular más esponjoso) o Tipo IV (hueso con trabéculas grandes y débiles).
	INDIC_PROTESICA	Determina la complejidad de la cirugía y la carga que tendrá el implante. Puede ser pieza unitaria (P_U), puente (P), prótesis completa fija (P_C_F) o prótesis completa removible (P_C_R).
	COMP_QUIRURGICAS	Este concepto determina si se produjo alguna complicación en la intervención quirúrgica.
Datos del Seguimiento Postoperatorio	SEGUI_POSTOP	Determina si el implante tuvo éxito en el proceso de oseointegración o el proceso fracasó.
	OBSERVACIONES	Establece porque razón el implante fracaso o simplemente que rasgos se notaron luego del proceso de curación.

La Tabla 10 detalla las variables transformadas del conjunto de datos según la dimensión a la que pertenece, tipo de dato, valor que contiene y distribución de cada variable respecto al total de los datos.

Tabla 10. Dimensión, tipo de dato, valor que contiene y distribución de cada variable.

Dimensión	Variable	Tipo de dato	Valores	Cant.	Dist. (%)
Datos del Paciente	RANGO_EDAD	Cadena	ENTRE 16 Y 39 ENTRE 40 Y 51 ENTRE 52 Y 90	264 370 531	23% 32% 46%

Dimensión	Variable	Tipo de dato	Valores	Cant.	Dist. (%)
Datos del Paciente	GENERO	Carácter	F M	649 516	56% 44%
	OCUPACION	Cadena	ACADEMICO ADMINISTRACION DESOCUPADO EMPLEADO INDEPENDIENTE PROFESIONAL TÉCNICO	153 168 309 113 248 124 50	13% 14% 27% 10% 21% 11% 4%
	O_S	Cadena	NO SI	360 805	31% 69%
	ANTECEDENTE	Cadena	ARTRITIS COLESTEROL DIABETES PROB_CARD TIROIDES OTROS NO	13 8 63 149 13 38 881	1% 1% 5% 13% 1% 3% 76%
	FUMA	Cadena	NO SI	175 990	15% 85%
	ALCOHOLISMO	Cadena	NO SI	1.150 15	99% 1%
	PERIO	Cadena	BAJO MEDIO ALTO NO	225 93 24 823	19% 8% 2% 71%
	DESDENTADO	Cadena	NO SI	1.007 158	86% 14%
	INGEST_MED	Cadena	AAA ANTIDIABETICO ANTIHIPERTENSIVO ANTIULCEROSO HORMONAS SUPLEMENTO OTROS NO	19 34 88 13 25 12 35 939	2% 3% 8% 1% 2% 1% 3% 81%
	ALERGIA	Cadena	ASPIRINETA PENICILINA POLEN UROBIOTIC_INY NO	12 14 10 9 1.120	1% 1% 1% 1% 96%

Dimensión	Variable	Tipo de dato	Valores	Cant.	Dist. (%)
Datos del Implante	DISENIO	Cadena	CILINDRICO CONICO	494 671	42% 58%
	LONGITUD	Cadena	4 A 6 MM 7 A 9 MM 10 A 12 MM 13 A 15 MM	17 282 739 127	2% 24% 63% 11%
	DIAMETRO	Cadena	2,5 A 3 MM 3 A 3,5 MM 3,5 A 4 MM 4 A 4,5 MM 4,5 A 5 MM	10 479 522 93 61	1% 41% 45% 8% 5%
	CONEXION	Cadena	EXTERNO INTERNO	497 668	43% 57%
	PROCEDENCIA	Cadena	IMPORTADO NACIONAL	342 823	29% 71%
	TRAT_SUP	Cadena	T1 T2 T3 T4 T5 T6 T7	553 60 100 255 65 41 91	48% 5% 9% 22% 6% 4% 8%
Datos de la Fase Quirúrgica	ESTACION_ANIO	Cadena	VERANO OTOÑO INVIERNO PRIMAVERA	296 309 308 252	25% 27% 26% 22%
	ZONA_PAC	Cadena	ZONA NORTE ZONA CENTRO ZONA SUR	104 254 807	9% 22% 69%
	REGISTRO	Cadena	NO SI	172 993	15% 85%
	PIEZA_DENTARIA	Cadena	MANDIBULA_A MANDIBULA_P MAXILAR_A MAXILAR_P	44 359 335 427	4% 31% 29% 37%
	PROT_CARGA	Cadena	INMEDIATO TARDIO TEMPRANO	29 1.105 31	3% 95% 3%
	EXODONCIA	Cadena	NO SI	1.003 162	86% 14%
	EXPAN_OSEA	Cadena	NO SI	884 281	76% 24%

Dimensión	Variable	Tipo de dato	Valores	Cant.	Dist. (%)
Datos de la Fase Quirúrgica	ELEV_SENO_MAXILAR	Cadena	NO SI	1.124 41	97% 4%
	REG_TEJ_DUROS	Cadena	NO SI	1.101 64	95% 6%
	REG_TEJ_BLANDOS	Cadena	NO SI	1.090 75	94% 6%
	PROC_ADICIONAL	Cadena	NO SI	840 325	72% 28%
	TIEM_COLOC	Cadena	INMEDIATO TARDIO TEMPRANO	93 1.045 27	8% 90% 2%
	TIPO_HUESO	Cadena	TIPO I TIPO II TIPO III TIPO IV	10 128 562 465	1% 11% 48% 40%
	INDIC_PROTESICA	Cadena	P P_C_F P_C_R P_U	68 100 9 988	6% 9% 1% 85%
	COMP_QUIRURGICAS	Cadena	FALTA_CIERRE_PRIM NINGUNA OTROS	38 1.117 10	3% 96% 1%
Datos de Seguimiento Postoperatorio	SEGUI_POSTOP	Cadena	EXITO FRACASO	1.009 156	87% 13%
	OBSERVACIONES	Cadena	EXPO_TORNILO INTOX_TABAQUISMO MOVILIDAD MUCOSITIS_PERI PERDIDA_OSEA PERIIMPLANTITIS NINGUNA	14 11 32 12 1.040 30 26	1% 1% 3% 1% 89% 3% 2%

La Figura 30 muestra la distribución de la variable clase a predecir. Donde se puede apreciar claramente que se trata de un conjunto desbalanceado.

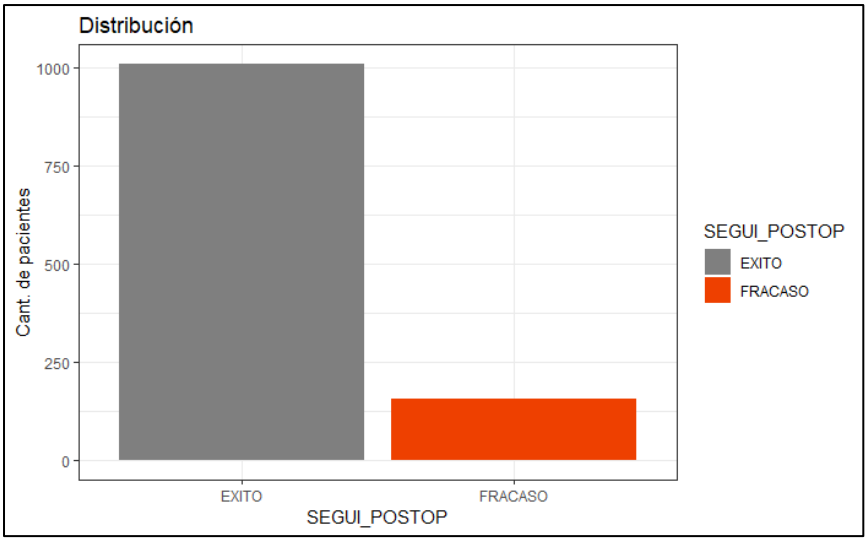


Figura 30. Distribución de la variable clase SEGUI_POSTOP del conjunto de datos de *Implantes Dentales*.

Se decidió no utilizar la variable OBSERVACION para el estudio de caso, debido a que la información que contiene es sobre el motivo del fracaso del proceso de oseointegración tejido/implante, razón por la que genera ruido en el proceso de extracción de conocimiento. Esto se apreció al realizar un ensayo exploratorio.

La Figura 31 detalla el resultado de medir la ganancia de información de todas las variables con el método Information Gain y la Figura 32 refleja el resultado logrado de comparar todas las variables con el método Gain Ratio.

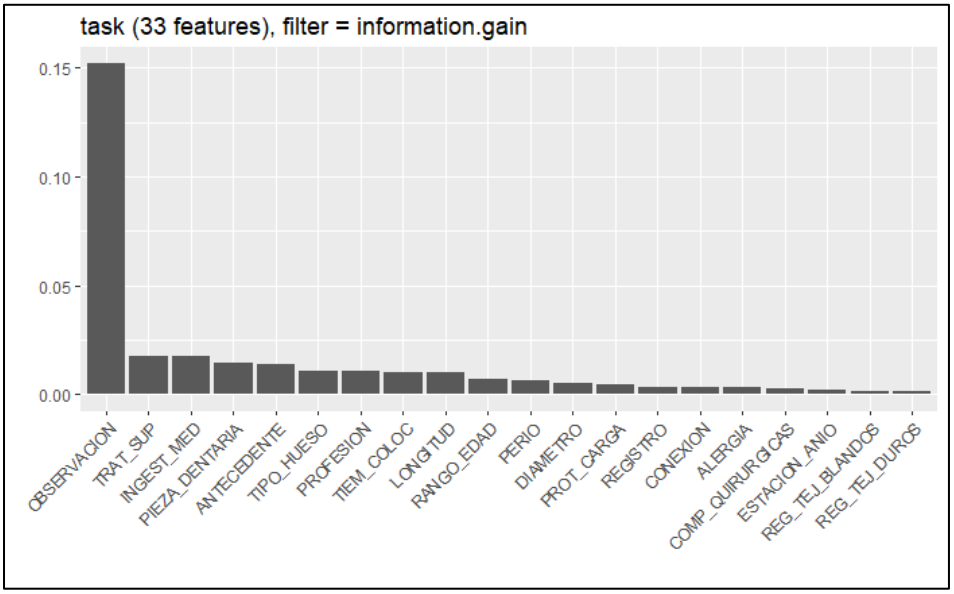


Figura 31. Ganancia de información de la variable OBSERVACION empleando el método Information Gain.

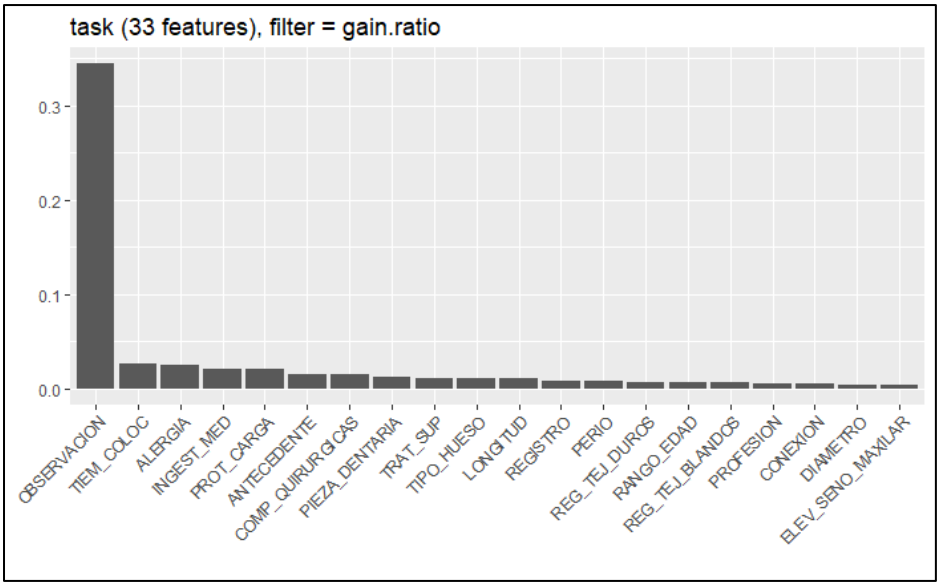


Figura 32. Ganancia de información de la variable OBSERVACION empleando el método Gain Ratio.

3.4 MODELADO

Esta fase se desarrolla en dos etapas. La primer parte se corresponde con un proceso para seleccionar las características más importantes y la segunda recae en la definición de la estrategia y el procedimiento para la clasificación e identificación de los casos de fracaso para el conjunto de datos de implantes dentales.

3.4.1 PROCEDIMIENTO PARA LA SELECCIÓN DE CARACTERÍSTICAS

Como se abordó en la sección *2.7 ANTECEDENTES EN LA UTILIZACIÓN DE MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS* del apartado del Marco Teórico, varios investigadores han utilizado los métodos de selección de características Information Gain (IG), Gain Ratio (GR), Random Forest importance (RFI), Relief (R) y Chi-Squared (ChiS) para reducir la dimensionalidad y seleccionar las características más importantes de un conjunto de datos [45], [138], [140], [146], [149]–[153]. Conjuntamente, los clasificadores comúnmente utilizados son Support Vector Machine (SVM), Random Forest (RF) y K Nearest Neighbors (KNN) con validación cruzada. En evaluaciones exploratorias, estos clasificadores obtuvieron el mejor rendimiento sobre los conjuntos de datos empleados. Además, estos algoritmos de aprendizaje se han seleccionado porque son representativos de diferentes tipos de clasificadores y son ampliamente utilizados en estudios similares para validar los subconjuntos de características resultantes. Mientras que las medidas de rendimiento comúnmente utilizadas son la matriz de confusión, precisión (accuracy) y exactitud equilibrada (bac) [75], [123], [124], [142]. Por lo tanto, estos clasificadores y medidas de

rendimiento se tuvieron en cuenta para el diseño y la validación del procedimiento propuesto en esta primera etapa.

En la Figura 33 se aprecia los pasos de este procedimiento. Cabe mencionar, que los métodos de selección de características así como los clasificadores utilizados, se los consideró con el mismo peso. Debido a que la finalidad de esta primera etapa fue conocer las características seleccionadas por cada método, para posteriormente corroborar el rendimiento que tuvo ese subconjunto de características con cada uno de los clasificadores utilizados.

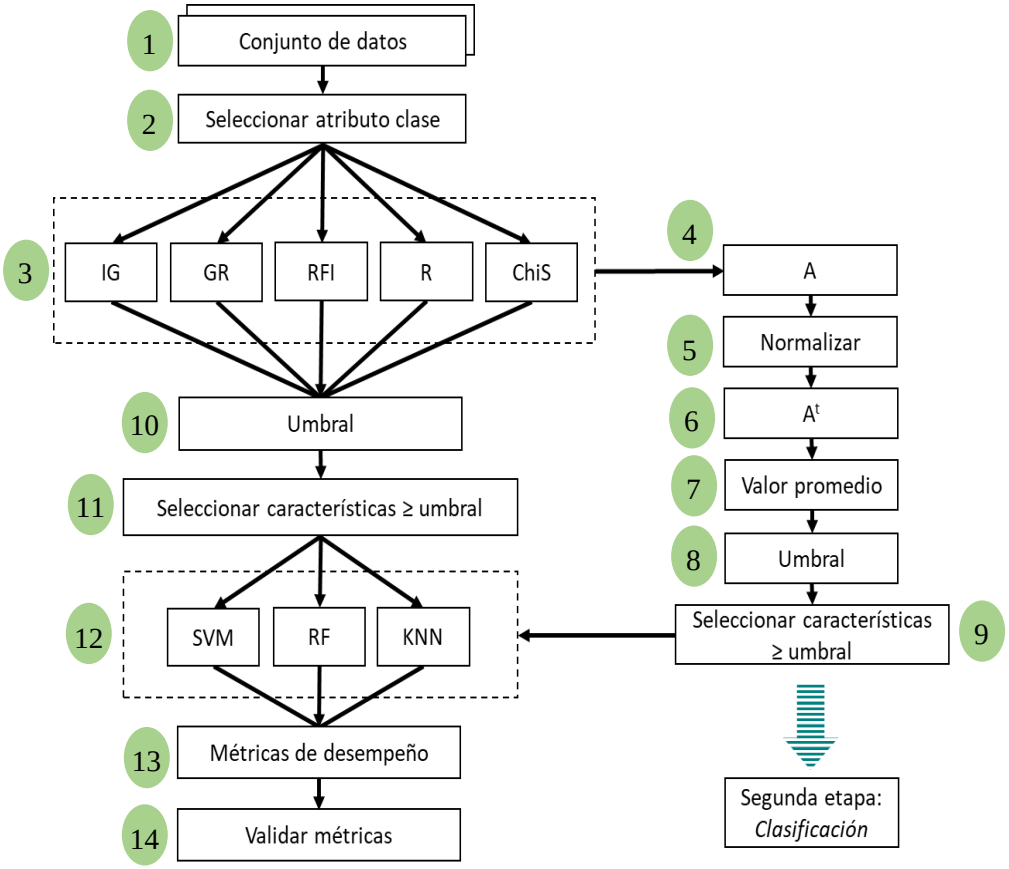


Figura 33. Procedimiento propuesto para la selección de características más relevantes.

El procedimiento de selección de características propuesto consiste en unificar los resultados de los métodos IG, GR, RFI, R y ChiS. A continuación, se detallan los pasos. Cabe remarcar que el procedimiento de esta sección fue elaborado sobre la herramienta software R.

Paso 1. Leer el conjunto de datos.

Paso 2. Seleccionar la característica objetivo para la predicción (variable clase).

Paso 3. Obtener los subconjuntos de características de los métodos de selección de características: IG, GR, RFI, R y ChiS.

Paso 4. Confeccionar una matriz (A) que concentre el valor de importancia obtenido por los cinco métodos para cada característica. Es decir, para una misma característica habrá cinco posibles valores diferentes de importancia.

Paso 5. Normalizar los valores, debido a que los métodos empleados se desempeñan en rangos diferentes, este paso es fundamental para lograr un valor medio de cada característica. Para este fin se utilizó la función *normalize* (ecuación 21), la cual permite normalizar valores en base al método mínimo-máximo.

La normalización mínimo-máximo regulariza las características en un rango [24]. Dado min_A y max_A valores mínimo y máximo de una característica A. La normalización mínimo-máximo mapea un valor v_i de A para v_i' en el rango $[new\_min_A, new\_max_A]$ mediante:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \qquad 21$$

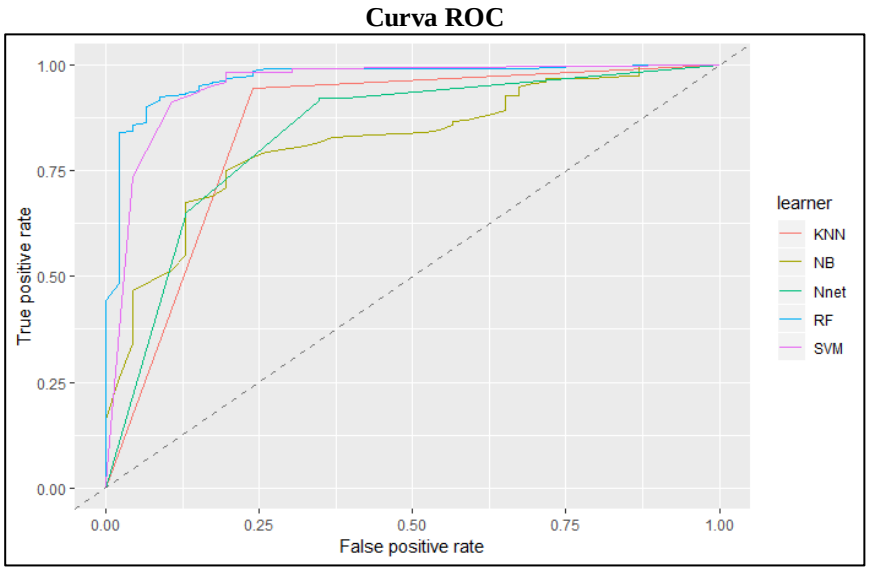
Se utilizó este criterio de normalización debido a que permite preservar todas las relaciones de los valores de los datos originales, es decir no introduce ningún sesgo potencial en los datos. Además, se encuentra demostrado que tiene mejor rendimiento en la clasificación [183], [184]. El rango empleado fue [0,1].

Paso 6. Transponer la matriz (A^t). Simplemente para facilitar las operaciones por columna. Ya que la matriz A concentra en su cabecera a los métodos (IG, GR, RFI, R y CHiS) y en sus filas a cada una de las características. Y el objetivo es obtener un valor medio por cada característica y no por cada método de selección de características.

Paso 7. Obtener un valor medio de cada característica en función de los valores obtenidos por los diferentes métodos. Se empleó la mediana [24] como medida de tendencia central debido a que los valores de importancia dados por los diferentes métodos no seguían una distribución normal. En el caso de que los valores sigan una distribución normal, se debe aplicar la media [24].

Paso 8. Obtener un umbral. Este umbral se determinó mediante una búsqueda en la cuadrícula utilizando un parámetro de prueba con valores entre 0,1 y 1, con incrementos de 0,1 en cada prueba. Esta búsqueda se sometió a una validación cruzada de 10 iteraciones. Este ensayo se realiza con los valores obtenidos en el paso 7. El valor del umbral seleccionado fue el que permitió obtener la mejor precisión (bac) en la clasificación con RF.

Se ha utilizado un clasificador RF para la búsqueda de los umbrales, ya que permitió lograr el mejor desempeño en comparación a otros clasificadores. Para determinar el rendimiento, se examinó varios tipos de clasificadores (con calibración) y se calculó el área bajo la curva (auc). Esto se refleja mediante la curva ROC de la Figura 34, la cual resume el desempeño de los clasificadores KNN, NB, Nnet, RF y SVM sobre el conjunto de datos de *Implantes Dentales* (sin selección de características). Este comportamiento se repite de manera equivalente para los otros conjuntos de datos empleados en la validación experimental.



	auc	bac
KNN	0.85	0.85
NB	0.82	0.63
Nnet	0.85	0.79
RF	0.97	0.87
SVM	0.95	0.84

Figura 34. Rendimiento de los clasificadores KNN, NB, Nnet, RF y SVM sobre el conjunto de datos *Implantes Dentales*.

Paso 9. Seleccionar las características que cumplan con la condición de ser igual o mayor al umbral obtenido en el paso 8.

Paso 10. Obtener un umbral óptimo para cada uno de los cinco métodos de selección de características. Esto se hizo a través de un ajuste de las características seleccionadas por cada método, contrastando su rendimiento en la clasificación con un RF (se utilizó una validación cruzada de 10 iteraciones). El valor del umbral seleccionado para cada método fue el que permitió obtener la mejor precisión (bac) en la clasificación.

Paso 11. Seleccionar las características para cada uno de los cinco métodos que cumplan con la condición de ser igual o mayor al umbral obtenido en el paso 10.

Paso 12. Aplicar los clasificadores SVM, RF y KNN con una validación cruzada de 10 iteraciones sobre los cinco conjuntos de características obtenidos en el paso 11 y a la matriz normalizada lograda en el paso 9. Se ha seleccionado estos tres clasificadores, debido a que permitieron lograr la mejor precisión (bac) sobre el conjunto de datos de *Implantes Dentales* (Figura 34) y los otros conjuntos utilizados para la validación.

Paso 13. Obtener las medidas de rendimiento: TP, FP, TN, FN, bac, auc y accuracy.

Paso 14. Validar las medidas TN y bac.

3.4.2 PROCEDIMIENTO PARA LA CLASIFICACIÓN

Para mejorar y asegurar un resultado más preciso en la clasificación, se propone en esta segunda etapa un procedimiento mediante la combinación de diversas técnicas. Cabe remarcar que el procedimiento de esta sección fue elaborado sobre la herramienta software Python.

En la Figura 35 se resumen los pasos del mecanismo propuesto para la integración de las predicciones de los cinco clasificadores empleados [252].

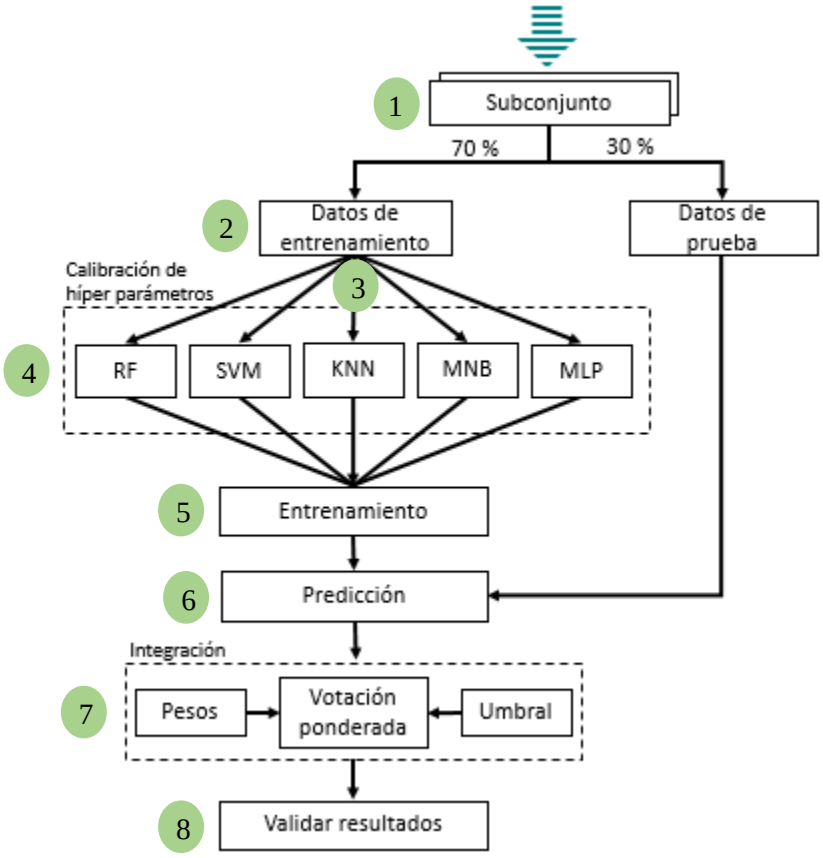


Figura 35. En esta representación se resumen los pasos del mecanismo propuesto para la integración de las predicciones de los siguientes clasificadores: Random Forest (RF), C-Support Vector (SVC), K-Nearest Neighbors (KNN), Multinomial Naive Bayes (MNB) y Multilayer Perceptron (MLP).

Seguidamente, se detallan cada uno de los pasos y las especificaciones que abarca el procedimiento planteado para la predicción de los casos de fracaso.

Paso 1. Leer el subconjunto logrado en el paso 9 del procedimiento propuesto para la selección de características más relevantes y seleccionar la característica objetivo para la predicción (variables clase).

Paso 2. Para realizar una tarea de clasificación, luego de la selección de las características más importantes de un conjunto, es necesario dividir los datos. Para el estudio de caso se dividió los datos de forma aleatoria para preservar la distribución de ambas clases en: 70 % para entrenamiento y 30 % para evaluación [44], [49], [173], [178], [194], [195], [198], [199]. Garantizando que todos los casos se encuentren representados en ambos conjuntos.

Paso 3. Un paso importante en toda tarea de clasificación es la búsqueda de los mejores clasificadores individuales para el estudio de caso. Después de examinar los tipos de clasificadores utilizados en los artículos relevados en la sección *2.8 ANTECEDENTES EN EL ENSAMBLE DE CLASIFICADORES* del apartado del Marco Teórico, proponemos el uso de los siguientes clasificadores: Random Forest (RF), Support Vector Machine (SVM), K Nearest Neighbors (KNN), Multinomial Naive Bayes (MNB) y Multi-layer Perceptron (MLP). Se utiliza más de un clasificador con el propósito de no declinar la decisión en base a los resultados de uno solo. Además, estos clasificadores obtuvieron en evaluaciones exploratorias el mejor desempeño en comparación con otros explorados, como: Rpart [88], Ada [79], Gradient Boosting Machine (GBM) [143] y distintos Naive Bayes [43].

Paso 4. Para obtener un modelo robusto y optimizar los resultados de los clasificadores, se realizó una búsqueda en cuadrícula para ajustar los híper parámetros [10], [170], [173], [182], [195]. Esta búsqueda se efectúa sobre los datos de entrenamiento del conjunto de datos. Para este proceso se especificó los siguientes requerimientos:

1. Un espacio de búsqueda, se definió rangos de valores (mínimos, máximos o posibles opciones a considerar) para cada uno de los híper parámetros y se fue ajustando en función de la medida de rendimiento seleccionada.
2. Un algoritmo de optimización o ajuste, se empleó el método GridSearchCV [253], es el más costoso en cuanto a rendimiento, pero permite cubrir todo el espacio de búsqueda definido. Este genera valores candidatos a partir del espacio de búsqueda definido para cada parámetro. Por lo que, cuando se ajusta a un conjunto de datos, se evalúan todas las combinaciones posibles de valores de los parámetro y se conserva la mejor combinación.
3. Un método de evaluación, como estrategia de remuestreo se utilizó una validación cruzada de 10 iteraciones.
4. Una medida de rendimiento, se fijó la métrica precisión de equilibrio, la cual está dada por los verdaderos positivos más los verdaderos negativos dividido por la totalidad de muestras del conjunto de datos [145].

En la Tabla 11, se exponen los híper parámetros que se buscó ajustar para cada clasificador sobre el conjunto de datos de estudio, además se detallan los espacios de búsquedas definidos para cada parámetro.

Tabla 11. Híper parámetros y rangos de búsqueda definido para los clasificadores RF, SVC, KNN, MNB y MLP.

Clasificador	Híper parámetros	Espacio de búsqueda
RF	*n_estimators*	*range (1, 150)*
	criterion	*gini, entropy*
	bootstrap	*True, False*
SVC	*kernel*	*linear, rbf, poly*
	C	*range (1, 10)*
	gamma	*range (1, 10)*
	degree	*range (1, 10)*
KNN	*n_neighbors*	*range (1, 100)*
	weights	*uniform, distance*
	p	*manhattan, euclidean*
MNB	*alpha*	*[0, 0.1, 0.2, 0.3, …, 0.9, 1]*
	fit_prior	*True, False*
	class_prior	*[0.5,0.5], [0.4,0.6], [0.6,0.4]*
MLP	*hidden_layer_sizes*	*range (1, 10)*
	activation	*logistic, tanh, relu*
	alpha	*[0.0001, 0.05]*
	solver	*lbfgs, sgd, adam*
	learning_rate	*constant, invscaling*

Paso 5. Realizar el entrenamiento de cada clasificador con los valores óptimos hallados para cada híper parámetro en el paso 4. Por ejemplo, para el clasificador RF:

Paso 6. Realizar la predicción con los datos del conjunto apartado para prueba.

Paso 7. Integrar las predicciones. Luego de examinar y evaluar las distintas técnicas empleadas en los trabajos citados de la sección *2.8 ANTECEDENTES EN EL ENSAMBLE DE CLASIFICADORES* del apartado del Marco Teórico, para la integración de los resultados de varios clasificadores y determinar la etiqueta de clase final, se aplica el método de votación suave ponderada [254], [255]. Esta regla permite lograr los mejores resultados de predicción para el estudio de caso.

Por lo tanto, la integración de las predicciones consistió en multiplicar para cada tupla el valor de probabilidad de la clase objetivo, obtenida por cada clasificador por el peso asignado al mismo. El peso fue determinado mediante una búsqueda en cuadricula utilizando un parámetro de prueba *w* con valores comprendidos entre 0 y 1. Esta búsqueda fue sometida a una validación cruzada de 10 iteraciones, donde se medió la precisión (accuracy) [144], [203] de cada clasificador, seleccionando el valor de *w* que logró la mejor precisión [148], [169], [171].

Una vez determinado los pesos, se aplicó el procedimiento del método votación suave ponderada [254], [255]. Este método recoge las probabilidades de clase predichas por cada clasificador, multiplica por el peso asignado al mismo y los promedia. La etiqueta de clase final se deriva de la etiqueta de clase con la probabilidad promedio más alta (ecuación 22). Está dado por:

$$\hat{y} = \arg max_i \sum_{j=1}^{m} w_j p_{ij} \qquad 22$$

donde p_{ij} es la probabilidad predicha por el jth clasificador, w_j es el peso asignado al jth clasificador. Este enfoque solo se recomienda si los clasificadores están bien calibrados.

Para el estudio de caso, en lugar de utilizar el promedio máximo se aplica un umbral [143], [171], ya que en evaluaciones exploratorias permitió lograr mejores resultados en la clasificación. Este umbral estuvo determinado por una búsqueda en cuadricula utilizando un parámetro de prueba μ con valores comprendidos entre 0,1 y 0,5 con incrementos de 0,1 en cada prueba. Se seleccionó el valor de μ que permitió obtener el mejor resultado de clasificación para el conjunto de datos utilizado.

Paso 8. Validar el accuracy de cada clasificador con los hallados en la integración del paso 7.

El código fuente de ambos procedimientos (*3.4.1 PROCEDIMIENTO PARA LA SELECCIÓN DE CARACTERÍSTICAS* y *3.4.2 PROCEDIMIENTO PARA LA CLASIFICACIÓN*), se encuentran alojados en un repositorio de GitHub[13].

3.5 EVALUACIÓN

En esta etapa se define la estrategia para validar los resultados obtenidos de los dos procedimientos propuestos en la etapa de modelado, el cual abarca la validación con otros conjuntos de datos de similares características y con expertos en el área de implantología.

3.5.1 CONJUNTOS DE DATOS DE VALIDACIÓN

Para la experimentación del enfoque propuesto se utilizó cuatro conjuntos de datos, un conjunto del estudio de caso y tres de validación. En la Tabla 12 se presentan las características resumidas de estos conjuntos.

[13] Código fuente de procedimientos. Disponible en https://github.com/nancyganz/Tesis. (Consultado el 02/11/2020).

Tabla 12. Características de los conjuntos de datos utilizados para la evaluación experimental. De izquierda a derecha se presenta: nombres de los conjuntos de datos, número de muestras y cantidad de atributos por tupla.

Conjunto de Datos	Muestra	Atributos
Implantes Dentales[1]	1.165	33
Artificial[2]	1.748	33
Heart Disease[3]	303	13
Breast Cancer[4]	277	10

[1] ***Implantes Dentales:*** consta de 1.165 tuplas de historias clínicas de pacientes sometido a procesos quirúrgicos de colocación de implantes dentales en la Provincia de Misiones, Argentina. Lo conforman 32 características categóricas y un atributo de clase binario desbalanceado (1.009 casos etiquetados como éxito y 156 como fracaso).

[2] ***Artificial:*** es un conjunto artificial generado con el algoritmo SMOTE, donde para obtener los casos artificiales de la clase minoritaria, la entrada residió en: T = 156 tuplas; *SMOTE N%* = 250%; k = 5. Mientras que para generar los casos artificiales de la clase mayoritaria, la entrada consistió en: T = 1.009 casos; *SMOTE N%* = 250%; k = 5. Para este último, en lugar de tomar el subconjunto de tuplas de menor índice, se modificó el algoritmo para que tome el subconjunto de índice mayor, que se corresponden con los casos de la clase éxito. El procedimiento de generación de los casos fue el mismo que el de la clase minoritaria. Finalmente, se extrajo los casos generados para ambas clases y se confeccionó un nuevo conjunto de datos artificial con una distribución similar al conjunto original de *Implantes Dentales*.

[3] ***Heart Disease:*** este conjunto de datos consta de un total de 303 tuplas con 12 atributos categóricos y un atributo de clase binario. Fue elaborado por las siguientes instituciones: Instituto Húngaro de Cardiología (Budapest), Hospital Universitario (Zúrich, Suiza), Hospital Universitario (Basilea, Suiza), Centro Médico VA y Fundación de Clínica Cleveland (EEUU). Cada tupla representa los datos obtenidos de un paciente. La característica objetivo hace referencia a la presencia o ausencia de enfermedad cardíaca. Lo conforman 138 casos de ausencia de la enfermedad y 165 con presencia de esta. Este conjunto fue obtenido del repositorio de datos kaggle[14].

[4] ***Breast Cancer:*** este conjunto de datos contiene registros de cáncer de mama que se obtuvieron en el Instituto de Oncología del Centro Médico Universitario de Ljubljana, Yugoslavia. Consta de 277 tuplas con 9 características categóricas y un atributo de clase binario. El atributo clase refleja los casos de recurrencia (81 casos) y no recurrencia (196

[14] Heart Disease, Kaggle. Disponible en https://www.kaggle.com/ronitf/heart-disease-uci/version/1# = . (Consultado el 30/07/2020).

casos) a la enfermedad. Este conjunto fue obtenido del repositorio Open Machine Learning (OpenML)[15].

3.5.2 RENDIMIENTO DE LA CLASIFICACIÓN A NIVEL HUMANO

Como se mencionó en el apartado *2.9.4 RENDIMIENTO DE CLASIFICACIÓN POR EXPERTOS HUMANOS*, el rendimiento a nivel humano permite estimar una tasa de error óptima y corroborar el funcionamiento del sistema de clasificación. Para evaluar el rendimiento del enfoque propuesto sobre el conjunto de datos *Implantes Dentales*, se realizó una comparación con la opinión de expertos humanos. Estos fueron seleccionados del *"Registro de Profesionales que practican Cirugía Buco maxilofacial, Implantología, Periodoncia y Manipulación de Tejidos"* del Colegio de Odontólogos de la provincia de Misiones, Argentina.

La evaluación estuvo sujeta a la clasificación por cuatro expertos del área, a cada uno de ellos se le suministro una muestra aleatoria distinta del 10% de prevalencia de los casos. Los casos fueron presentados sin la etiqueta para que el experto lo clasifique en función de su experiencia, y de esta manera poder contrastar con los valores hallados por el enfoque propuesto en esta tesis.

3.6 INTERPRETACIÓN

En este estudio, se sigue dos propósitos importantes. Uno es mostrar si la combinación de algoritmos o técnicas de aprendizaje automático, logran un mayor rendimiento que la aplicación de métodos individuales para el estudio de caso. El otro propósito es aumentar la predicción de los implantes que fracasan y definir los factor que mayor influencia ejercen en el proceso de extracción de conocimiento.

Para fundamentar el primer propósito, se utilizaron cinco clasificadores utilizando el conjunto de datos de paciente que se sometieron a la colocación de implantes dentales y luego se combinaron los clasificadores de manera que permita lograr un mayor rendimiento. Los resultados de este estudio demostraron que el algoritmo híbrido (procedimiento propuesto), alcanza mayor precisión que el uso de un solo algoritmo individual para la clasificación de los casos de fracasos contenidos en las historias clínicas del conjunto de datos. Además, aumentó la métrica sensibilidad (Tabla 25) significativamente y dado que es muy importante identificar a los pacientes cuyo implantes fracasan, se ha logrado el segundo propósito de este estudio.

Según la opinión de los expertos, los factores más importantes que influyen en el fracaso de los implantes dentales son diferentes. Para evaluar la efectividad del procedimiento

[15] Breast Cancer, OpenML. Disponible en https://www.openml.org/d/13. (Consultado el 30/07/2020).

propuesto, se utilizaron 1.165 casos de pacientes que se colocaron al menos un implante dental. Este conjunto de datos fue recolectado de las historias clínicas perteneciente a pacientes de varios Odontólogos e Implantólogos de la Provincia de Misiones, Argentina y consta de 33 características relacionados al proceso de colocación del implante dental y un atributo clase binario (el cual permite definir si el caso en cuestión tuvo éxito o fracaso).

A partir de los resultados del procedimiento propuesto en la etapa de modelado, los cuales se muestran en el siguiente capítulo, surgen recomendaciones de aplicación de los resultados, ejecución del modelo en otros conjuntos de datos y preguntas que podrían generar nuevas investigaciones. En el capítulo *4. RESULTADOS Y DISCUSIÓN* se presenta un informe concluyente y una revisión de los resultados logrados y del proyecto de Ciencia de Datos.

4. RESULTADOS Y DISCUSIÓN

4.1 PROCEDIMIENTO DE SELECCIÓN DE CARACTERÍSTICAS

En esta sección, se evalúa y presentan los resultados obtenidos de aplicar el procedimiento propuesto para la selección de las características más importantes sobre los cuatro conjuntos de datos propuestos (*Implantes Dentales, Artificial, Heart Disease* y *Breast Cancer*). El procedimiento estuvo sujeto a la combinación de los métodos: Information Gain (IG), Gain Ratio (GR), Random Forest importance (RFI), Relief (R) y Chi-Squared (ChiS). Los cuales fueron abordados en detalle en la sección *2.7.1 MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS SELECCIONADOS* del apartado del Marco Teórico. Para la validación, se utilizó los clasificadores: Support Vector Machine (SVM), Random Forest (RF) y K Nearest Neighbors (KNN), explicados en la sección *3.4.1 PROCEDIMIENTO PARA LA SELECCIÓN DE CARACTERÍSTICAS.*

Conjuntamente, en esta sección se presentan los resultados y la comparación de la validación realizada con los expertos humanos (Odontólogos e Implantólogos), para la selección de características del conjunto de datos de *Implantes Dentales*.

4.1.1 CONJUNTO DE DATOS *IMPLANTES DENTALES*

En la Tabla 13 se listan las características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto para el conjunto de datos *Implantes Dentales*.

Tabla 13. Características seleccionadas por los métodos IG, GR, RFI, R, y ChiS así como por el procedimiento propuesto para el conjunto de datos de *Implantes Dentales*.

Método	Características seleccionadas
IG	Ingesta de medicamentos, antecedentes médicos, ocupación, tipo de hueso, longitud, tratamiento de superficie, pieza dentaria, tiempo de colocación, rango de edad, conexión, diámetro, estación del año, protocolo de carga, periodontitis, alergia y registro.
GR	Alergia, ingesta de medicamentos, tiempo de colocación, protocolo de carga, longitud, tipo de hueso, complicación quirúrgica, pieza dentaria, regeneración de tejidos blandos, conexión, tratamiento de superficie, ocupación, registro, rango de edad y periodontitis.
RFI	Ocupación, tratamiento de superficie, estación del año, pieza dentaria, longitud, tipo de hueso, zona del paciente, rango edad, conexión, ingesta de medicamentos, periodontitis y antecedentes médicos.
R	Estación del año, diámetro, tratamiento de superficie, ocupación, pieza dentaria, rango de edad, genero, zona del paciente, registro, tiempo de colocación, procedencia, periodontitis, tipo de hueso y conexión.

Método	Características seleccionadas
ChiS	Ingesta de medicamentos, antecedentes médicos, longitud, ocupación, tipo de hueso, pieza dentaria, tratamiento de superficie, tiempo de colocación, conexión, rango de edad, periodontitis, estación del año y diámetro.
Procedimiento Propuesto	Ingesta de medicamentos, ocupación, antecedentes médicos, tratamiento de superficie, pieza dentaria, tipo de hueso, longitud, estación del año, tiempo de colocación, rango de edad, diámetro, conexión, periodontitis, complicación quirúrgica, registro, y regeneración de tejidos blandos.

La Tabla 14 muestra el número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto. Además, se reflejan los valores derivados de realizar la tarea de clasificación con los clasificadores SVM, RF y KNN sobre el conjunto de datos de *Implantes Dentales*.

Tabla 14. Número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto, junto con los resultados obtenidos en la clasificación con los clasificadores SVM, RF y KNN para el conjunto de datos de *Implantes Dentales*.

Método	Nro.	Clasificador	Medidas de rendimiento						
			TP	*FN*	*TN*	*FP*	*bac*	*auc*	*acc*
IG	16	SVM	97%	3%	34%	66%	66%	85%	89%
		RF	97%	3%	65%	35%	81%	92%	93%
		KNN	95%	5%	56%	44%	76%	90%	90%
GR	16	SVM	97%	3%	34%	66%	66%	85%	89%
		RF	97%	3%	52%	48%	74%	91%	91%
		KNN	96%	4%	50%	50%	73%	89%	90%
RFI	12	SVM	98%	2%	30%	70%	64%	83%	89%
		RF	97%	3%	57%	43%	77%	90%	92%
		KNN	95%	5%	50%	50%	73%	87%	89%
R	14	SVM	98%	2%	32%	68%	65%	83%	89%
		RF	98%	2%	54%	46%	76%	90%	92%
		KNN	95%	5%	54%	46%	74%	88%	89%
ChiS	13	SVM	98%	2%	35%	65%	66%	84%	89%
		RF	98%	2%	61%	39%	79%	91%	93%
		KNN	95%	5%	55%	45%	75%	88%	90%

Método	Nro.	Clasificador	Medidas de rendimiento						
			TP	*FN*	*TN*	*FP*	*bac*	*auc*	*acc*
Procedimiento Propuesto	16	SVM	97%	3%	**37%**	63%	**67%**	85%	89%
		RF	98%	2%	**66%**	34%	**82%**	92%	93%
		KNN	95%	5%	**58%**	42%	**77%**	88%	90%

En la Figura 36 se muestra el porcentaje de verdaderos negativos (TN) obtenido con los clasificadores SVM, RF y KNN, basado en las características seleccionadas por cada método de selección de características y el procedimiento propuesto. En esta figura se muestra que el procedimiento propuesto logró el mejor rendimiento en la predicción de la clase minoritaria en comparación con el logrado por los otros cinco métodos.

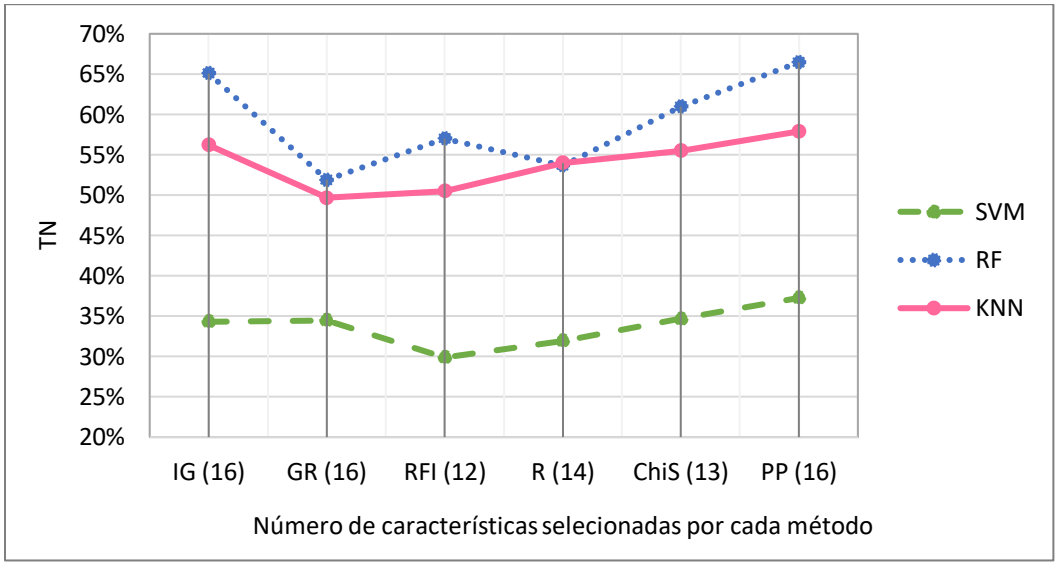


Figura 36. Porcentaje de verdaderos negativos obtenido por los clasificadores SVM, RF y KNN, basada en el subconjunto de características seleccionadas por los métodos de selección de características, así como por el procedimiento propuesto (PP).

En la Figura 37 se muestra el porcentaje de exactitud (bac) de los casos correctamente identificados para ambas clases, es decir, los éxitos y fracasos del conjunto de datos de *Implantes Dentales*, obtenidos por los clasificadores SVM, RF y KNN, en función de las características seleccionadas por cada método de selección de características y el procedimiento propuesto.

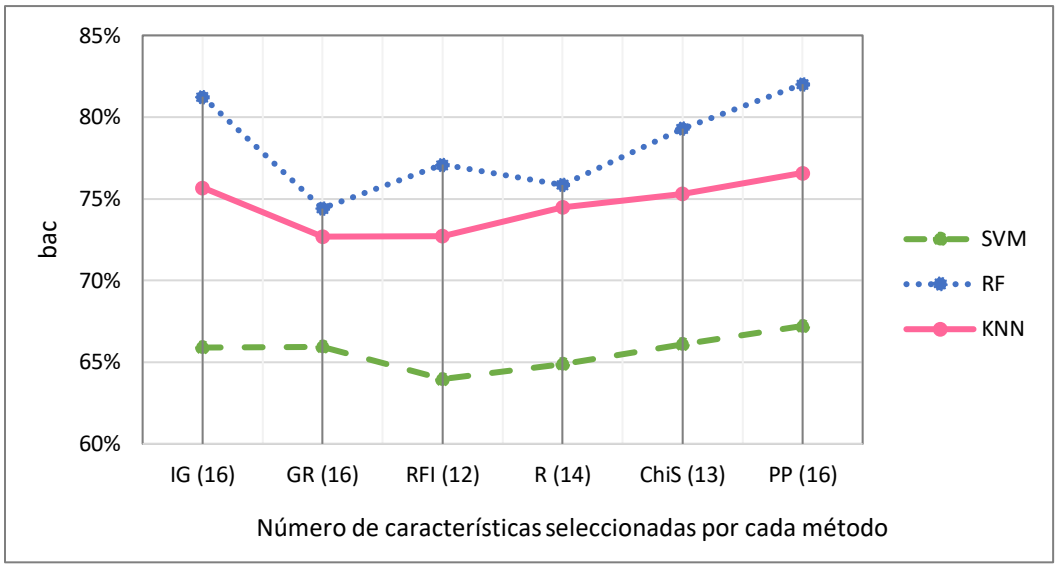


Figura 37. Exactitud equilibrada (bac) obtenido por los clasificadores SVM, RF y KNN, basado en el subconjunto de características seleccionadas por los métodos de selección de características, así como por el procedimiento propuesto (PP).

Cuando se seleccionan características de un conjunto de datos, el proceso debe incluir el análisis y ensayo de varios métodos de selección de características, para no sesgar la decisión en función de los resultados de un solo método. Por ello es conveniente utilizar una combinación de métodos para tener una visión más completa.

A través de la experimentación con el conjunto de datos de *Implantes Dentales*, el procedimiento propuesto logró el mejor rendimiento de la clase fracaso, superando los métodos IG, GR, RFI, R y ChiS. Sin descuidar el porcentaje de precisión para la clase de éxito.

Es importante tener en cuenta que no siempre es bueno reducir el conjunto de características al mínimo posible, puesto que se pueden descartar características sustanciales para el estudio y reducir la eficiencia en la predicción. Además, en términos de costo y rendimiento de cálculo, no siempre es bueno seleccionar todas las características posibles, debido a que a veces introducen ruido y no proporcionan entropía.

Podemos argumentar que, en el caso de conjuntos de datos desequilibrados, el TN y el bac son naturalmente las mejores medidas que se pueden buscar y medir, ya que benefician a la clase minoritaria, que es la que a menudo se intenta predecir, como han afirmado varios investigadores [181], [203], [256], [257].

Conjuntamente, se aprecia que el clasificador Random Forest ha tenido un muy buen desempeño en la clasificación, logrando buenos resultados de precisión, en comparación de

SVM y KNN. Esto puede depender en gran medida de la cantidad de datos, tipo y otras cuestiones relacionados al conjunto de datos.

4.1.1.1 VALIDACIÓN CON EXPERTOS HUMANOS

Para evaluar el rendimiento del procedimiento propuesto de selección de características se hizo una comparación con la opinión de los expertos. Éstos fueron seleccionados del *"Registro Provincial de Profesionales que practican Cirugía Buco Maxilofacial, Implantología, Periodoncia y Manipulación de Tejidos"* del Colegio de Odontólogos de la provincia de Misiones, Argentina.

La evaluación estuvo sujeta a la selección de características por parte de cuatro expertos, a cada uno de los cuales se les entregó un formulario donde debían marcar las características que según su experiencia consideraban más importantes a la hora de planificar una cirugía de colocación de implantes dentales. En la Tabla 15 se muestran las características seleccionadas por los expertos en comparación con el procedimiento propuesto de selección de características.

Tabla 15. Características seleccionadas por los expertos y el procedimiento propuesto.

Experto 1	Experto 2	Experto 3	Experto 4	Procedimiento propuesto
Antecedente	Edad	Edad	Antecedente	Rango de Edad
Tabaquismo	Antecedente	Antecedente	Tabaquismo	Ocupación
Alcoholismo	Tabaquismo	Tabaquismo	Periodontitis	Antecedente
Periodontitis	Alcoholismo	Periodontitis	Desdentado	Periodontitis
Desdentado	Periodontitis	Ingesta de Medicamento	Ingesta de Medicamento	Ingesta de Medicamento
Ingesta de Medicamento	Desdentado	Alergia	Diseño	Trat. de Superficie
Alergia	Ingesta de Medicamento	Diseño	Longitud	Longitud
Trat. de Superficie	Diseño	Longitud	Diámetro	Diámetro
Diseño	Longitud	Diámetro	Trat. de Superficie	Conexión
Longitud	Diámetro	Pieza Dental	Pieza Dental	Estación del Año
Diámetro	Pieza Dental	Protocolo de Carga	Protocolo de Carga	Pieza Dental
Pieza Dental	Protocolo de Carga	Exodoncia	Expansión Ósea	Registro
Protocolo de Carga	Expansión Ósea	Expansión Ósea	Elev. de Seno Maxilar	Reg. de Tejidos Blandos
Exodoncia	Elev. de Seno Maxilar	Elev. de Seno Maxilar	Reg. de Tejidos Duros	Tiempo de Colocación
Expansión Ósea	Reg. de Tejidos Duros	Tipo de Hueso	Reg. de Tejidos Blandos	Tipo de Hueso
Elev. de Seno Maxilar	Tiempo de Colocación	Complicación Quirúrgica	Tiempo de Colocación	Complicación Quirúrgica
Tipo de Hueso	Tipo de Hueso	Habilidad del cirujano	Tipo de Hueso	
Complicación Quirúrgica	Complicación Quirúrgica		Complicación Quirúrgica	
Habilidad del cirujano	Habilidad del cirujano		Habilidad del cirujano	

	Características seleccionadas por todos los expertos.
	Características seleccionadas por algunos expertos.
	Características no seleccionadas por los expertos.

Además, se proporcionó una lista de las características seleccionadas por el procedimiento propuesto y los expertos formularon observaciones al respecto. Sus opiniones se resumen a continuación:

- Los factores pieza dental, antecedentes médicos, periodontitis, tipo de hueso, edad, tabaquismo, características del implante, protocolo de carga y mejora del lecho óseo, son decisivos a la hora de planificar una cirugía oral para la colocación de un implante dental.
- Destacaron que la ubicación de la pieza dental a reemplazar es importante, ya que no es lo mismo colocar un implante en el maxilar anterior que en la zona posterior, debido a la calidad del lecho óseo.
- También afirmaron que se debe tener en cuenta la historia (antecedentes) y los medicamentos del paciente, debido a las diferencias entre un medicamento para controlar la hipertensión, la diabetes, un antineoplásico, un inhibidor de la reabsorción ósea o simplemente un analgésico.
- La periodontitis es un factor de gran importancia, ya que determina el grado de afección del tejido/hueso alrededor de uno o varios dientes.
- También, afirman que la mejora del lecho óseo ofrece una mayor probabilidad de éxito a un proceso de este tipo y que la complejidad de la cirugía es otro componente a tener en cuenta, ya que la planificación es diferente cuando se trata de una pieza unitaria, un puente o una prótesis completa fija o removible.
- Las características que no seleccionó el procedimiento propuesto, puede ser por razones de que se requiere mayor cantidad de casos con esas condiciones para analizar. Además, aseguran que los especialistas implantólogos controlan determinadas características previamente a una intervención quirúrgica (como por ejemplo el tabaquismo).
- En cuanto a otras características como la ocupación del paciente, la estación del año y el tratamiento de la superficie de los implantes, llegaron a la conclusión de que se necesita un estudio más detallado.
- Finalmente, todos concluyeron que la habilidad del cirujano (Registro) es crucial para que la cirugía tenga éxito.

En base a los resultados, podemos afirmar que los factores que más influyen en el fracaso de los implantes dentales, según el conjunto de datos utilizados, la opinión de los expertos y el procedimiento de selección de características de esta etapa, son: ingesta de medicamentos, ocupación, antecedente médicos, tratamiento de superficie, pieza dental, tipo de hueso, longitud del implante, estación del año, tiempo de colocación, rango de edad, diámetro del implante, conexión, periodontitis, complicación quirúrgica, registro y regeneración de los tejidos blandos.

4.1.2 VALIDACIÓN CON OTROS CONJUNTOS DE DATOS

En la Tabla 16 se listan las características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto para el conjunto de datos *Artificial.*

Tabla 16. Características seleccionadas por los métodos IG, GR, RFI, R, y ChiS así como por el procedimiento propuesto para el conjunto de datos *Artificial*.

Método	Características seleccionadas
IG	Tipo hueso, ingesta de medicamentos, antecedentes médicos, tratamiento de superficie, ocupación, pieza dentaria, rango de edad, tiempo de colocación, periodontitis, estación del año, longitud, diámetro, conexión y regeneración de tejidos blandos.
GR	Ingesta de medicamentos, tiempo de colocación, alergia, antecedentes médicos, regeneración de tejidos blandos, tipo de hueso, protocolo de carga, complicación quirúrgica, pieza dentaria, periodontitis, rango de edad, tratamiento de superficie, longitud, ocupación, regeneración de tejidos duros, elevación de seno maxilar, diámetro, conexión, y estación año.
RFI	Ocupación, tratamiento de superficie, estación del año, pieza dentaria, rango edad, tipo de hueso, periodontitis, diámetro, longitud, antecedentes médicos, ingesta de medicamentos y tiempo de colocación.
R	Ocupación, tratamiento de superficie, género, estación del año, longitud, diámetro, zona del paciente, conexión, pieza dentaria, protocolo de carga, expansión ósea, procedimiento adicional, antecedentes médicos y tipo de hueso.
ChiS	Ingesta de medicamentos, antecedentes médicos, tratamiento de superficie, tipo de hueso, pieza dentaria, periodontitis, ocupación, tiempo de colocación, rango de edad, estación del año, longitud, diámetro, conexión, y regeneración de tejidos blandos.
Procedimiento Propuesto	Ocupación, tratamiento de superficie, ingesta de medicamentos, tipo de hueso, antecedentes médicos, pieza dentaria, estación del año, tiempo de colocación, longitud, rango de edad, periodontitis, diámetro, genero, conexión, regeneración de tejidos blandos, protocolo de carga, complicación quirúrgica, expansión ósea, procedimiento adicional y registro.

La Tabla 17 especifica el número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto. Además, se reflejan los valores

derivados de realizar la tarea de clasificación con los clasificadores SVM, RF y KNN sobre el conjunto de datos *Artificial*.

Tabla 17. Número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto, junto con los resultados obtenidos en la clasificación con los clasificadores SVM, RF y KNN para el conjunto de datos de *Artificial*.

Método	**Nro.**	**Clasificador**	**Medidas de rendimiento**						
			TP	*FN*	*TN*	*FP*	*bac*	*auc*	*acc*
IG	14	SVM	98%	2%	56%	44%	77%	91%	92%
		RF	98%	2%	71%	29%	84%	97%	94%
		KNN	98%	2%	63%	37%	80%	95%	93%
GR	19	SVM	98%	2%	60%	40%	79%	92%	93%
		RF	98%	2%	74%	26%	86%	98%	95%
		KNN	98%	2%	67%	33%	83%	95%	94%
RFI	12	SVM	98%	2%	62%	38%	80%	92%	93%
		RF	98%	2%	76%	24%	87%	97%	95%
		KNN	98%	2%	67%	33%	82%	95%	94%
R	14	SVM	98%	2%	59%	41%	79%	90%	93%
		RF	99%	1%	70%	30%	84%	97%	95%
		KNN	97%	3%	64%	36%	81%	93%	93%
ChiS	14	SVM	98%	2%	58%	42%	78%	91%	92%
		RF	98%	2%	71%	29%	85%	97%	95%
		KNN	98%	2%	63%	37%	81%	95%	93%
Procedimiento Propuesto	20	SVM	98%	2%	**63%**	37%	**81%**	93%	94%
		RF	99%	1%	**76%**	24%	**88%**	98%	96%
		KNN	97%	3%	**69%**	31%	**83%**	95%	94%

En la Tabla 17 se observa que RFI con el clasificador RF logra el mismo porcentaje de acierto que el procedimiento propuesto con RF, pero disminuye notablemente el acierto con los clasificadores SVM y KNN.

En la Tabla 18 se listan las características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto para el conjunto de datos *Heart Disease.*

Tabla 18. Características seleccionadas por los métodos IG, GR, RFI, R, y ChiS así como por el procedimiento propuesto para el conjunto de datos *Heart Disease*.

Método	Características seleccionadas
IG	Thal, cp, ca, exang, slope, oldpeak, sex, age, restecg y trestbps.
GR	Thal, exang, cp, ca, oldpeak y slope.
RFI	Ca, tal, cp, exang, sex, slope, oldpeak, age, restecg, trestbps y fbs.
R	Cp, sex, ca, oldpeak, thal, age, exang y slope.
ChiS	Thal, cp, ca, exang, slope, oldpeak, sex, age, restecg.
Procedimiento Propuesto	Cp, thal, ca, exang, oldpeak, slope, sex, age, restecg y trestbps.

La Tabla 19 muestra el número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto. Además, se reflejan los valores derivados de realizar la tarea de clasificación con los clasificadores SVM, RF y KNN sobre el conjunto de datos *Heart Disease*. En esta tabla se observa que tanto ChiS como el procedimiento propuesto seleccionan la misma cantidad y las mismas características, por lo que logran resultados similares en la clasificación. Además, para este conjunto de datos RFI logra buenos resultados de precisión con RF y KNN, pero esto se debe a que seleccionó más características.

Tabla 19. Número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto, junto con los resultados obtenidos en la clasificación con los clasificadores SVM, RF y KNN para el conjunto de datos *Heart Disease*.

Método	Nro.	Clasificador	Medidas de rendimiento						
			TP	*FN*	*TN*	*FP*	*bac*	*auc*	*acc*
IG	10	SVM	84%	16%	**86%**	14%	**85%**	91%	85%
		RF	81%	19%	**85%**	15%	**83%**	90%	83%
		KNN	76%	24%	**82%**	18%	**79%**	88%	79%
GR	6	SVM	81%	19%	84%	16%	83%	90%	83%
		RF	79%	21%	86%	14%	82%	90%	83%
		KNN	83%	17%	65%	35%	74%	85%	73%

Método	Nro.	Clasificador	Medidas de rendimiento						
			TP	*FN*	*TN*	*FP*	*bac*	*auc*	*acc*
RFI	11	SVM	83%	17%	85%	15%	84%	91%	84%
		RF	82%	18%	87%	13%	84%	90%	84%
		KNN	76%	24%	83%	17%	79%	88%	80%
R	8	SVM	82%	18%	86%	14%	84%	91%	84%
		RF	82%	18%	85%	15%	83%	91%	83%
		KNN	81%	19%	81%	19%	81%	88%	81%
ChiS	9	SVM	82%	18%	86%	14%	84%	91%	84%
		RF	82%	18%	84%	16%	83%	90%	83%
		KNN	78%	22%	80%	20%	79%	87%	79%
Procedimiento Propuesto	10	SVM	84%	16%	**86%**	14%	**85%**	91%	85%
		RF	81%	19%	**85%**	15%	**83%**	90%	83%
		KNN	76%	24%	**82%**	18%	**79%**	88%	79%

En la Tabla 20 se listan las características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto para el conjunto de datos *Breast Cancer*.

Tabla 20. Características seleccionadas por los métodos IG, GR, RFI, R, y ChiS así como por el procedimiento propuesto para el conjunto de datos *Breast Cancer*.

Método	Características seleccionadas
IG	Deg.malig, inv.nodes, tumor.size, node.caps y irradiat.
GR	Deg.malig, node.caps, inv.nodes, irradiat y tumor.size.
RFI	Deg.malig, inv.nodes, node.caps y irradiat.
R	Node.caps, tumor.size, age, breast, inv.nodes, breast.qua y deg.malig.
ChiS	inv.nodes, deg.malig, node.caps, tumor.size y irradiat.
Procedimiento Propuesto	Deg.malig, inv.nodes, node.caps, tumor.size y irradiat.

La Tabla 21 muestra el número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto. Además, se reflejan los valores derivados

de realizar la tarea de clasificación con los clasificadores SVM, RF y KNN sobre el conjunto de datos *Breast Cancer*.

Tabla 21. Número de características seleccionadas por los métodos IG, GR, RFI, R y ChiS, así como por el procedimiento propuesto, junto con los resultados obtenidos en la clasificación con los clasificadores SVM, RF y KNN para el conjunto de datos de *Breast Cancer*.

Método	**Nro.**	**Clasificador**	**Medidas de rendimiento**						
			TP	*FN*	*TN*	*FP*	*bac*	*auc*	*acc*
IG	5	SVM	94%	6%	**33%**	67%	**64%**	70%	76%
		RF	91%	9%	**37%**	63%	**64%**	70%	75%
		KNN	92%	8%	**26%**	74%	**59%**	66%	72%
GR	5	SVM	94%	6%	**33%**	67%	**64%**	69%	76%
		RF	91%	9%	**37%**	63%	**64%**	70%	75%
		KNN	92%	8%	**26%**	74%	**59%**	66%	72%
RFI	4	SVM	95%	5%	33%	67%	64%	67%	77%
		RF	93%	7%	33%	67%	63%	72%	75%
		KNN	95%	5%	20%	80%	57%	69%	73%
R	7	SVM	93%	7%	29%	71%	61%	72%	74%
		RF	90%	10%	36%	64%	63%	69%	74%
		KNN	91%	9%	31%	69%	61%	66%	74%
ChiS	5	SVM	94%	6%	**33%**	67%	**64%**	69%	76%
		RF	91%	9%	**37%**	63%	**64%**	70%	75%
		KNN	92%	8%	**26%**	74%	**59%**	66%	72%
Procedimiento Propuesto	5	SVM	94%	6%	**33%**	67%	**64%**	69%	76%
		RF	91%	9%	**37%**	63%	**64%**	70%	75%
		KNN	92%	8%	**26%**	74%	**59%**	66%	72%

En la Tabla 21 se aprecia que IG, GR, ChiS y el procedimiento propuesto seleccionan las mismas características. En la Tabla 21 se observa que las características seleccionadas con estos métodos de selección de características logran con SVM y RF los mejores resultados de precisión para la clase minoritaria (TN). Mientras que R con KNN logra mejorar la precisión (TN), esto se debe a que seleccionó más características que los otros métodos.

4.2 PROCEDIMIENTO DE CLASIFICACIÓN

En esta sección, se evalúa y presentan los resultados obtenidos de aplicar el enfoque propuesto de integración de las predicciones sobre los cuatro conjuntos de datos propuestos. Así como, la validación realizada con los expertos humanos.

En la Tabla 22 se aprecia un resumen de los conjuntos de datos utilizados, así como la cantidad de características finales luego de aplicar el procedimiento propuesto para la selección de características.

Tabla 22. Características resumidas de los conjuntos de datos utilizados para la evaluación experimental. De izquierda a derecha se presenta: nombres de los conjuntos de datos, número de muestras, número de atributos por tupla, cantidad de características seleccionadas por el procedimiento de selección de características propuesto y tamaño de los conjuntos de entrenamiento y de prueba.

Conjunto de Datos	Muestra	Características	Nro.	Entrenamiento	Prueba
Implantes Dentales[1]	1.165	33	16	815	350
Artificial[2]	1.748	33	20	1.223	525
Heart Disease[3]	303	13	10	212	91
Breast Cancer[4]	277	10	5	193	84

Como se describió en el apartado *3.4.2 PROCEDIMIENTO PARA LA CLASIFICACIÓN* del capítulo materiales y métodos, la integración de las predicciones se realiza a través del método de votación suave ponderada [254], [255].

En la Tabla 23 se muestran los valores óptimos hallados en el entrenamiento para cada uno de los clasificadores (RF, SVC, KNN, MNB y MLP) con los datos de entrenamiento de cada conjunto de datos utilizado para el estudio de caso (*Implantes Dentales, Artificial, Heart Disease* y *Breast Cancer*).

La Tabla 24, refleja los valores más apropiados a ser asignados como pesos, a cada uno de los clasificadores (RF, SVC, KNN, MNB y MLP) y el umbral (*threshold*) más óptimo para definir la etiqueta final de clase de los conjuntos de datos utilizados (*Implantes Dentales, Artificial, Heart Disease* y *Breast Cancer*). Cualquiera de las combinaciones listadas (pesos y umbral) permitirá lograr por ejemplo el 79% de acierto de la clase Fracaso para el conjunto de datos de *Implantes Dentales*. Esta tabla de valores fue obtenida para cada conjunto de datos, con el objetivo de lograr el mejor ajuste y rendimiento de clasificación de las etiquetas de clase correspondientes.

En la Tabla 25 se presentan los porcentajes de acierto obtenidos por cada clasificador de forma individual y el obtenido del enfoque propuesto sobre los datos de prueba de los

conjuntos de datos utilizados. En esta tabla (Tabla 25), se observa que los clasificadores SVC y KNN logran el mejor rendimiento sobre la clase no objetivo para todos los conjuntos de datos en comparación con los demás clasificadores, incluso superan al enfoque propuesto en todos los casos. Para la clase objetivo, se aprecia que la integración de las predicciones de los cinco clasificadores permitió alcanzar el mayor porcentaje de acierto. Para esta clase, además se observa que el rendimiento de los clasificadores individuales fue variado. Si bien, el rendimiento de la integración de las predicciones no fue la mejor opción para la clase no objetivo, no quiere decir que haya sido la peor en comparación con las predicciones individuales. Mientras que la integración de las probabilidades para la clase objetivo fue la mejor opción, ya que permitió obtener el mayor porcentaje de acierto.

Tabla 23. Híper parámetros y valores óptimos encontrados para los clasificadores RF, SVC, KNN, MNB y MLP sobre los conjuntos de datos *Implantes Dentales*, *Artificial*, *Heart Disease* y *Breast Cancer*.

Clasificador	**Híper parámetros**	**Valores óptimos**			
		Implantes Dentales	*Artificial*	*Heart Disease*	*Breast Cancer*
RF	*n_estimators* *criterion* *bootstrap*	*8* *entropy* *True*	*2* *entropy* *False*	*7* *gini* *True*	*7* *gini* *True*
SVC	*kernel* *C* *gamma* *degree*	*rbf* *1* *1* *0*	*rbf* *1* *1* *0*	*rbf* *1* *1* *0*	*liner* *1* *1* *0*
KNN	*n_neighbors* *weights* *p*	*20* *distance* *euclidean*	*40* *distance* *euclidean*	*2* *uniform* *manhattan*	*50* *uniform* *manhattan*
MNB	*alpha* *fit_prior* *class_prior*	*1* *True* *[0.6,0.4]*	*0.7* *True* *[0.6,0.4]*	*0* *True* *[0.6,0.4]*	*0.2* *True* *[0.6,0.4]*
MLP	*hidden_layer_sizes* *activation* *alpha* *solver* *learning_rate*	*10* *logistic* *0.05* *lbfgs* *constant*	*10* *logistic* *0.05* *lbfgs* *constant*	*10* *relu* *0.0001* *lbfgs* *constant*	*10* *logistic* *0.0001* *lbfgs* *constant*

Tabla 24. Pesos de los clasificadores y umbral (*threshold*) óptimo para los conjuntos de datos *Implantes Dentales*, *Artificial*, *Heart Disease* y *Breast Cancer*.

Conjunto de Datos	RF	SVC	KNN	NB	MLP	Accuracy	Threshold	TN
Implantes Dentales	0,90	0,60	0,80	0,70	0,70	0,93	0,50	0,79
	0,90	0,70	0,80	0,70	0,60		0,50	
	0,90	1,00	1,00	0,70	0,70		0,60	
	1,00	0,60	0,70	0,60	0,70		0,50	
	1,00	0,60	0,70	0,70	0,70		0,50	
	1,00	0,70	0,70	0,70	0,60		0,50	
	1,00	0,80	1,00	0,80	0,80		0,60	
	1,00	0,80	1,00	0,90	0,80		0,60	
	1,00	0,90	1,00	0,80	0,70		0,60	
	1,00	1,00	1,00	0,70	0,60		0,60	
	1,00	1,00	1,00	0,80	0,60		0,60	
Artificial	0,60	0,60	0,80	0,80	1,00	0,96	0,50	0,90
	0,60	0,60	0,80	0,90	1,00		0,50	
	0,60	0,70	0,70	0,80	1,00		0,60	
	0,60	0,70	0,70	0,90	1,00		0,50	
	0,60	0,90	0,70	0,60	0,80		0,50	
Heart Disease	0,60	1,00	0,80	1,00	0,60	0,79	0,60	0,94
	0,70	0,90	0,80	1,00	0,60		0,60	
	0,70	1,00	0,80	0,90	0,60		0,60	
	0,80	0,90	0,80	0,90	0,60		0,60	
	0,80	1,00	0,70	0,90	0,60		0,60	
	0,90	1,00	0,90	0,60	0,60		0,60	
Breast Cancer	0,70	0,60	1,00	1,00	0,80	0,73	0,50	0,53
	0,80	0,60	1,00	1,00	0,70		0,50	
	0,80	0,70	0,90	1,00	0,70		0,50	
	0,90	0,60	0,90	1,00	0,70		0,50	
	0,90	0,60	1,00	0,90	0,70		0,50	
	1,00	0,60	0,90	0,90	0,70		0,50	
	1,00	0,60	0,90	1,00	0,60		0,50	
	1,00	0,60	1,00	0,90	0,60		0,50	
	1,00	0,70	0,90	0,90	0,60		0,50	
	1,00	0,80	0,80	0,90	0,60		0,50	

Tabla 25. Eficiencia en el acierto de los clasificadores RF, SVC, KNN, MNB, MLP y el procedimiento propuesto (PP) sobre los conjuntos de datos *Implantes Dentales*, *Artificial*, *Heart Disease* y *Breast Cancer*.

Conjunto de Datos	**Clasificador**	*Clase Objetivo*	*Clase no-Objetivo*
		Sensitivity	**Specificity**
Implantes Dentales	RF	59 %	98 %
	SVC	64 %	**99 %**
	KNN	64 %	**99 %**
	MNB	72 %	79 %
	MLP	66 %	97 %
	PP	**79 %**	96 %
Artificial	RF	81 %	97 %
	SVC	81 %	**99 %**
	KNN	81 %	**99 %**
	MNB	60 %	81 %
	MLP	82 %	97 %
	PP	**90 %**	97 %
Heart Disease	RF	81 %	71 %
	SVC	70 %	**79 %**
	KNN	70 %	76 %
	MNB	77 %	74 %
	MLP	72 %	68 %
	PP	**94 %**	58 %
Breast Cancer	RF	36 %	78 %
	SVC	36 %	83 %
	KNN	20 %	**97 %**
	MNB	52 %	76 %
	MLP	32 %	80 %
	PP	**53 %**	81 %

En la Figura 38 se aprecia claramente el porcentaje de acierto de la clase minoritaria en la clasificación con los clasificadores RF, SVC, KNN, MNB, MLP y procedimiento propuesto (PP) sobre el conjunto de datos *Implantes Dentales*.

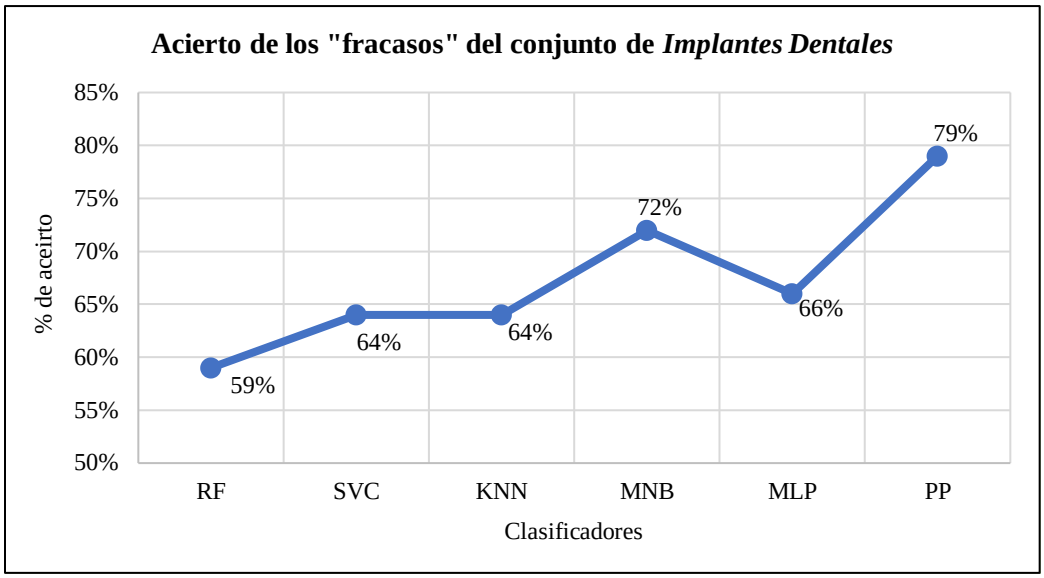
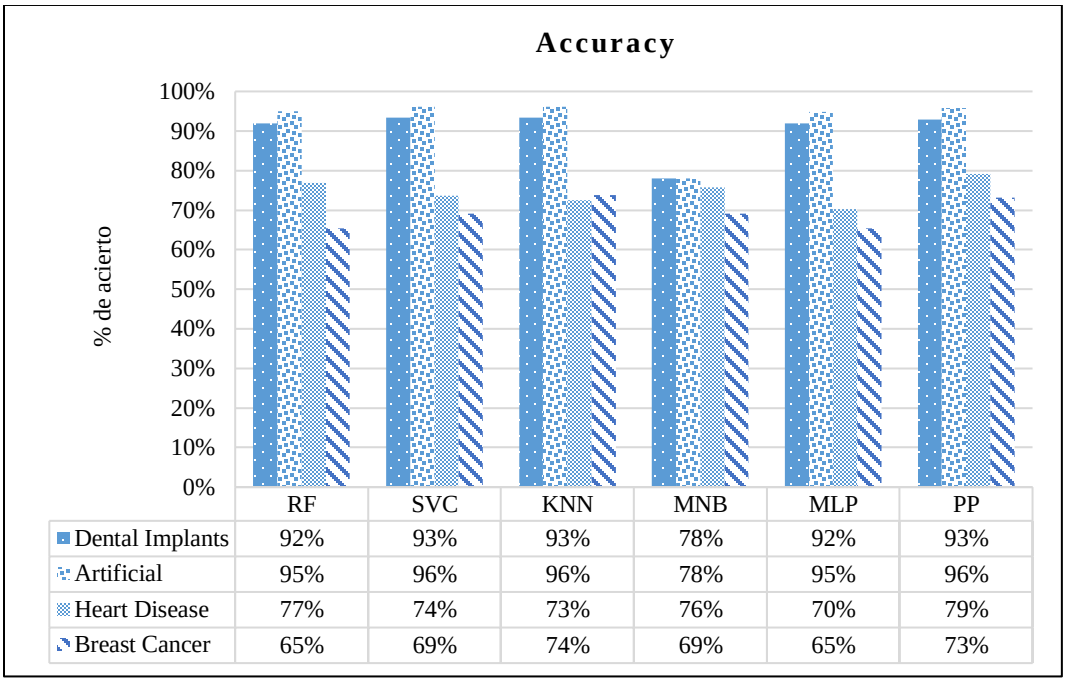


Figura 38. Porcentaje de acierto de la clase minoritaria (TN) en la clasificación con los clasificadores RF, SVC, KNN, MNB, MLP y procedimiento propuesto (PP) sobre el conjunto de datos *Implantes Dentales*.

En la Figura 39, se presenta el porcentaje de la métrica accuracy alcanzada para cada clasificador y el enfoque propuesto (PP) sobre los cuatro conjuntos de datos utilizados.



	RF	SVC	KNN	MNB	MLP	PP
Dental Implants	92%	93%	93%	78%	92%	93%
Artificial	95%	96%	96%	78%	95%	96%
Heart Disease	77%	74%	73%	76%	70%	79%
Breast Cancer	65%	69%	74%	69%	65%	73%

Figura 39. Accuracy de los clasificadores RF, SVC, KNN, MNB, MLP, así como el procedimiento propuesto (PP) sobre los conjuntos de datos *Implantes Dentales*, *Artificial*, *Heart Disease* y *Breast Cancer*.

En la Figura 39 se aprecia que los modelos SVC, KNN y el procedimiento propuesto (PP) fueron los de mejor desempeño sobre los conjuntos *Implantes Dentales* y *Artificial*. Así mismo, el enfoque propuesto logró la mejor precisión sobre el conjunto de datos *Heart Disease*. Mientras que sobre el conjunto de datos *Breast Cancer*, el procedimiento propuesto obtuvo un accuracy un poco inferior (1 centésimo) al obtenido por KNN, aunque es una buena precisión en comparación con los resultados de los demás clasificadores individuales.

4.2.1 VALIDACIÓN CON EXPERTOS HUMANOS

Para evaluar el rendimiento del procedimiento propuesto de integración de los clasificadores, se realiza una comparación con la opinión de cuatro expertos humanos. Éstos fueron seleccionados del *"Registro Provincial de Profesionales que practican Cirugía Buco Maxilofacial, Implantología, Periodoncia y Manipulación de Tejidos"* del Colegio de Odontólogos de la provincia de Misiones, Argentina.

La evaluación estuvo sujeta a la clasificación de casos de historias clínicas por estos cuatro expertos en el área. A cada uno de los cuales se les proporcionó una muestra aleatoria distinta del 10% de prevalencia de casos. Los casos se presentaron sin la etiqueta para que los expertos pudieran clasificarlos según su experiencia y así poder contrastar con los valores encontrados por nuestro enfoque de clasificación.

Finalmente, se comparó los resultados logrados por el enfoque propuesto de integración de las predicciones sobre el conjunto de datos *Implantes Dentales*, con la precisión lograda en la clasificación por los expertos humanos (Figura 40). El modelo propuesto logra un 93% de precisión total, con un error del 7%. Mientras que en promedio la clasificación realizada por los expertos, logra una precisión total del 87%, con un error promedio del 13% (Tabla 26).

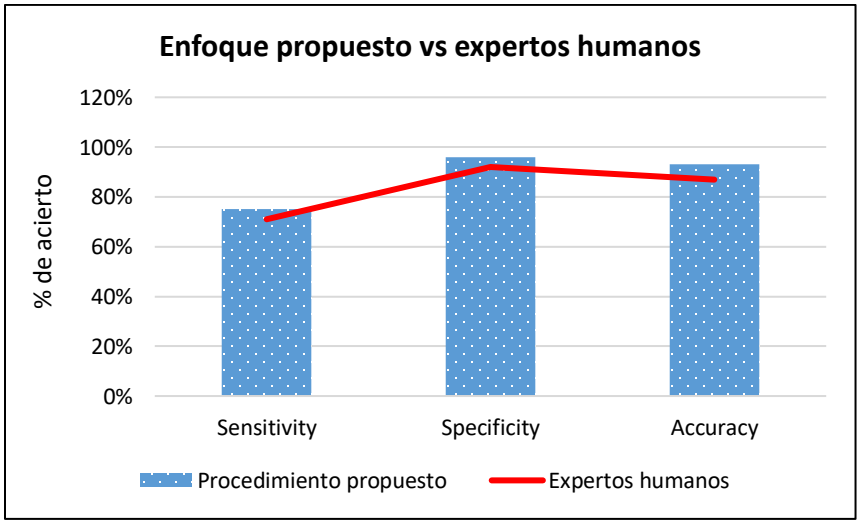


Figura 40. Valores de las métricas Sensitivity, Specificity y Accuracy logradas por el enfoque propuesto en comparación a la clasificación realizada por los expertos humanos.

Tabla 26. Comparación de los parámetros de evaluación logrados por el enfoque propuesto y la clasificación de los expertos sobre el conjunto de datos *Implantes Dentales*.

Modelo	Sensitivity	Specificity	Accuracy	Error
Enfoque propuesto	75 %	96 %	93 %	7 %
Expertos	71 %	92 %	87 %	13 %

La segunda parte de esta tesis tuvo la finalidad de aplicar múltiples clasificadores, para aumentar el acierto de los fracasos del conjunto de datos de historias clínicas, de pacientes que se han sometido a procesos quirúrgicos de colocación de implantes dentales en la Provincia de Misiones, Argentina. Se demostró que en este dominio es mejor integrar las predicciones de los clasificadores, para no sesgar la decisión sobre un solo resultado. Asimismo, utilizar predicciones integradas permite conocer diversos puntos de vistas o resultados para un mismo caso, ya que se utiliza más de un clasificador, permitiendo asegurar una asignación de etiqueta o clasificación más precisa.

El enfoque propuesto fue validado con un conjunto de datos artificiales generados para el estudio de caso y otros dos conjuntos de datos de prueba.

Mediante la experimentación realizada sobre el conjunto de datos original de *Implantes Dentales*, el enfoque propuesto logró el mejor porcentaje de acierto de la clase objetivo (fracasos), en comparación con el rendimiento de los clasificadores de forma individual y la clasificación realizada por los expertos humanos.

Los expertos consultados en patologías bucales y rehabilitación compleja en implantología oral, de distintos puntos de la provincia de Misiones, Argentina, coincidieron y remarcaron que en este campo de estudio es menos delicado etiquetar un caso como fracaso, que etiquetarlo como éxito cuando era un eventual fracaso.

Como resultado, cada clasificador logró hasta un 72% de acierto de la clase objetivo del conjunto de datos *Implantes Dentales* (Tabla 25) y el experto humano un 71% (Tabla 26), mientras que el enfoque propuesto (PP) permitió alcanzar el 75% de casos correctamente identificados como fracasos (Tabla 25).

Los clasificadores SVC y KNN lograron el mejor rendimiento sobre la clase no objetivo para todos los conjuntos de datos en comparación con los demás clasificadores, incluso superan al enfoque propuesto. Para la clase objetivo, se apreció que el enfoque propuesto permitió alcanzar el mayor porcentaje de acierto y la menor tasa de error para todos los casos.

5. CONCLUSIONES

Este trabajo de tesis permitió lograr la creación de un registro novedoso de historias clínicas de pacientes, que se han sometido a procesos quirúrgicos de colocación de implantes dentales en la Provincia de Misiones, Argentina. Además, permitió estudiar la aplicación de múltiples métodos de Ciencia de Datos en un ámbito de poco conocimiento.

En base a los objetivos planteados y al desarrollo de la presente Tesis Doctoral, se pudieron extraer las siguientes conclusiones:

- Se logró caracterizar los distintos tipos de tratamientos de superficie de los implantes dentales. Esta característica fue central a la hora de analizar y definir las variables de mayor ganancia de información para el estudio de caso.
- Se logró proponer un modelo de aprendizaje automático para identificar y descartar las características redundantes e irrelevantes en conjuntos de datos desbalanceados.
- Se detectaron los factores que ejercen una mayor influencia en el proceso de osteointegración (tejido óseo / implante dental) en base al conjunto de datos utilizados y a través de la validación con expertos humanos (especialistas Implantólogos y Odontólogos). Estos factores son: ingesta de medicamentos, ocupación del paciente, antecedentes médicos, tratamiento de superficie, pieza dental, tipo de hueso, longitud del implante, estación del año, tiempo de colocación, rango de edad del paciente, diámetro del implante, conexión, periodontitis, complicación quirúrgica, registro (habilidad del cirujano) y regeneración de los tejidos blandos.
- El procedimiento propuesto para la selección de características permitió conocer los atributos más relevantes del conjunto de datos y mejorar el rendimiento en la predicción de los fracasos, sin descuidar el porcentaje de precisión para la clase de éxito, en comparación con los métodos individuales de selección de características utilizados (IG, GR, RFI, R y ChiS).
- Cuando se seleccionan características de un conjunto de datos, el proceso debe incluir el análisis y aplicación de varios métodos de selección de características, para no sesgar la decisión en función de los resultados de un solo método. Por ello, es conveniente utilizar una combinación de métodos para tener una visión más certera.
- Es importante tener en cuenta que no siempre es bueno reducir el conjunto de características al mínimo posible, puesto que se pueden descartar características sustanciales para el estudio y reducir la eficiencia en la predicción. Además, en términos de costo y rendimiento de cálculo, no siempre es bueno seleccionar todas las características posibles, debido a que en ocasiones introduce ruido y no proporciona ganancia de información.

- Se puede argumentar que, en el caso de conjuntos de datos desequilibrados, la tasa de verdaderos negativos (TN) y la exactitud equilibrada (bac) son naturalmente las mejores medidas que se pueden buscar y medir, ya que benefician a la clase minoritaria, que es la que a menudo se intenta predecir, como lo han afirmado varios investigadores [181], [203], [256], [257].
- Conjuntamente, se apreció que en la validación del procedimiento propuesto de selección de características, el clasificador RF ha obtenido un muy buen desempeño en la clasificación para todos los conjuntos de datos utilizados, logrando buenos resultados de precisión, en comparación de los clasificadores SVM y KNN. Esto puede depender en gran medida de la cantidad de datos, tipo y otras cuestiones relacionados al conjunto de datos.
- Se logró proponer un modelo de aprendizaje automático mediante la aplicación de múltiples clasificadores, para mejorar el rendimiento de predicción en conjuntos de datos desbalanceados.
- El enfoque de múltiple clasificadores permitió mejorar el rendimiento de predicción de los fracasos del conjunto de datos de *Implantes Dentales*, así como de la clase minoritaria de los conjuntos de validación utilizados.
- En base a los resultados de la clasificación por parte de los expertos humanos, se puede alegar que el enfoque propuesto de integración de las predicciones permitió lograr un rendimiento de clasificación superior. Por lo tanto, se logró proponer un procedimiento de extracción de conocimiento validado por expertos humanos sobre un dominio de poco conocimiento.

6. TRABAJOS FUTUROS

Evaluar la posibilidad de incorporar una nueva dimensión al conjunto de datos relacionada al análisis del agua, ya que esta variable puede ser de interés para el estudio de caso.

Validar el enfoque propuesto de selección de características, así como el de integración de las predicciones sobre otros conjuntos de datos del área de la salud o la medicina. Además, se podría proponer la aplicación de alguna ponderación a los métodos de selección de características utilizados en función de su rendimiento, para evaluar la posibilidad de ajustar la tasa de precisión de las etiquetas de clases. Conjuntamente, se podría evaluar la inclusión o ampliación de los clasificadores utilizados, para mejorar el porcentaje de acierto.

Estudiar con más detalle el tratamiento de superficie de los implantes dentales. Así como, extender el relevamiento de casos de historias clínicas de implantes dentales a otras partes del territorio nacional e internacional.

Desarrollar e implementar un sistema de apoyo a la toma de decisión empleando Lógica Difusa, con la finalidad de proporcionar a los especialistas implantólogos, una alternativa a la hora de analizar situaciones poco regulares o complejas. Permitiendo modelar y analizar de ante mano, en base a determinadas especificaciones, cual podría ser el resultado postoperatorio de la intervención quirúrgica del Implante Dental. Evaluando el beneficio e impacto social en los especialistas Odontólogos del Nordeste argentino, al disponer de un asistente virtual que los ayude a evaluar y determinar las condiciones específicas del paciente y la técnica más adecuada a utilizar, en cada caso particular.

7. PRODUCCIÓN CIENTÍFICA

7.1 PRESENTACIONES A CONGRESOS

[1] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Sistema de clasificación para predicción de fracasos en implantes dentales validado por expertos humanos," in *Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA) de las Jornadas Argentinas de Informática (49 JAIIO - SADIO)*. Buenos Aires, Argentina. 2020.

[2] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Predicción del resultado de oseointegración en implantes dentales mediante múltiples clasificadores," in *XXII Workshop de Investigadores en Ciencias de la Computación (WICC)*, RedUNCI, ISBN 978-987-3714-82-5., pp. 180–184. El Calafate, Santa Cruz, Argentina. 2020.

[3] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Predicción del resultado de oseointegración en implantes dentales mediante múltiples clasificadores," in *XXII Workshop de Investigadores en Ciencias de la Computación (WICC)*, RedUNCI, ISBN 978-950-34-1906-9, (Poster). El Calafate, Santa Cruz, Argentina. 2020.

[4] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Evaluación de la precisión de acierto de un conjunto desbalanceado mediante la combinación de clasificadores," in *XXV Congreso Argentino de Ciencias de la Computación (CACIC)*, ISBN 978-987-688-377-1, pp. 497–506. Río Cuarto, Córdoba, Argentina. 2019.

[5] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Evaluación de características de implantes dentales para su codificación en tipos de tratamientos de superficie," in *19° Congreso Internacional de Metalurgia y Materiales (CONAMET- SAM)*. Valdivia, Chile. 2019.

[6] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Aplicación de un modelo de aprendizaje automático para la predicción de fracasos en implantes dentales," in *4º Congresso de Engenharia e Ciências Aplicadas nas Três Fronteiras (MEC3F)*, ISSN 2675-4452, p. 27. Foz do Iguaçu, Brasil. 2019.

[7] N. B. Ganz, F. A. Domínguez, A. E. Ares, and H. D. Kuna, "Selección de características mediante la combinación de métodos para evaluar la precisión de clasificación en un conjunto de datos de implantes dentales," in *XXI Workshop de Investigadores en Ciencias de la Computación (WICC)*, RedUNCI, ISBN 978-987-3619-27-4, pp. 263–267. San Juan, Argentina. 2019.

[8] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Análisis de set de datos de implantes dentales aplicando técnicas de minería de datos," in *Jornada Científico – Tecnológica (JCT)*, ISBN 978-950-579-495-9, p. 269. Misiones, Argentina. 2018.

[9] N. B. Ganz, F. A. Domínguez, A. E. Ares, and H. D. Kuna, "Avances en Selección de Biomateriales utilizados en Implantes Dentales aplicando Técnicas de Minería de Datos," in *XX Workshop de Investigadores en Ciencias de la Computación (WICC)*, RedUNCI, ISBN 978-987-3619-27-4, pp. 209–213. Corrientes, Argentina. 2018.

[10] N. B. Ganz, A. E. Ares, and H. D. Kuna, “Selección de biomateriales para implantes dentales utilizando la minería de datos,” in *103ª Reunión de la Asociación Física Argentina (AFA)*, p. 225. Buenos Aires, Argentina. 2018.

[11] N. B. Ganz, A. E. Ares, and H. D. Kuna, “Aplicación de la Minería de Datos para la Selección de Biomateriales en Implantes Dentales,” in *18º Congreso Internacional de Metalurgia y Materiales (SAM-CONAMET)*, ISBN 978-987-1323-62-3, pp. 1061–1063. San Carlos de Bariloche, Argentina. 2018.

[12] N. B. Ganz, H. D. Kuna, and A. E. Ares, “Selección de Biomateriales utilizados en Implantes Dentales aplicando Técnicas de Minería de Datos,” in *XIX Workshop de Investigadores en Ciencias de la Computación (WICC)*, RedUNCI, ISBN 978-987-3619-27-4, pp. 344–348. Buenos Aires, Argentina. 2017.

7.2 PUBLICACIONES

[1] N. B. Ganz, A. E. Ares, and H. D. Kuna, “Detection of failure factors in dental implants by using multiple feature selection methods,” *Revista Perspectivas em Ciência da Informação*, ISSN 1981-5344. 2020. (En revisión).

[2] N. B. Ganz, A. E. Ares, and H. D. Kuna, “Clinical histories dataset of patients undergone surgical processes of dental implant placement in the province of Misiones, Argentina,” *Data in Brief*, ISSN 2352-3409. 2020. (En revisión).

[3] N. B. Ganz, A. E. Ares, and H. D. Kuna, “Predicting dental implant failures by integrating multiple classifiers,” *Revista de Ciencia y Tecnología*, ISSN 1851-7587, vol. 34, pp. 13–23. 2020.

[4] N. B. Ganz, A. E. Ares, and H. D. Kuna, “Aplicación de la minería de datos para la selección de biomateriales en implantes dentales,” *Revista SAM*, ISSN 1668-4788, vol. 1, pp. 40–44. 2019.

8. FINANCIAMIENTO

Esta tesis fue financiado por el *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) a través de una “Beca Interna Doctoral” otorgada por Resolución D Nº 4869.

Proyectos de Investigación en los que se enmarcó la Tesis

Proyecto Incentivado: ALEACIONES BASE ALUMINIO, ZINC Y ESTAÑO: OBTENCIÓN POR SOLIDIFICACIÓN DIRECCIONAL Y EVALUACIÓN DE PROPIEDADES. Código: 16/Q548. Directora: Alicia E. Ares. Aprobado por Resolución N°

207/14 del Consejo Directivo de fecha 13/06/2014. Facultad de Ciencias Exactas, Químicas y Naturales. Universidad Nacional de Misiones. Programa Nacional de Incentivos a Docentes Investigadores. Período de Desarrollo: 01/01/2014 al 31/12/2016.

Proyecto Incentivado: OBTENCIÓN DE ALEACIONES BASE ALUMINIO Y ÓXIDOS PARA APLICACIONES TECNOLÓGICAS. Código: 16/Q628. Directora: Alicia E. Ares. Aprobado por Resolución N° 415/17 del Consejo Directivo de fecha 28/08/2017. Facultad de Ciencias Exactas, Químicas y Naturales. Universidad Nacional de Misiones. Programa Nacional de Incentivos a Docentes Investigadores. Período de Desarrollo: 01/01/2017 al 31/12/2019.

Proyecto Incentivado: PRODUCCIÓN Y CARACTERIZACIÓN DE ALEACIONES LIGERAS Y RECUBRIMIENTOS MICRO/NANO ESTRUCTURADOS. Código: 16/Q1225-PI. Directora: Alicia E. Ares. Aprobado por Resolución N° 219/20 del Consejo Directivo. Facultad de Ciencias Exactas, Químicas y Naturales. Universidad Nacional de Misiones. Programa Nacional de Incentivos a Docentes Investigadores. Período de Desarrollo: 01/01/2020 al 31/12/2023.

Subsidios con los que se financió el desarrollo de la Tesis

Proyecto ANPCyT: OBTENCIÓN, CARACTERIZACIÓN Y PROPIEDADES DE ALEACIONES DE TITANIO PARA LA SUSTITUCIÓN DE TEJIDOS DUROS. Código: PICT-E-2014-0170. Investigadora Responsable: Alicia E. Ares. Aprobado por Resolución Nº 472/14 del 08/09/2014. Agencia Nacional de Promoción Científica y Tecnológica - Universidad Nacional de Misiones. Período de Desarrollo: 01/10/2014 al 22/04/2016.

Proyecto ANPCyT: OBTENCIÓN DE ALEACIONES BASE ALUMINIO CON DIFERENTES ESTRUCTURAS DE GRANOS Y RECUBRIMIENTOS NANOESTRUCTURADOS DE Al2O3 PARA DISTINTAS APLICACIONES TECNOLÓGICAS. Código: PICT-2017-0079. Investigadora Responsable: Alicia E. Ares. Aprobado por Resolución Nº 310/18 del 08/06/2018. Agencia Nacional de Promoción Científica y Tecnológica. Universidad Nacional de Misiones. Período de Desarrollo: 10/12/2018 al 10/12/2021.

9. REFERENCIAS

[1] L. Cao, "Data Science : A Comprehensive Overview," *ACM Comput. Surv.*, vol. 50, no. 3, pp. 1–42, 2017.

[2] J. E. B. Tamez, F. N. Zilli, L. A. Fandiño, and J. M. Guizar, "Factores relacionados con el éxito o el fracaso de los implantes dentales colocados en la especialidad de Prostodoncia e Implantología en la Universidad de La Salle Bajío," *Revi. Esp. Cirugía Oral y Maxilofac.*, vol. 286, pp. 1–9, 2016.

[3] J. Domínguez, J. Acuña, M. Rojas, J. Bahamondes, and S. Matus, "Study of association between systemic diseases and dental implant failure," *Rev. Clínica Periodoncia, Implantol. y Rehabil. Oral*, vol. 6, no. 1, pp. 9–13, 2013.

[4] A. L. I. Oliveira, C. Baldisserotto, and J. Baldisserotto, "A comparative study on machine learning techniques for prediction of success of dental implants," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3789 LNAI, pp. 939–948, 2005.

[5] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[6] B. Irie and Sei Miyake, "Capabilities of Three-layered Perceptrons," *IEEE nternational Conf. Neural Networks*, pp. 641–648, 1988.

[7] R. S. Moayeri, M. Khalili, and M. Nazari, "A Hybrid Method to Predict Success of Dental Implants," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 1–6, 2016.

[8] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: A review," *Procedia Manuf.*, vol. 35, pp. 698–703, 2019.

[9] F. Harrou, A. Zeroual, and Y. Sun, "Traffic congestion monitoring using an improved kNN strategy," *Meas. J. Int. Meas. Confed.*, vol. 156, p. 107534, 2020.

[10] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, 2018.

[11] A. C. Braga, P. Vaz, J. C. Sampaio-Fernandes, A. Felino, and M. P. Tavares, "Decision model to predict the implant success," *Proc. 12th Int. Conf. Comput. Sci. Its Appl.*, vol. 7333 LNCS, no. PART 1, pp. 665–674, 2012.

[12] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, 2014.

[13] J. D. Kelleher and B. Tierney, "What is Data Science?," in *Data science*, London, England: The MIT Press, 2018, pp. 1–38.

[14] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, 2018.

[15] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework.,” *Int Conf Knowl. Discov. Data Min.*, pp. 82–88, 1996.

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “The KDD Process for Extracting Useful Knowledge from Volumes of Data,” *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[18] H. Kuna, R. G. Martinez, and F. Villatoro, “Automatic Outliers Fields Detection in Databases,” *J. Model. Simul. Syst. HyperSciences Publ.*, vol. 3, no. 1, pp. 14-20., 2012.

[19] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, “Benchmark for filter methods for feature selection in high-dimensional classification data,” *Comput. Stat. Data Anal.*, vol. 143, p. 106839, 2020.

[20] U. Fyyaad, G. Piatetsky-Shapiro, and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework,” *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, pp. 82–88, 1996.

[21] U. Fayyad, “Knowledge Discovery in Databases: An Overview,” in *Relational Data Mining*, 2001, pp. 28–47.

[22] S. Sethi, D. Malhotra, and N. Verma, “Data Mining : Current Applications & Trends,” *Int. J. Innov. Eng. Technol.*, vol. 6, no. 4, pp. 667–673, 2016.

[23] Mehmed Kantardzic, “Introduction to Data Mining,” in *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd. ed., Wiley-IEEE Press, 2011, pp. 1–10.

[24] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2012.

[25] O. Simeone, “A Very Brief Introduction to Machine Learning with Applications to Communication Systems,” *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 648–664, 2018.

[26] S. Fathi, M. Ahmadi, B. Birashk, and A. Dehnad, “Development and use of a clinical decision support system for the diagnosis of social anxiety disorder,” *Comput. Methods Programs Biomed.*, vol. 190, no. 4, p. 105354, 2020.

[27] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018.

[28] G. Pitolli, L. Aniello, G. Laurenza, L. Querzoni, and R. Baldoni, “Malware family identification with BIRCH clustering,” *Proc. - Int. Carnahan Conf. Secur. Technol.*, vol. 2017-Octob, pp. 1–6, 2017.

[29] Y. Min and Y. Li, “Vehicles recognition based on the size characteristics and the CURE clustering algorithm,” in *2015 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2015*, 2015.

[30] J. Jiang, G. Ananthanarayanan, P. Bodik, S. Sen, and I. Stoica, “Chameleon : Scalable Adaptation of Video Analytics,” *SIGCOMM ’18 Proc. 2018 Conf. ACM Spec. Interes.*

Gr. Data Commun., pp. 253–266, 2018.

[31] A. K. Dubey, U. Gupta, and S. Jain, "Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 1, pp. 18–29, 2018.

[32] E. Schubert and P. J. Rousseeuw, "Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11807 LNCS, pp. 171–187, 2019.

[33] M. Daoudi and S. Meshoul, "Revisiting BFR Clustering Algorithm for Large Scale Gene Regulatory Network Reconstruction using MapReduce," in *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*, 2017.

[34] N. Price-Jones and J. Bovy, "Blind chemical tagging with DBSCAN: prospects for spectroscopic surveys," *Mon. Not. R. Astron. Soc.*, vol. 487, no. 1, pp. 871–886, 2019.

[35] O. Metaxas, H. Dimitropoulos, Y. Ioannidis, and M. Paedigree, "AITION: A scalable KDD platform for Big Data Healthcare," *2014 IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2014*, vol. 600932, pp. 601–604, 2014.

[36] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection using Gain Ratio and Correlation based Feature Slection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[37] H. K. Han, H. S. Kim, and S. Y. Sohn, "Sequential association rules for forecasting failure patterns of aircrafts in Korean airforce," *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 1129–1133, 2009.

[38] C. Rudin, B. Letham, A. Salleb-Aouissi, E. Kogan, and D. Madigan, "Sequential event prediction with association rules," *J. Mach. Learn. Res.*, vol. 19, pp. 615–634, 2011.

[39] P. Lin, K. Ye, M. Chen, and C. Xu, "DCSA : Using Density-Based Clustering and Sequential Association Analysis to Predict Alarms in Telecommunication Networks," *2019 IEEE 25th Int. Conf. Parallel Distrib. Syst.*, 2019.

[40] Adamo Jean-Marc, "Sequential and Parallel Algorithms," in *Data Mining for Association Rules and Sequential Patterns*, Springer-Verlag New York, 2001.

[41] R. Devika, C. Koushik, and V. Subramaniyaswamy, "An Event Detection on Twitter using ECLAT (Equivalence Class Transformation) algorithm with TRCM (Transaction based Rule Change Mining)," *Int. J. Pure Appl. Math.*, vol. 119, no. 12, pp. 13347–13355, 2018.

[42] B. Richhariya and M. Tanveer, "EEG signal classification using universum support vector machine," *Expert Syst. Appl.*, vol. 106, pp. 169–182, 2018.

[43] C. D. Manning, P. Raghavan, and H. Schutze, "Text classification and Naive Bayes," in *Introduction to Information Retrieval*, Cambridge University Press, 2009, pp. 253–287.

[44] K. Bhattacharjee and M. Pant, "Hybrid Particle Swarm Optimization-Genetic

Algorithm trained Multi-Layer Perceptron for Classification of Human Glioma from Molecular Brain Neoplasia Data," *Cogn. Syst. Res.*, vol. 58, pp. 173–194, 2019.

[45] L. Gao, M. Ye, X. Lu, and D. Huang, "Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification," *Genomics, Proteomics Bioinforma.*, vol. 15, no. 6, pp. 389–395, 2017.

[46] M. Antonie, A. Coman, and O. R. Zaiane, "Application of Data Mining Techniques for Medical Image Classification," *Proc. Second Int. Work. Multimida Data Min.*, pp. 94–101, 2001.

[47] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," *Appl. Soft Comput. J.*, vol. 50, pp. 135–141, 2017.

[48] M. Abbas, K. Ali Memon, A. Aleem Jamali, S. Memon, and A. Ahmed, "Multinomial Naive Bayes Classification Model for Sentiment Analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 3, pp. 62–67, 2019.

[49] W. Chen *et al.*, "GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models," *Sci. Total Environ.*, vol. 634, pp. 853–867, 2018.

[50] W. Chen, X. Yan, Z. Zhao, H. Hong, D. T. Bui, and B. Pradhan, "Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression , naive Bayes and RBFNetwork models for the Long County area (China)," *Bull. Eng. Geol. Environ.*, vol. 78, pp. 247–266, 2018.

[51] D. Tien Bui, T. A. Tuan, H. Klempe, B. Pradhan, and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides*, vol. 13, no. 2, pp. 361–378, 2016.

[52] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. USA, 2011.

[53] Z. Ceylan, "Assessment of agricultural energy consumption of Turkey by MLR and Bayesian optimized SVR and GPR models," *J. Forecast.*, vol. 39, no. 6, pp. 944–956, 2020.

[54] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 2005.

[55] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *SIGMOD '96 Proc. 1996 ACM SIGMOD Int. Conf. Manag. data*, vol. 25, no. 2, pp. 103–114, 1996.

[56] S. Guha, R. Rastogi, and K. Shim, "CURE : An Efficient Clustering Algorithm for Large Databases," *SIGMOD Rec.*, vol. 27, no. 2, pp. 73–84, 1998.

[57] G. Karypis, E.-H. (Sam) Han, and V. Kumar, "Chameleon: Hierarchical Clustering Using Dynamic Modeling," *Computer (Long. Beach. Calif).*, vol. 32, no. 8, pp. 68–75, 1999.

[58] S. Guha, R. Rastogi, and K. Shim, “Rock: a robust clustering algorithm for categorical attributes,” *Proc. 15th Int. Conf. Data Eng.*, pp. 512–521, 1999.

[59] A. Patidar, R. Joshi, and S. Mishra, “Implementation of distributed ROCK algorithm for clustering of large categorical datasets and its performance analysis,” in *2011 3rd International Conference on Electronics Computer Technology*, 2011, pp. 79–83.

[60] Y. Li, Q. Huang, S. Liu, and P. Liu, *Load Pattern Analysis of Key Accounts based on Two- step Clustering*. New York, NY, USA: Association for Computing Machinery, 2016.

[61] J. MacQUEEN, “Some methods for classification and analysis of multivariate observations,” in *5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 233, no. 233, pp. 281–297.

[62] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.

[63] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *KDD-96*, 1996, pp. 226–231.

[64] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, “OPTICS : Ordering Points To Identify the Clustering Structure,” *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, 1999.

[65] A. Hinneburg and D. A. Keim, “An efficient approach to clustering in large multimedia databases with noise,” in *KDD’98: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998, vol. 5, no. 4, pp. 58–65.

[66] E. A. Patrick, “Clustering Using a Similarity Measure Based on Shared Near Neighbors,” *IEEE Trans. Comput.*, vol. C–22, no. 11, pp. 1025–1034, 1973.

[67] T. N. Tran, R. Wehrens, and L. M. C. Buydens, “Knn Density-Based Clustering for High Dimensional Multispectral Images,” in *2003 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, 2003, pp. 147–151.

[68] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” in *94: Proceedings of the 20th International Conference on Very Large Data Bases*, 1994, pp. 487–499.

[69] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach,” *Data Min. Knowl. Discov.*, vol. 8, pp. 53–87, 2004.

[70] G. I. Webb, “Efficient search for association rules,” in *KDD ’00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 99–107.

[71] M. J. Zaki and C.-J. Hsiao, “CHARM : An Efficient Algorithm for Closed Itemset Mining,” *SDM*, pp. 457–473, 2002.

[72] J. Pei, J. Han, and R. Mao, "CLOSET : An Efficient Algorithm for Mining Frequent Closed Itemsets," *DMKD*, no. 1, pp. 1–10, 2000.

[73] J. Schwarz, "Correlation Coefficients According to Bravais-Pearson, Spearman, and Kendall," *Intell. Instruments Comput.*, no. June, pp. 114–126, 1987.

[74] B. Schölkopf and A. J. Smola, "Support Vector Machines and Kernel Algorithms," *Handb. Brain Theory Neural Networks*, pp. 1119–1125, 2002.

[75] K. R. Pradeep and N. C. Naveen, "Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics," *Procedia Comput. Sci.*, vol. 132, pp. 412–420, 2018.

[76] W. Xiaohu, W. Lele, and L. Nianfeng, "An Application of Decision Tree Based on ID3," *Phys. Procedia*, vol. 25, pp. 1017–1021, 2012.

[77] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., vol. 12. Morgan Kaufmann, 2011.

[78] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[79] E. Alfaro, M. Gamez, and N. García, "adabag : An R Package for Classification with Boosting and Bagging," *J. Stat. Softw.*, vol. 54, no. 2, 2013.

[80] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.

[81] J. Chaki, N. Dey, L. Moraru, and F. Shi, "Fragmented plant leaf recognition: Bag-of-features, fuzzy-color and edge-texture histogram descriptors with multi-layer perceptron," *Optik (Stuttg).*, vol. 181, no. December 2018, pp. 639–650, 2019.

[82] T. Zarei and R. Behyad, "Predicting the water production of a solar seawater greenhouse desalination unit using multi-layer perceptron model," *Sol. Energy*, vol. 177, no. October 2018, pp. 595–603, 2019.

[83] T. Sarkar and M. Mishra, "Soil Erosion Susceptibility Mapping with the Application of Logistic Regression and Artificial Neural Network," *J. Geovisualization Spat. Anal.*, vol. 2, no. 1, pp. 2–17, 2018.

[84] A. Borucka, "Application of the Logistic Regression Model to Study Customer Loyalty in an Online Store," in *ICBIM '19: Proceedings of the 3rd International Conference on Business and Information Management*, 2019, pp. 21–26.

[85] W. Yu and C. Zhao, "Sparse Exponential Discriminant Analysis and Its Application to Fault Diagnosis," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5931–5940, 2018.

[86] N. Metawa, M. K. Hassan, and M. Elhoseny, "Genetic algorithm based model for optimizing bank lending decisions," *Expert Syst. Appl.*, vol. 80, pp. 75–82, 2017.

[87] H. Lingaraj, "A Study on Genetic Algorithms and its Applications," *Int. J. Comput. Sci. Eng.*, vol. 4, no. 10, pp. 139–143, 2019.

[88] T. Therneau, B. Atkinson, and B. Ripley, "rpart: Recursive Partitioning and Regression Trees," *R Packag. version*, 2019.

[89] G. D. Merkel, R. J. Povinelli, and R. H. Brown, "Short-term load forecasting of natural gas with deep neural network regression," *Energies*, vol. 11, no. 8, 2018.

[90] L. K. Song, G. C. Bai, and C. W. Fei, "Probabilistic LCF life assessment for turbine discs with DC strategy-based wavelet neural network regression," *Int. J. Fatigue*, vol. 119, no. October 2018, pp. 204–219, 2019.

[91] Y. He, Y. Qin, S. Wang, X. Wang, and C. Wang, "Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network," *Appl. Energy*, vol. 233–234, no. October 2018, pp. 565–575, 2019.

[92] Y. Li, W. Zheng, Z. Cui, and T. Zhang, "Face recognition based on recurrent regression neural network," *Neurocomputing*, vol. 297, pp. 50–58, 2018.

[93] O. Maimon and L. Rokach, *The Data Mining and Knowledge Discovery Handbook*. Publishers, Springer Science + Business Media, 2005.

[94] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step data mining guide," 2000.

[95] S. V. Europe, "CRISP-DM Methodology," 2015. [Online]. Available: http://crisp-dm.eu/home/crisp-dm-methodology/.

[96] N. W. Grady, "Knowledge Discovery in Data Science KDD meets Big Data," *IEEE Int. Conf. Big Data*, pp. 1603–1608, 2016.

[97] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: JOHN WILEY & SONS, 2005.

[98] R. Figueroa-diaz, S. Chamba, R. Guaman-quinche, and M. Cueva-hurtado, "Mapas Auto-Organizados aplicados a la segmentación de clientes en entornos empresariales," *Revista Tecnológica ESPOL*, vol. 29, pp. 130–143, 2016.

[99] Ó. Marbán, G. Mariscal, and J. Segovia, "A Data Mining & Knowledge Discovery Process Model," in *Data Mining and Knowledge Discovery in Real Life Applications*, no. February, I-Tech, Ed. Vienna, Austria: Julio Ponce y Adem Karahoca, 2009, pp. 1–17.

[100] Microsoft, "What is the Team Data Science Process?," *Team Data Science Process*, 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview. [Accessed: 05-Oct-2020].

[101] Microsoft, "The Team Data Science Process: Lifecycle," *The Team Data Science Process*, 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle. [Accessed: 05-Oct-2020].

[102] K. Schwaber and M. Beedle, *Agile Software Development with Scrum*. 2002.

[103] IBM Analytics, "Analytics Solutions Unified Method (ASUM)." pp. 1–3, 2016.

[104] D. Nogueira, "Agile Data Mining: Uma metodologia ágil para o desenvolvimento de

projetos de data mining," Faculdade de Engenharia da Universidade Do Porto, 2014.

[105] K. Schwaber and J. Sutherland, "The Scrum Guide," no. November. pp. 1–19, 2017.

[106] SAS Institute Inc., "Introduction to SEMMA," 2020. [Online]. Available: https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=15.1&locale=en. [Accessed: 09-Jul-2020].

[107] SAS Institute Inc., "Data Mining and SEMMA," 2020. [Online]. Available: https://documentation.sas.com/?docsetId=emcs&docsetTarget=n0pejm83csbja4n1xueveo2uoujy.htm&docsetVersion=14.3&locale=en. [Accessed: 09-Jul-2020].

[108] Dorian Pyle, *Business Modeling and Data Mining*. San Francisco, 2003.

[109] G. Fois, G. A. Agüero Crovella, and P. V. Britos, "Evaluación comparativa de las metodologías Team Data Science Process TDSP y Analytics Solutions Unified Method for Data Mining ASUM-DM desde la perspectiva de la ciencia de datos," in *Investigación Formativa en Ingeniería*, 4a. ed., Medellín, Antioquia, Colombia: Editorial Instituto Antioqueño de Investigación, 2020, pp. 264–270.

[110] E. Rivo, J. De La Fuente, Á. Rivo, E. García-Fontán, M. Á. Cañizares, and P. Gil, "Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management," *Clin. Transl. Oncol.*, vol. 14, no. 1, pp. 73–79, 2012.

[111] F. Portela, M. F. Santos, J. Machado, A. Abelha, F. Rua, and Á. Silva, "Real-Time Decision Support Using Data Mining to Predict Blood Pressure Critical Events in Intensive Medicine Patients," *Ambient Intell. Heal.*, vol. 9456, pp. 77–79, 2015.

[112] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J. Healthc. Inf. Manag.*, vol. 19, no. 2, pp. 64–72, 2005.

[113] A. Viloria *et al.*, "Determinating student interactions in a virtual learning environment using data mining," *Procedia Comput. Sci.*, vol. 155, no. 2018, pp. 587–592, 2019.

[114] A. Morais, H. Peixoto, C. Coimbra, A. Abelha, and J. Machado, "Predicting the need of Neonatal Resuscitation using Data Mining," *Procedia Comput. Sci.*, vol. 113, pp. 571–576, 2017.

[115] A. Brandão, E. Pereira, F. Portela, M. F. Santos, A. Abelha, and J. Machado, "Managing Voluntary Interruption of Pregnancy Using Data Mining," *Procedia Technol.*, vol. 16, pp. 1297–1306, 2014.

[116] S. Pereira, F. Portela, M. F. Santos, J. Machado, and A. Abelha, "Predicting Type of Delivery by Identification of Obstetric Risk Factors through Data Mining," *Procedia Comput. Sci.*, vol. 64, pp. 601–609, 2015.

[117] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput. Biol. Med.*, vol. 112, no. July, p. 103375, 2019.

[118] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[119] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[120] F. Macedo, M. Rosário Oliveira, A. Pacheco, and R. Valadas, “Theoretical foundations of forward feature selection methods based on mutual information,” *Neurocomputing*, vol. 325, pp. 67–89, 2019.

[121] N. Dessì and B. Pes, “Similarity of feature selection methods: An empirical study across data intensive classification tasks,” *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4632–4642, 2015.

[122] N. Abe and M. Kudo, “Entropy Criterion for Classifier-Independent Feature Selection,” *Int. Conf. Knowledge-Based Intell. Inf. Eng. Syst.*, no. 1, pp. 689–695, 2005.

[123] U. M. Khaire and R. Dhanalakshmi, “Stability of feature selection algorithm: A review,” *J. King Saud Univ. - Comput. Inf. Sci.*, 2019.

[124] R. Zhang, F. Nie, X. Li, and X. Wei, “Feature selection with multi-view data: A survey,” *Inf. Fusion*, vol. 50, no. May 2018, pp. 158–167, 2019.

[125] P. Banerjee, “Comprehensive Guide on Feature Selection,” *kaggle*, 2020. [Online]. Available: https://www.kaggle.com/prashant111/comprehensive-guide-on-feature-selection. [Accessed: 20-Oct-2020].

[126] M. Bennasar, Y. Hicks, and R. Setchi, “Feature selection using Joint Mutual Information Maximisation,” *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8520–8532, 2015.

[127] J. R. Quinlan, “Induction of Decision Trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[128] S. S. Shreem, S. Abdullah, and M. Z. A. Nazri, “Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm,” *Int. J. Syst. Sci.*, vol. 47, no. 6, pp. 1312–1329, 2016.

[129] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” *J. Biomed. Inform.*, vol. 85, no. July, pp. 189–203, 2018.

[130] M. Robnik-Sikonja and I. Kononenko, “Theoretical and empirical analysis of ReliefF and RRiefF,” *J. Mach. Learn.*, vol. 53, pp. 23–69, 2003.

[131] K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 50, no. 302, pp. 157–175, 1900.

[132] A. Jovic, K. Brkic, and N. Bogunovic, “A review of feature selection methods with applications,” *38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron.*, no. May, pp. 1200–1205, 2015.

[133] U. Stańczyk, “Feature evaluation by filter, Wrapper and embedded approaches,” *Stud. Comput. Intell.*, vol. 584, pp. 29–44, 2015.

[134] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.

[135] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[136] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. CRC Press, 1984.

[137] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," *Adv. Neural Inf. Process. Syst. 26*, pp. 431–439, 2013.

[138] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Appl. Soft Comput. J.*, vol. 86, p. 105836, 2020.

[139] G. Biau, "Analysis of a Random Forests Model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063–1095, 2012.

[140] A. Chaudhary, S. Kolhe, and Rajkamal, "Performance Evaluation of feature selection methods for Mobile devices," *J. Eng. Res. Appl.*, vol. 3, no. 6, pp. 587–594, 2013.

[141] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992.

[142] M. Togaçar, B. Ergen, and Z. Cömert, "Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 23–39, 2020.

[143] M. Mohandes, M. Deriche, and S. O. Aliyu, "Classifiers Combination Techniques: A Comprehensive Review," *IEEE Access*, vol. 6, pp. 19626–19639, 2018.

[144] R. Susmaga, "Confusion Matrix Visualization," in *Intelligent Information Processing and Web Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 107–116.

[145] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[146] Z. Karimi, M. Mansour, R. Kashani, and A. Harounabadi, "Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods," *Int. J. Comput. Appl.*, vol. 78, no. 4, pp. 21–27, 2013.

[147] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Leaming*, vol. 20, no. 3, pp. 273–297, 1995.

[148] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, vol. 46, no. June 2018, pp. 147–170, 2019.

[149] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugosl. J. Oper. Res.*, vol. 21, no. 1,

pp. 119–135, 2011.

[150] T. Z. Phyu and N. N. Oo, “Performance Comparison of Feature Selection Methods,” *MATEC Web Conf.*, vol. 42, p. 06002, 2016.

[151] H. Dag, K. E. Sayin, I. Yenidogan, S. Albayrak, and C. Acar, “Comparison of Feature Selection Algorithms for Medical Data,” *Int. Symp. Innov. Intell. Syst. Appl.*, pp. 1–5, 2012.

[152] M. Peker, A. Arslan, B. Sen, F. V. Celebi, and A. But, “A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF),” *Int. Symp. Innov. Intell. Syst. Appl.*, 2015.

[153] R. Parimala and R. Nallaswamy, “A Study of Spam E-mail classification using Feature Selection package,” *Glob. J. Comput. Sci. Technol.*, vol. 11, no. 7, pp. 45–54, 2011.

[154] P. R., V. M.L., and S. S., “Gain Ratio Based Feature Selection Method for Privacy Preservation,” *ICTACT J. Soft Comput.*, vol. 01, no. 04, pp. 201–205, 2011.

[155] I. Sumaiya Thaseen and C. Aswani Kumar, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017.

[156] M. Moran and G. Gordon, “Curious Feature Selection,” *Inf. Sci. (Ny).*, vol. 485, pp. 42–54, 2019.

[157] F. Kamalov and F. Thabtah, “A Feature Selection Method Based on Ranked Vector Scores of Features for Classification,” *Ann. Data Sci.*, vol. 4, no. 4, pp. 483–502, 2017.

[158] C. B. Gokulnath and S. P. Shantharajah, “An optimized feature selection based on genetic approach and support vector machine for heart disease,” *Cluster Comput.*, vol. 22, pp. 14777–14787, 2018.

[159] Y. Lu, “Knowledge integration in a multiple classifier system,” *Appl. Intell.*, vol. 6, no. 2, pp. 75–86, 1996.

[160] L. I. Kuncheva, “Combining Pattern Classifiers: Methods and Algorithms,” in *Combining Pattern Classifiers*, 2nd ed., John Wiley & Sons, 2014, pp. 290–325.

[161] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, “Multiple classifiers in biometrics. part 1: Fundamentals and review,” *Inf. Fusion*, vol. 44, no. December 2017, pp. 57–64, 2018.

[162] Y. Miao, H. Jiang, H. Liu, and Y. dong Yao, “An Alzheimers disease related genes identification method based on multiple classifier integration,” *Comput. Methods Programs Biomed.*, vol. 150, pp. 107–115, 2017.

[163] L. R. Ren, Y. L. Gao, J. X. Liu, R. Zhu, and X. Z. Kong, “Extreme Learning Machine: An Efficient Robust Classifier for Tumor Classification,” *Comput. Biol. Chem.*, vol. 89, no. September, p. 107368, 2020.

[164] M. Pandey and S. Taruna, “Towards the integration of multiple classifier pertaining to the Student’s performance prediction,” *Perspect. Sci.*, vol. 8, pp. 364–366, 2016.

[165] M. W. L. Moreira, J. J. P. C. Rodrigues, V. Furtado, N. Kumar, and V. V. Korotaev, "Averaged one-dependence estimators on edge devices for smart pregnancy data analysis," *Comput. Electr. Eng.*, vol. 77, no. January, pp. 435–444, 2019.

[166] J. Yan, D. B. Bracewell, F. Ren, and S. Kuroiwa, "Integration of Multiple Classifiers for Chinese Semantic Dependency Analysis," *Electron. Notes Theor. Comput. Sci.*, vol. 225, no. C, pp. 457–468, 2009.

[167] A. M. Holder, A. Markarian, J. M. Doyle, and J. R. Olson, "Predicting geographic distributions of fishes in remote stream networks using maximum entropy modeling and landscape characterizations," *Ecol. Modell.*, vol. 433, no. July, p. 109231, 2020.

[168] H. Lou and R. I. Cukier, "A maximum entropy principle approach to a joint probability model for sequences with known neighbor and next neighbor pair probabilities," *Chem. Phys.*, vol. 538, no. April, p. 110872, 2020.

[169] D. Ruano-Ordás, I. Yevseyeva, V. B. Fernandes, J. R. Méndez, and M. T. M. Emmerich, "Improving the drug discovery process by using multiple classifier systems," *Expert Syst. Appl.*, vol. 121, pp. 292–303, 2019.

[170] J. Novakovic and A. Veljovic, "C-support vector classification: Selection of kernel and parameters in medical diagnosis," *IEEE 9th Int. Symp. Intell. Syst. Informatics*, pp. 465–470, 2011.

[171] L. Oliveira, U. Nunes, and P. Peixoto, "On Exploration of Classifier Ensemble Synergism in Pedestrian Detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 16–27, 2010.

[172] G. Beliakov and D. Divakov, "On representation of fuzzy measures for learning Choquet and Sugeno integrals," *Knowledge-Based Syst.*, vol. 189, no. xxxx, p. 105134, 2020.

[173] D. Chong, N. Zhu, W. Luo, and X. Pan, "Human thermal risk prediction in indoor hyperthermal environments based on random forest," *Sustain. Cities Soc.*, vol. 49, no. April, p. 101595, 2019.

[174] L. Chen, C. Wang, J. Chen, Z. Xiang, and X. Hu, "Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K Nearest Neighbor (KNN)," *J. Voice*, 2020.

[175] K. Adem, "Diagnosis of breast cancer with Stacked autoencoder and Subspace kNN," *Phys. A Stat. Mech. its Appl.*, vol. 551, p. 124591, 2020.

[176] Y. Pan, H. Gao, H. Lin, Z. Liu, L. Tang, and S. Li, "Identification of bacteriophage virion proteins using multinomial Naïve bayes with g-gap feature tree," *Int. J. Mol. Sci.*, vol. 19, no. 6, 2018.

[177] Y. S. Kong, S. Abdullah, D. Schramm, M. Z. Omar, and S. M. Haris, "Optimization of spring fatigue life prediction model for vehicle ride using hybrid multi-layer perceptron artificial neural networks," *Mech. Syst. Signal Process.*, vol. 122, pp. 597–621, 2019.

[178] S. Naeem, S. Shahhosseini, and A. Ghaemi, "Simulation of CO2 capture using sodium hydroxide solid sorbent in a fluidized bed reactor by a multi-layer perceptron neural network," *J. Nat. Gas Sci. Eng.*, vol. 31, pp. 305–312, 2016.

[179] X. Fan, L. Wang, and S. Li, "Predicting chaotic coal prices using a multi-layer perceptron network model," *Resour. Policy*, vol. 50, pp. 86–92, 2016.

[180] C. Chang and C. Lin, "LIBSVM : A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–39, 2011.

[181] J. Liu and E. Zio, "Integration of feature vector selection and support vector machine for classification of imbalanced data," *Appl. Soft Comput. J.*, vol. 75, pp. 702–711, 2019.

[182] J. Chorowski, J. Wang, and J. M. Zurada, "Review and performance comparison of SVM- and ELM-based classifiers," *Neurocomputing*, vol. 128, pp. 507–516, 2014.

[183] S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," *Expert Syst. Appl.*, vol. 106, pp. 252–262, 2018.

[184] W. Li and Z. Liu, "A method of SVM with normalization in intrusion detection," *Procedia Environ. Sci.*, vol. 11, no. PART A, pp. 256–262, 2011.

[185] M. M. Rahman, B. C. Desai, and P. Bhattacharya, "Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion," *Comput. Med. Imaging Graph.*, vol. 32, no. 2, pp. 95–108, 2008.

[186] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[187] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognit.*, vol. 44, no. 2, pp. 330–349, 2011.

[188] D. S. Cao, J. H. Huang, Y. Z. Liang, Q. S. Xu, and L. X. Zhang, "Tree-based ensemble methods and their applications in analytical chemistry," *Trends Anal. Chem.*, vol. 40, no. 2, pp. 158–167, 2012.

[189] F. B. de Santana, W. Borges Neto, and R. J. Poppi, "Random forest as one-class classifier and infrared spectroscopy for food adulteration detection," *Food Chem.*, vol. 293, no. July 2018, pp. 323–332, 2019.

[190] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019.

[191] H. Alshalabi, S. Tiun, N. Omar, and M. Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization," *Procedia Technol.*, vol. 11, pp. 748–754, 2013.

[192] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecol. Modell.*, vol. 406, no. April, pp. 109–120, 2019.

[193] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 Int. Conf. Autom. Comput. Technol. Manag.*, pp. 593–596, 2019.

[194] G. Isabelle, W. Maharani, and I. Asror, "Analysis on Opinion Mining Using Combining Lexicon-Based Method and Multinomial Naïve Bayes," *2018 Int. Conf. Ind. Enterp. Syst. Eng. (ICoIESE 2018)*, vol. 2, no. IcoIESE 2018, pp. 214–219, 2019.

[195] B. T. Pham, M. D. Nguyen, K. T. T. Bui, I. Prakash, K. Chapi, and D. T. Bui, "A novel artificial intelligence approach based on Multi-layer Perceptron Neural Network and Biogeography-based Optimization for predicting coefficient of consolidation of soil," *Catena*, vol. 173, no. September 2018, pp. 302–311, 2019.

[196] H. Heo, H. Park, N. Kim, and J. Lee, "Prediction of credit delinquents using locally transductive multi-layer perceptron," *Neurocomputing*, vol. 73, no. 1–3, pp. 169–175, 2009.

[197] Y. Quan, Y. Xu, Y. Sun, and Y. Huang, "Supervised dictionary learning with multiple classifier integration," *Pattern Recognit.*, vol. 55, pp. 247–260, 2016.

[198] V. Gholami, K. W. Chau, F. Fadaee, J. Torkaman, and A. Ghaffari, "Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers," *J. Hydrol.*, vol. 529, no. March 2019, pp. 1060–1069, 2015.

[199] W. Chen *et al.*, "Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China," *Sci. Total Environ.*, vol. 626, pp. 1121–1135, 2018.

[200] I. H. Witten and E. Frank, "Data mining: Pratical Machine Learning Tools and Techniques with Java Implementations," *Morgan Kaufmann Publ.*, pp. 149–151, 2000.

[201] F. Bartel, R. Hielscher, and D. Potts, "Fast cross-validation in harmonic approximation," *Appl. Comput. Harmon. Anal.*, vol. 49, no. 2, pp. 415–437, 2020.

[202] B. Bischl *et al.*, "Resampling," *mlr*, 2017. [Online]. Available: https://mlr.mlr-org.com/articles/tutorial/resample.html.

[203] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[204] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[205] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling TEchnique (SMOTE) for Handling Class Imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019.

[206] S. Susan and A. Kumar, "SSO Maj -SMOTE-SSO Min : Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets," *Appl. Soft Comput. J.*, vol. 78, pp. 141–149, 2019.

[207] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined SMOTE and PSO based

RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456–3466, 2011.

[208] J. Sun, H. Li, H. Fujita, B. Fu, and W. Ai, "Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting," *Inf. Fusion*, vol. 54, no. July 2019, pp. 128–144, 2019.

[209] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "Viewpoint: When will ai exceed human performance? Evidence from ai experts," *J. Artif. Intell. Res.*, vol. 62, pp. 729–754, 2018.

[210] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, "Detecting hip fractures with radiologist-level performance using deep neural networks," *arXiv - Comput. Vis. Pattern Recognit.*, 2017.

[211] R. Grundkiewicz and M. Junczys-Dowmunt, "Near human-level performance in grammatical error correction with hybrid machine translation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, vol. 2, pp. 284–290.

[212] P. P. Angelov and X. Gu, "Deep rule-based classifier with human-level performance and characteristics," *Inf. Sci. (Ny).*, vol. 463–464, pp. 196–213, 2018.

[213] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 1026–1034, 2015.

[214] E. M. Rojas, "Machine Learning: analysis of programming languages and development tools," *Iber. J. Inf. Syst. Technol.*, pp. 586–599, 2020.

[215] scikit-learn, "scikit-learn: Machine Learning in Python," 2020. [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 09-Oct-2020].

[216] G. S. Duffó, *Biomateriales: una mejor calidad de vida*. EUDEBA, 2005.

[217] B. D. Ratner, A. S. Hoffman, F. J. Schoen, and J. E. Lemons, *Biomaterials Science. An introduction to Materials in Medicine*. San Diego, California - USA: Elsevier Academic Press, 1996.

[218] R. L. Macchi, *Materiales dentales*. Buenos Aires: Editorial Médica Panamericana, 2007.

[219] M. C. Piña Barba, *La física en la medicina*, 2da. Edici. México: Secretaría de Educación Pública, 1998.

[220] M. V. Regi, *Biomateriales: Repuestos para el Cuerpo Humano*. Madrid, España: Discurso de Ingreso Real Academia de Ingeniería, 2004.

[221] Universidad de Clemson, "The history of the Annual International Biomaterials Symposium & Annual Meeting of the Society of Biomaterials." [Online]. Available: http://www.clemson.edu/centers-institutes/cwhall/index.html. [Accessed: 09-Jul-2020].

[222] R. H. Alvarez, "Válvulas cardíacas protésicas: Revisión actualizada," *Revista de Posgrado de la Via Cátedra de Medicina*, vol. 137, pp. 19–32, 2004.

[223] G. P. Kothiyal and A. Srinivasan, "Trends in Biomaterials," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2016.

[224] T. R. Cuadrado, "Biomateriales y Dispositivos Biomedicos hacia la Sofisticación y el Reuso," *Instituto de Investigación en Ciencia y Tecnología de Materiales, Facultad de Ingenierías, Universidad Nacional del Mar de Plata, CONICET*. [Online]. Available: http://www.hpc.org.ar/images/revista/169-v3p86.pdf. [Accessed: 09-Jul-2020].

[225] J. Wirth, M. Tahriri, K. Khoshroo, M. Rasoulianboroujeni, and A. R. Dentino, "Surface modification of dental implants," *Biomater. Oral Dent. Tissue Eng.*, pp. 85–96, Jan. 2017.

[226] P. Bicudo, J. Reis, A. M. Deus, L. Reis, and M. F. Vaz, "Mechanical behaviour of dental implants," *Procedia Struct. Integr.*, vol. 1, pp. 26–33, 2016.

[227] G. Singh, "Surface Treatment of dental implants: A review," *IOSR J. Dent. Med. Sci.*, vol. 17, no. 2, pp. 49–53, 2018.

[228] J. Black and G. Hastings, "Handbook of Biomaterial Properties," *J. Control. Release*, vol. 65, no. 3, p. 439, 2000.

[229] S. Abullais, N. AlQahtani, N. Kudyar, and N. Priyanka, "Success of dental implants: Must-know prognostic factors," *J. Dent. Implant.*, vol. 6, no. 1, p. 44, 2016.

[230] J. R. Davis, "Handbook of Materials for Medical Devices," in *ASM International*, 1st ed., 2003, pp. 195–220.

[231] B. D. Ratner, A. S. Hoffman, F. J. Schoen, and J. E. Lemons, *Biomaterials Science: An Introduction to Materials in Medicine*. 2013.

[232] P. Altuna, E. Lucas-Taulé, J. Gargallo-Albiol, O. Figueras-Álvarez, F. Hernández-Alfaro, and J. Nart, "Clinical evidence on titanium-zirconium dental implants: A systematic review and meta-analysis," *Int. J. Oral Maxillofac. Surg.*, vol. 45, no. 7, pp. 842–850, 2016.

[233] S. AM Negm, "Implant Success versus Implant Survival," *Dentistry*, vol. 06, no. 02, pp. 2–6, 2016.

[234] C. Leyens and M. Peters, *Titanium an Titanium Alloys Fundamentals and Applications*. Wiley- VCH, 2003.

[235] D. N. R. Vootla and D. K. V. Reddy, "Osseointegration- Key Factors Affecting Its Success-An Overview," *IOSR J. Dent. Med. Sci.*, vol. 16, no. 04, pp. 62–68, 2017.

[236] J. Li, J. A. Jansen, X. F. Walboomers, and J. J. van den Beucken, "Mechanical aspects of dental implants and osseointegration: A narrative review," *J. Mech. Behav. Biomed. Mater.*, vol. 103, no. September 2019, p. 103574, 2020.

[237] A. Barfeie, J. Wilson, and J. Rees, "Implant surface characteristics and their effect on

osseointegration," *Br. Dent. J.*, vol. 218, no. 5, pp. 1–9, 2015.

[238] L. Le Guéhennec, A. Soueidan, P. Layrolle, and Y. Amouriq, "Surface treatments of titanium dental implants for rapid osseointegration," *Dent. Mater.*, vol. 23, no. 7, pp. 844–854, Jul. 2007.

[239] O. E. Ogle, "Implant Surface Material, Design, and Osseointegration," *Dent. Clin. North Am.*, vol. 59, no. 2, pp. 505–520, 2015.

[240] Clinica Lauden, "Implantología," 2018. [Online]. Available: https://www.clinicalauden.com/implantologia/.

[241] Q. Chen and G. A. Thouas, "Metallic implant biomaterials," *Mater. Sci. Eng. R Reports*, vol. 87, pp. 1–57, 2015.

[242] D. R. Prithviraj, S. Deeksha, K. M. Regish, and N. Anoop, "A systematic review of zirconia as an implant material," *Indian J Dent Res*, vol. 23, no. 5, pp. 643–649, 2012.

[243] A. Apratim, P. Eachempati, K. K. K. Salian, V. Singh, S. Chhabra, and S. Shah, "Zirconia in dental implantology: A review," *J. Int. Soc. Prev. Community Dent. Vol.*, vol. 5, no. 3, pp. 147–156, 2015.

[244] P. A. Assal, "The osseointegration of zirconia dental implants," *Schweiz Monatsschr Zahnmed*, vol. 123, no. 7, pp. 644–654, 2013.

[245] M. Saini, "Implant biomaterials: A comprehensive review," *World J. Clin. Cases*, vol. 3, no. 1, pp. 52–57, 2015.

[246] Z. Özkurt and E. Kazazoğlu, "Zirconia dental implants: A literature review," *J. Oral Implantol.*, vol. 37, no. 3, pp. 367–376, 2011.

[247] H. I. Arbildo-Vega, C. A. Lamas-Lara, and H. Vásquez-Rodrigo, "Survival rate of zirconium oxide dental implants. A systematic review and meta-analysis," *Rev. Esp. Cir. Oral y Maxilofac.*, vol. 39, no. 3, pp. 132–142, 2017.

[248] S. Saridag, O. Tak, and G. Alniacik, "Basic properties and types of zirconia: An overview," *World J. Stomatol*, vol. 2, no. 3, pp. 40–47, 2013.

[249] C. E. Misch, *Implantologia Contemporanea*. Elsevier, 2009.

[250] Honorable Congreso de la Nación Argentina, "Ley 25.326: Protección de Datos Personales," *Boletín Oficial*. Buenos Aires, Argentina., pp. 1–29, 2000.

[251] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Aplicación de la minería de datos para la selección de biomateriales en implantes dentales," *Rev. SAM*, vol. 1, pp. 40–44, 2019.

[252] N. B. Ganz, A. E. Ares, and H. D. Kuna, "Predicting dental implant failures by integrating multiple classifiers," *Rev. Cienc. y Tecnol.*, vol. 34, 2020.

[253] sciki-learn, "Tuning the hyper-parameters of an estimator," 2019. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html#grid-search. [Accessed: 30-Jul-2020].

[254] X. Fan and H. Shin, "Road vanishing point detection using weber adaptive local filter

and salient-block-wise weighted soft voting," *IET Comput. Vis.*, vol. 10, no. 6, pp. 503–512, 2016.

[255] L. N. Eeti and K. M. Buddhiraju, "A modified class-specific weighted soft voting for bagging ensemble," *Int. Geosci. Remote Sens. Symp.*, vol. November, pp. 2622–2625, 2016.

[256] R. Qiong Wei and J. L. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS One*, vol. 8, no. 7, pp. 1–12, 2013.

[257] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci. (Ny).*, vol. 513, pp. 429–441, 2020.

[258] A. y T. M. (anmat) Administración Nacional de Medicamentos, "Vademecum Nacional de Medicamentos," 2020. [Online]. Available: http://anmatvademecum.servicios.pami.org.ar/index.html. [Accessed: 13-Jul-2020].

[259] Institut Straumann AG, "Straumann SLA®," 2020. [Online]. Available: https://www.straumann.com/ar/es/profesionales-de-la-odontologia/ciencia/bibliografia/sla.html. [Accessed: 13-Jul-2020].

[260] Institut Straumann AG, "Las superficies del Straumann® Dental Implant System," 2020. [Online]. Available: https://www.straumann.com/ar/es/profesionales-de-la-odontologia/productos-y-soluciones/implantes-dentales/superficies-de-los-implantes-dentales.html. [Accessed: 13-Jul-2020].

[261] ML Implant System, "Implantes SHe," 2020. [Online]. Available: https://www.mlimplantsystem.com.ar/implantes-she. [Accessed: 13-Jul-2020].

[262] ROSTERDENT, "Catálogo de Productos." pp. 3–4, 2017.

[263] Alpha-Bio Tec, "Catálogo de productos." 2017.

[264] Tree-Oss, "Superficie OXALIFE®," 2020. [Online]. Available: https://tree-oss.com/superficie-oxalife/. [Accessed: 13-Jul-2020].

[265] bio com, "Sistema de Implante Dental - Catálogo," 2012. [Online]. Available: https://issuu.com/allimplant5/docs/catalogo-allimplant-2012. [Accessed: 13-Jul-2020].

[266] Richard J. Lazzara, "Implants System Design and Its Potential Impact on the Establishment and Sustainability of Aesthetics," *J. Implant Reconstr. Dent.*, vol. 1, 2012.

[267] B&W Group, "Generalidades de los implantes: Grabado Bi-ácido," 2012. [Online]. Available: http://bywgroup.com/sitio_anterior/generalidades.html. [Accessed: 13-Jul-2020].

[268] Biohorizons, "Tapered Internal Catálogo de Producto," 2010. [Online]. Available: https://issuu.com/biodental/docs/tapered_internal__espa_ol_/6. [Accessed: 13-Jul-2020].

[269] SmileTech, "SmileTech: Tratamiento de Superficie," 2013. [Online]. Available:

http://www.red-dental.com/O_N67901.HTM. [Accessed: 13-Jul-2020].

[270] Biounite, “Biounite® Nanotecnología a su Alcance,” 2014. [Online]. Available: http://www.red-dental.com/O_N72601.HTM. [Accessed: 13-Jul-2020].

[271] Nobel Biocare, “TiUnite,” 2020. [Online]. Available: https://www.nobelbiocare.com/en-us/tiunite. [Accessed: 13-Jul-2020].

10. ANEXOS

10.1 ANEXO I – INGEST_MED

En este anexo se describe el estudio realizado sobre la acción terapéutica, las indicaciones y tipificación final de los medicamentos o drogas [258] que contiene la variable INGEST_MED del conjunto de datos utilizado para el estudio de caso (ver Tabla 27).

ANTIHIPERTENSIVO: grupo de medicamentos utilizados para el tratamiento de la hipertensión arterial (HTA). Estos fármacos tienen la finalidad de normalizar la presión arterial irregular (frecuentemente presión arterial alta), para prevenir enfermedades cardiovasculares.

ANTIDIABETICO: medicamentos usados para reducir los niveles de glucosa en sangre. La selección de los diferentes tipos de antidiabéticos depende de la enfermedad, la edad y condición de salud del paciente, así como otros factores.

HORMONAS: sustancias segregadas por células especializadas, localizadas en glándulas endocrinas (carentes de conductos), o también por células epiteliales e intersticiales cuyo fin es el de influir en la función de otras células. Así también, este grupo de medicamentos engloba a los anticonceptivos hormonales, siendo el método más eficaz para controlar la fertilidad y evitar el embarazo.

AAA: conocido como las triples A (Analgésico, Antiinflamatorio y Antipirético). Los medicamentos que se emplean para tratar el dolor se corresponde a los analgésicos, los medicamentos usados para prevenir o disminuir la inflamación de los tejidos son antiinflamatorios, y los que actúan reduciendo la fiebre se conocen como antipiréticos.

ANSIOLITICO: un ansiolítico o tranquilizante es un fármaco psicotrópico con acción depresora del sistema nervioso central, destinado a disminuir o eliminar los síntomas de la ansiedad, angustia, nerviosismo y del insomnio.

ANTIBIOTICO: los antibióticos son medicamentos potentes que combaten las infecciones bacterianas. Actúan matando las bacterias o impidiendo que se reproduzcan.

ANTIARTROSICO: medicamentos específicos para tratar la artrosis. Tienen la capacidad de disminuir la intensidad del dolor y mejorar la movilidad del paciente afectado. La eficacia de estos medicamentos está demostrada en la osteoartritis de cadera y rodilla.

HIPOLIPEMIANTE: medicamentos que tienen la propiedad de disminuir los niveles de lípidos en sangre. La importancia de estas sustancias viene dada porque el exceso de algunos tipos de lípidos (colesterol o triglicéridos) o de las lipoproteínas que es uno de los principales factores de riesgo para enfermedades cardiovasculares.

SUPLEMENTO: suplementos nutricionales indicados para el tratamiento de anemias (hierro), osteoporosis (calcio), prevención y tratamiento de los estados carenciales de vitamina A y vitamina D para ayudar al cuerpo a absorber el calcio. También, vitamina B para el aumento de melanina a causas de problemas de pigmentación de la piel y sensibilidad a los rayos UV.

ANTIHISTAMINICO: fármaco que sirve para reducir o eliminar los efectos de las alergias. Actúa bloqueando la acción de la histamina en las reacciones alérgicas, a través del bloqueo de sus receptores. Permiten tratar los síntomas de la congestión, secreción nasal, estornudos, picazón, hinchazón de las vías respiratorias, urticaria, erupciones cutáneas y secreción de los ojos.

ANTIULCEROSO: medicamentos que buscan conseguir el alivio de los síntomas, curan o facilitan la cicatrización de úlceras gástricas o/y ulceras duodenales. También llamados agentes antibacterianos. Estos medicamentos actúan evitando el crecimiento y la propagación de la bacteria Helicobacter pylori que, a menudo, ocurre con las úlceras. Tratar esta infección previene que las úlceras regresen.

HIPNOTICO: los fármacos somníferos e hipnóticos son drogas psicotrópicas y psicoactivas que inducen al sueño o somnolencia. Se utilizan regularmente cuando el paciente presenta dificultades a la hora de dormir, para prevenir trastornos que provoquen malestar o alteraciones que interfieran en las actividades cotidianas.

ANTINEOPLÁSICO: los medicamentos antineoplásicos o inmunomodulador, tienen la capacidad de impedir el desarrollo, crecimiento o proliferación de células tumorales malignas. Se utilizan en la quimioterapia del cáncer. Estos fármacos pueden actuar sobre una o varias fases del ciclo celular o sobre los mecanismos de control de la reproducción de las células vivas.

Las búsquedas y descripción de los medicamentos fueron realizadas en el Diccionario MedlinePlus[16] y en el Vademécum Nacional de Medicamentos (VNM) – ANMAT[17].

[16] MedlinePlus. Disponible en https://medlineplus.gov/spanish/. (Consultado el 14/10/2020).
[17] VNM-ANMAT. Disponible en http://anmatvademecum.servicios.pami.org.ar/index.html. (Consultado el 14/10/2020).

Tabla 27. Descripción de medicamentos [258].

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
LOTRIAL	Enalapril maleato	Inhibidor de la enzima convertidora de la angiotensina (ECA).	Para todos los grados de hipertensión esencial y en la hipertensión vasculorrenal. Puede emplearse como indicación inicial o asociado con otros agentes antihipertensivos, sobre todo diuréticos.	ANTIHIPERTENSIVO
ATENOLOL	Atenolol	Bloqueante beta-adrenérgico.	Hipertensión esencial. Angor pectoris. Arritmias cardíacas. Coadyuvante del tratamiento de la estenosis subaórtica hipertrófica.	
ATACAND	Candesartán cilexetil	Antihipertensivo	Hipertensión arterial.	
LOSACOR	Losartan	Antihipertensivo	Hipertensión arterial esencial leve a moderada. Insuficiencia cardíaca congestiva.	
MICARDIS	Telmisartan	Antihipertensivo	Tratamiento de la hipertensión arterial esencial. Prevención de la morbilidad y la mortalidad cardiovascular, en pacientes mayores de 55 años de edad con riesgo elevado de enfermedad cardiovascular que no pueden recibir tratamiento con inhibidores de la enzima convertidora de angiotensina.	
MICARDIS PLUS	Telmisartán + hidroclorotiazida	Antihipertensivo diurético	Tratamiento de la hipertensión esencial. Como combinación fija, se indica en aquellos pacientes en los que no se logra un adecuado control de la presión sanguínea con telmisartan o hidroclorotiazida solos.	

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
AMLODIPINA	Amlodipina	Antihipertensivo Antianginoso	Para hipertensión arterial (sola o en combinación con otros antihipertensivos). Angina crónica estable (sola o en combinación con otros agentes antianginosos). Angina vasospástica, de Prinzmetal o variante (sola o en combinación con otros agentes antianginosos).	ANTIHIPERTENSIVO
LISINAL	Lisinopril	Antihipertensivo	Hipertensión leve, moderada, grave y vasculorrenal. Insuficiencia cardíaca congestiva como coadyuvante de diuréticos y digitálicos.	
DIOVAN	Valsartan	Antihipertensivo	Insuficiencia cardíaca (NYHA clase II-IV). Hipertensión arterial leve a moderada de diferente etiología. Posinfarto de miocardio.	
MECTIN	Metformina	Hipoglucemiante	Para diabetes mellitus no dependiente de insulina (tipo II) leve o moderada; utilizada particularmente en pacientes obesos o con tendencia al sobrepeso.	ANTIDIABETICO
INSULINA (Este no es exactamente el nombre del medicamento. Existen varias marcas nacionales e importadas)	Insulina	Hipoglucemiante	Tratamiento de la diabetes mellitus dependiente de insulina o como suplemento de la producción fisiológica de insulina endógena en pacientes con diabetes mellitus no dependiente de insulina. También puede agregarse a soluciones de hiperalimentación para facilitar la utilización de glucosa en pacientes con poca tolerancia a ella.	

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
LEVOTIROXINA	Levotiroxina	Hormona tiroidea. Agente de diagnóstico de la función tiroidea.	Tratamiento del hipotiroidismo por deficiencia de hormona tiroidea de cualquier etiología, así como el bocio simple (no endémico) y en la tiroiditis linfocítica crónica (de Hashimoto). Supresión del crecimiento de bocios adenomatosos y para prevenir los efectos bociogénicos de otros fármacos (litio, ácido aminosalicílico y algunos compuestos tipo sulfamida). Carcinoma de glándula tiroides, dependiente de tirotropinas.	HORMONAS
ANTICONCEPTIVOS	drospirenona +etinilestradio levonorgestrel +etinilestradiol (Asociaciones más utilizadas en Argentina)	Anticoncepción, contracepción o control de la natalidad.	La anticoncepción, contracepción o control de la natalidad es cualquier método o dispositivo para prevenir el embarazo.	
ASPIRINA	Ácido acetilsalisílico	Analgésico Antiinflamatorio Antipirético	Procesos dolorosos somáticos, inflamación de distinto tipo, fiebre. Profilaxis y tratamiento de trombosis venosas y arteriales. Artritis reumatoidea y juvenil. Profilaxis del infarto de miocardio en pacientes con angor pectoris inestable.	AAA
PARACETAMOL	Paracetamol	Analgésico Antipirético	Cefalea, odontalgia y fiebre.	
DIOXAFLEX PLUS	Diclofenac + pridinol	Analgésico Antiinflamatorio Antirreumático Miorrelajante	Procesos agudos o crónicos que se acompañan de dolor e inflamación.	

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
DORIXINA RELAX	Clonixinato lisina + Asoc.	Analgésico Antiinflamatorio Antipirético	Está destinado para el tratamiento del dolor de origen musculoesquelético, en especial cuando se acompaña de contractura muscular.	AAA
DIROLAC	Ketorolaco	Analgésico Antiinflamatorio	Para tratamientos a corto plazo del dolor agudo, de moderado a severo.	
CORTISONA (Este no es exactamente el nombre medicamento. Existen varias marcas comerciales)	Hidrocortisona sola o asociada a otras drogas	Corticosteroide Antiinflamatorio esteroide Inmunosupresor	Insuficiencia adrenocortical aguda o primaria crónica, síndrome adrenogenital, enfermedades alérgicas, enfermedades del colágeno, anemia hemolítica adquirida, anemia hipoplásica congénita, trombocitopenia secundaria en adultos, enfermedades reumáticas, enfermedades oftálmicas, tratamiento del shock. Enfermedades respiratorias, neoplásicas (manejo paliativo de leucemias y linfomas en adultos, y de leucemia aguda en la niñez), estados edematosos, enfermedades gastrointestinales (para ayudar al paciente a superar períodos críticos en colitis ulcerativa y enteritis regional), triquinosis con compromiso miocárdico.	
RIVOTRIL	Clonazepam	Ansiolítico Anticonvulsivante	Está indicado en los trastornos de angustia (ataque de pánico) con o sin agorafobia. Está indicado solo o como adyuvante, en el tratamiento del síndrome de Lennox-Gastaut (variante del petit mal), crisis convulsivas acinéticas y mioclónicas. Puede ser empleado en pacientes con crisis de ausencia (petit mal) refractarias a las succinimidas.	ANSIOLITICO

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
ALPLAX	Alprazolam	Ansiolítico	Trastorno de ansiedad generalizada (DSM-IV). Ansiedad asociada con depresión. Trastorno de angustia (Ataque de pánico) con o sin agorafobia.	ANSIOLITICO
Varios nombres comerciales	Lorazepam	Ansiolítico Sedante / hipnótico Relajante musculo-esquelético.	Trastornos por ansiedad. Ansiedad asociada con depresión mental. Síntomas de supresión alcohólica aguda. Insomnio por ansiedad o situaciones pasajeras de estrés.	
Varios nombre comerciales entre ellos Valium	Diazepam	Ansiolítico Miorrelajante Anticonvulsivo	Para tratamiento de epilepsia, ansiedad, trastornos psicosomáticos, tortícolis, espasmos musculares.	
Varios nombre comerciales	Citalopram	Ansiolítico Antidepresivo	Tratamiento de la depresión y abuso del alcohol.	
TOBRADEX	Tobramicina + dexametasona	Antibiótico Antiinflamatorio oftálmico de aplicación tópica.	Indicado para la prevención y tratamiento de inflamación y prevención de la infección asociada con cirugía de catarata en adultos y niños con edad de 2 años y mayores. Indicado para condiciones inflamatorias que responden a esteroides para las que está indicado un corticosteroide y en las que existe una infección ocular bacteriana superficial o un riesgo de infección ocular bacteriana.	ANTIBIOTICO
TETRALYSAL	Limeciclina	Antibiótico de amplio espectro.	Antibiótico sistémico de amplio espectro.	

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
Varios nombres comerciales	Glucosamina	Antiartrósico	Artrosis (primaria y secundaria). Osteocondroartrosis. Espondilosis. Condromalacia rotular. Periartritis escapulohumeral.	ANTIARTROSICO
LIPITOR	Atorvastatina	Hipolipemiante Hipocolesterolemiante	Hipercolesterolemias. Hipercolesterolemia familiar. Dislipidemias. Tratamiento adyuvante a la dieta para disminuir los niveles elevados de colesterol total, LDL-colesterol, apobetalipoproteínas y triglicéridos en pacientes con hipercolesterolemia primaria (heterocigota familiar y no familiar) y en la dislipemia mixta (Fredrickson tipos IIa y IIb). Es un agente sintético que reduce los lípidos.	HIPOLIPEMIANTE

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
Varios nombres comerciales	Vitamina A (Retinol)	Suplemento nutricional	Prevención y tratamiento de los estados carenciales de vitamina A. Síndrome de malabsorción, queratomalacia, xeroftalmía y nictalopía.	SUPLEMENTO
Varios nombres comerciales	Vitamina D	Suplemento	La vitamina D ayuda al cuerpo a absorber el calcio, uno de los principales elementos que constituyen los huesos.	
Varios nombres comerciales	Calcio	Suplemento	Utilizado en pacientes con dieta deficiente en calcio o cuando los requerimientos normales se encuentran incrementados, por ejemplo en embarazadas o durante la lactancia. Para tratamiento de osteoporosis, hipocalcemia crónica, hiperfosfatemia e hiperacidez gástrica.	
Varios nombres comerciales	Hierro	Antianémico	Está indicado en el tratamiento de anemias ferropénicas y como preventivo de la deficiencia de hierro.	
Varios nombres comerciales	Vitamina B	Vitamínico	Vitamina esencial para el aumento de melanina, esta proporciona a la piel la protección que necesita ante la acción de los rayos UV.	
LORATADINA	Loratadina	Antialérgico	Síntomas asociados con rinitis alérgica, como estornudos, secreción nasal (rinorrea) y prurito. Los signos y síntomas oculares y nasales son aliviados rápidamente después de la administración oral. Urticaria crónica y otras afecciones dermatológicas alérgicas.	ANTIHISTAMINICO

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
PANTUS GASTROMAX	Pantoprazol	Antiulceroso. Inhibidor selectivo de la bomba de protones.	Ulcera duodenal. Ulcera gástrica. Esofagitis por reflujo gastroesofágico. Síndrome de Zollinger-Ellison y otras condiciones hipersecretorias patológicas de ácido clorhídrico.	ANTIULCEROSO
OMEPRAZOL	Omeprazol	Antiulceroso	Está indicado para el tratamiento de úlcera duodenal, úlcera gástrica, enfermedad ulcerosa péptica con histología antral o cultivo positivo para Helicobacter pylori, esofagitis por reflujo, síndrome de Zollinger-Ellison, pacientes con riesgo de aspiración del contenido gástrico durante anestesia general (profilaxis de aspiración).	
ESOMEPRAZOL	Esomeprazol	Antiulceroso	Está indicado para: enfermedad de reflujo gastroesofágico (GERD). Tratamiento de esofagitis erosiva por reflujo. Tratamiento prolongado de pacientes con esofagitis curada para evitar recidivas. Tratamiento sintomático de la enfermedad de reflujo gastroesofágico (GERD). En combinación con regímenes de terapia antibacteriana apropiados para la erradicación de Helicobacter pylori y tratamiento de la úlcera duodenal asociada a Helicobacter pylori. Prevención de recaídas de úlcera péptica en pacientes con úlcera asociada a Helicobacter pylori.	

Medicamento	Droga (principio activo)	Acción Terapéutica	Indicaciones	Tipificación
SOMIT	Zolpidem hemitartrato + zolpidem tartrato	Hipnótico no benzodiazepínico	Tratamiento a corto plazo del insomnio primario.	HIPNOTICO
Varios nombres comerciales	Metotrexato	Citostático. Antimetabolito	Leucemia linfocítica aguda, leucemia mieloblástica aguda, carcinoma de mama, carcinoma de pulmón de células escamosas y de células pequeñas, linfomas no Hodgkin, linfosarcomas, psoriasis refractaria y severa que no responde en forma adecuada a otros tratamientos, artritis reumatoidea que no responde a otras terapéuticas, tumores trofoblásticos y carcinoma de cabeza y cuello (epidermoide).	ANTINEOPLÁSICO
Varios nombres comerciales	Anastrozol	Inhibidor selectivo no esteroideo de la aromatasa.	Cáncer de mama avanzado en mujeres posmenopáusicas. Tratamiento adyuvante en mujeres posmenopáusicas con cáncer de mama con receptor hormonal positivo precoz. Tratamiento de primera línea en mujeres posmenopáusicas con receptor hormonal positivo o desconocido con cáncer de mama localizado o metastásico. Tratamiento en mujeres posmenopáusicas con cáncer de mama avanzado con progresión lesional posterapia con tamoxifeno.	

10.2 ANEXO II – TRAT_SUP

En este anexo se presenta toda la información relevada y obtenida sobre los distintos tipos de tratamientos de superficie de las marcas de implantes dentales contenidos en el conjunto de datos utilizado.

SLA (Blasting + doble grabado ácido)

El tratamiento de superficie S.L.A. (Sandblasted, Large grit, Acid-etched) es la técnica más estudiada y documentada en implantología. Se encuentra patentada por Straumann [259]. Las siglas en ingles hacen referencia a:

- Sandblasted: arenado, granallado o chorreado abrasivo, es la operación de propulsar a alta presión un fluido, que puede ser agua, aire o una fuerza centrífuga con fuerza abrasiva, contra una superficie para alisarla o eliminar materiales contaminantes.
- Large grit: grano grueso o grande.
- Acid-etched: ataque ácido o grabado ácido.

Esta técnica consiste en el chorreado con arena de grano grueso, la cual genera una macro-rugosidad en la superficie del titanio. Posteriormente, se realiza un grabado con ácido que superpone una micro-rugosidad (ver Figura 41). La topografía resultante ofrece una estructura ideal para la adhesión celular.

El arenado permite crear una rugosidad óptima y aporta fijación mecánica, mientras que el grabado ácido suaviza las elevaciones creadas y mejora la adhesión de proteínas. Las características de la superficie obtenida dependen del tipo de partículas, su dureza, tamaño, y velocidad de impacto. El grabado ácido elimina varias capas atómicas de la superficie y reduce la posibilidad de contaminación de las partículas sobrantes del proceso de limpieza. Los agentes de grabado ácido más utilizados son ácido fluorhídrico, nítrico, sulfúrico y sus combinaciones. La técnica más usada es el doble grabado con ácido, que se lleva a cabo en dos fases: en una primera inmersión de los implantes en soluciones de Ácido Clorhídrico + Ácido Sulfúrico, Ácido Nítrico + Ácido Fluorhídrico o Ácido Nítrico. Después el implante es nuevamente inmerso en una solución acuosa de Ácido Nítrico que estabiliza la capa de óxido de titanio superficial. El arenado se realiza mediante la propulsión de partículas de diferentes tamaños de sílice, alúmina y óxido de Titanio. SLA proporciona la mejor opción de capacidad de humectabilidad y un buen ángulo de contacto [259].

Figura 41. Imágenes SEM del trabamiento SLA. [260]

Las marcas que utilizan este tipo de técnica para el tratamiento de superficie de sus implantes son: FIA[18], ML[19], STRAUMANN[20], ROSTERDENT[21], Q-IMPLANT[22 23], ALPHA-BIO[24] y ODONTIT[25].

Los implantes de las marcas FIA, Q-IMPLANT y ODONTIT aplican de igual manera los pasos descriptos del proceso SLA para el tratamiento de superficie de sus implantes.

Los implantes ML son tratados con un proceso físico-químico que condiciona la superficie y proporcionan una porosidad ideal para aumentar el área de contacto con el tejido óseo y así mejorar el proceso de oseointegración. El proceso con el cual se desarrolla el tratamiento se encuentra validado y es realizado en equipos diseñados especialmente para esta tarea, por lo que posee un control automatizado de todos los parámetros durante el arenado (velocidad, dirección, presión y tamaño de partícula), el grabado acido (temperatura, tiempo y potencia) y el pasivado (temperatura y tiempo) [261].

El proceso radica en:

1. Arenado: proyectar partículas a gran velocidad sobre una superficie, esto provoca la deformación plástica localizada sobre la superficie impactada y arranque del material.
2. Grabado Ácido: procedimiento basado en la corrosión controlada del titanio sumergido en una solución acida lo cual produce superficies rugosas y una topografía con micro huecos. Este tratamiento sumado al arenado permite conseguir dos escalas de rugosidad, los micro huecos provocados por el grabado ácido y una rugosidad de mayor escala asociada al arenado, por lo que, consigue mayor superficie de contacto hueso-implante comparado a los que se obtienen realizando solo uno de ellos.
3. Pasivado: la resistencia a la corrosión del titanio se debe a una película pasiva de óxido que se forma de manera natural y espontanea en contacto con en el aire y otros medios. Como consecuencia los agentes químicos y biológicos no interaccionan directamente con el titanio. A pesar de que el pasivado del titanio se produce de manera espontánea y natural es necesario favorecer el proceso mediante un tratamiento ácido para lograr una capa de óxido de mayor espesor y más homogéneo a lo largo de la superficie (ver Figura 42).

[18] FIA Implantes. Sistema BIOMEC. Disponible en https://implantesfia.com/biomec/. (Consultado el 14/10/2020).
[19] ML Implant System. Implantes SHe. Disponible en https://www.mlimplantsystem.com.ar/implantes-she. (Consultado el 14/10/2020).
[20] Institut Straumann AG. Straumann SLA®. Disponible en https://www.straumann.com/ar/es/profesionales-de-la-odontologia/ciencia/bibliografia/sla.html. (Consultado el 14/10/2020).
[21] ROSTERDENT. Catálogo de Productos. Disponible en http://www.rosterdent.com/CATALOGO-2017.pdf. (Consultado el 14/10/2020).
[22] Q-implant. Tratamiento y Envasado. Disponible en https://q-implant.ar/web/. (Consultado el 14/10/2020).
[23] Q-implant. Catálogo 2016 - Implantes, componentes protéticos e instrumental quirúrgico. Disponible en http://q-implant.com.ar/webml/wp-content/uploads/2018/08/CatalogoWeb.pdf. (Consultado el 14/10/2020).
[24] Alpha-Bio Tec. Catálogo de productos 2017. Disponible en http://www.alpha-bio.com.ar/alpha_newsletter/catalogo17.pdf. (Consultado el 14/10/2020).
[25] Odontit Implant Systems. Disponible en https://odontit.com/sistemasDeImplantes-es.php?btn=SISTEMAS-DE-IMPLANTES. (Consultado el 14/10/2020).

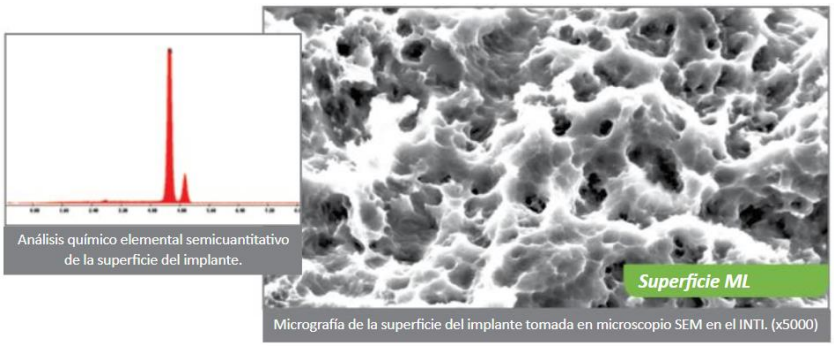


Figura 42. Micrografía de la superficie del implante ML, tomada en microscopio SEM en el INTI (x5000). [261]

Los implantes ROSTERDENT [262] cuentan con tratamiento S.L.A., el cual se realiza en 3 procesos generales: blasting, tratamiento ácido y neutralizado, los primeros crean macro y micro rugosidad, imitando la textura del hueso, facilitando así el proceso de óseo integración. El proceso culmina con exposición a radiación, para eliminar toda bacteria no deseada. La superficie se encuentra tratada mediante blasting más grabado ácido, el cual sirve para crear cavidades en la superficie del implante que son posteriormente tratadas con la técnica de pasivado. La técnica de pasivado químico tiene por objetivo aumentar la resistencia a la corrosión del implante. La inmersión en disoluciones ácidas con ligero carácter oxidante crea una capa inerte y estable de óxido de titanio de aproximadamente 5 a 10 micrómetros.

La superficie de los implantes ALPHA-BIO [263] son de tipo híbrido y surge de un complejo proceso de arenado con partículas de gran tamaño (de 20 a 40 micrones) y doble grabado térmico para la creación de microporos (de 1 a 5 micrones). Este proceso exclusivo crea una superficie altamente diferenciada, incrementa el área tridimensional (3D) y por lo tanto permite una absorción más intensa de sangre y proteínas de plasma directamente a los microporos del implante, inmediatamente después de su colocación (ver Figura 43).

La microestructura y las propiedades de rugosidad de la superficie de implante creadas mediante el proceso de arenado y doble grabado ácido, influyen considerablemente en el contacto inicial con el huésped (paciente) [259].

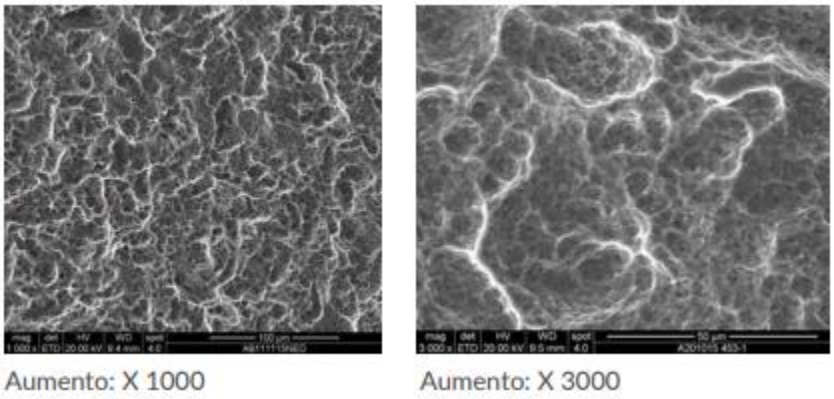


Figura 43. SEM de la superficie del implante ALPHA-BIO. [263]

Blasting + Grabado ácido + Tratamiento térmico

Esta técnica se logra mediante tres procedimientos:

1. Blasting para macro rugosidad, es decir proyección de partículas cerámicas por medio de un flujo de aire.
2. Grabado ácido para micro rugosidad.
3. Tratamiento térmico para una capa aumentada de óxido de titanio.

Las marcas que utilizan este tipo de técnica para el tratamiento de superficie de sus implantes son: TREE-OSS[26] y BIOCOM[27].

Los implantes TREE-OSS poseen el tratamiento de superficie OXALIFE [264]. Este novedoso tratamiento confiere al titanio comercialmente puro la rugosidad y porosidad ideales para una óptima respuesta biológica. La combinación de macro y micro porosidad aumenta la capacidad de humectabilidad superficial, lo que aloja los factores de crecimiento brindando una superficie altamente osteoconductiva. Su capa de óxido aumentada garantiza una excelente respuesta biológica favoreciendo el entorno para una predecible oseointegración temprana. El tratamiento superficial OXALIFE se logra mediante blasting, posteriormente un grabado ácido y finalmente un tratamiento térmico.

Los implantes BIOCOM utilizan el tratamiento de superficie OXACID [265]. Esta superficie confiere textura y porosidad ideal para el óptimo efecto biológico, favoreciendo el entorno para la oseointegración en menor plazo (ver Figura Figura 44). Esta superficie es obtenida mediante un tratamiento combinado de blastinado, grabado ácido y oxidación térmica.

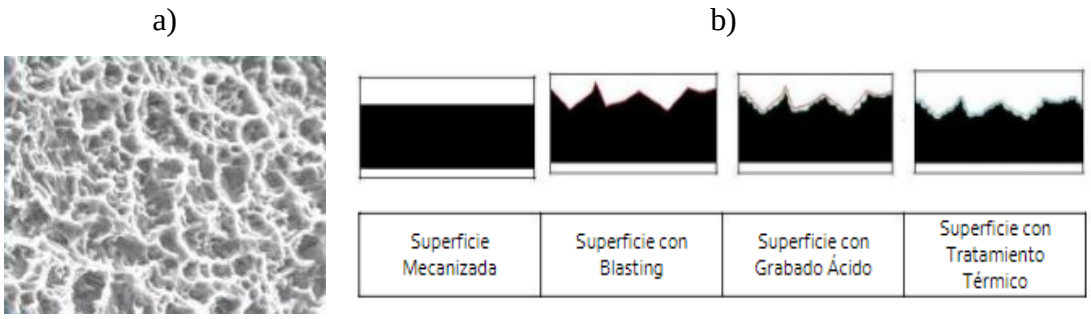


Figura 44. a) Microscopía electrónica de barrido de una superficie con tratamiento OXACID. b) Proceso del tratamiento OXACID. [265]

[26] Tree-Oss. Superficie OXALIFE®. Disponible en https://tree-oss.com/superficie-oxalife/. (Consultado el 14/10/2020).

[27] BIOCOM. Catálogo de productos 2012. Disponible en https://issuu.com/allimplant5/docs/catalogo-allimplant-2012. (Consultado el 14/10/2020).

Blasting + doble grabado ácido + depósitos de nano cristales de fosfato cálcico

La marca que utiliza este tipo de técnica para el tratamiento de superficie de sus implantes es BIOMET-3I[28].

Los implantes BIOMET-3I cuentan con superficie OSSEOTITE, patentada por Zimmer Biomet. Esta consiste en superficies arenadas con fosfato cálcico, doble grabado ácido y depósitos de nano cristales de fosfato cálcico. El tratamiento de la superficie consiste en un depósito discreto de cristales (DCD) de nanopartículas de cristales de fosfato cálcico sobre la Superficie OSSEOTITE, no es un recubrimiento rociado con plasma.

Proceso DCD (ver Figura 45):

1. Las partículas extremadamente pequeñas (de escala nano) de fosfato cálcico altamente cristalino, se encuentran suspendidas en la solución.
2. A continuación se induce a estas partículas a "auto-ensamblarse" sobre la superficie de óxido de titanio del implante.
3. Esto produce depósitos discretos de cristales de 20-100 nanómetros de longitud sobre la superficie del implante OSSEOTITE tratada con doble grabado ácido. La fuerza de unión de los cristales a la superficie OSSEOTITE supera el valor mínimo de resistencia, de 34,5 MPa, establecido por la norma ASTM F 1609-03[29] para la fijación de los revestimientos tradicionales de HA a las superficies de implantes.

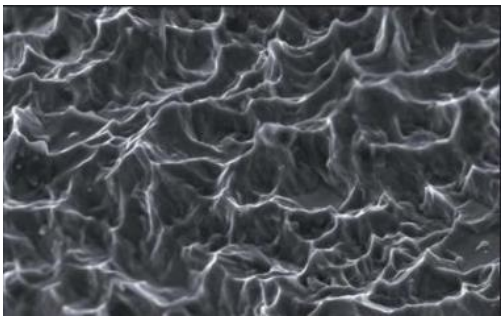


Figura 45. Superficie OSSEOTITE a 20.000 aumentos. [266]

Grabado bi-ácido

La marca que utiliza este tipo de técnica para el tratamiento de superficie de sus implantes es B&W[30].

[28] Zimmer Biomet, BIOMET 3I. Disponible en https://www.zimmerbiometdental.com/en-US/wps/wcm/connect/dental/6cf1ce94-1784-4add-a2a7-d053692932b2/ZB0067ES_REV_A_Osseotite_Implant_Brochure_final_SECURED.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE.Z18_1O041O02L8PAF0A9JPRUH520H76cf1ce94-1784-4add-a2a7-d053692932b2. (Consultado el 14/10/2020).

[29] ASTM. Disponible en https://www.astm.org/FAQ/index-spanish.html. (Consultado el 14/10/2020).

[30] B&W Group. Generalidades de los implantes: Grabado Bi-ácido. Disponible en http://bywgroup.com/sitio_anterior/generalidades.html. (Consultado el 14/10/2020).

Todos los implantes B&W están fabricados con Ti comercialmente puro grado IV, conforme a las normas ASTM, asegurando una perfecta biocompatibilidad con el organismo. Los implantes B&W son sometidos a un tratamiento de grabado bi-ácido, mezcla de ácido nítrico y clorhídrico a temperatura. Este tratamiento permite crear una micro rugosidad homogénea y controlada, logrando una mayor superficie de contacto entre el implante y el hueso. Favoreciendo la respuesta biológica de adhesión celular y asegurando una mejor y más rápida oseointegración [267] (ver Figura 46).

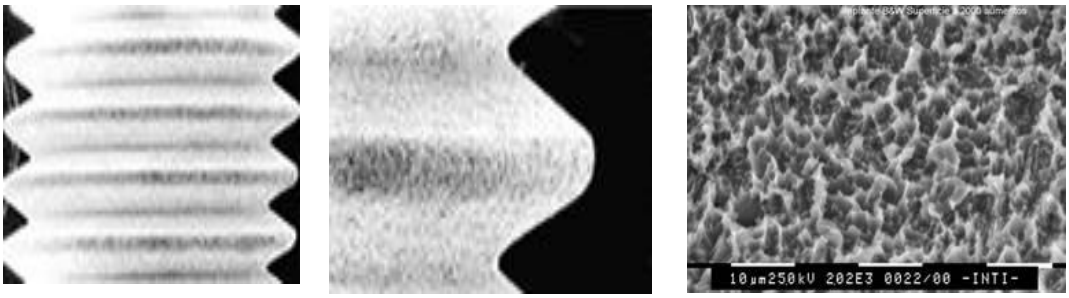


Figura 46. Superficie del implante a diferentes aumentos (microscopio electrónico). [267]

Blasting + grabado ácido

Las marcas que utilizan este tipo de técnica para el tratamiento de superficie de sus implantes son: MICROFIT[31], NEODENT[32] y FEDERA[33].

Blasting con fosfato tricálcico + grabado ácido

La marca que utiliza este tipo de técnica para el tratamiento de superficie de sus implantes es BIOHORIZONS[34].

Los implantes BIOHORIZONS cuentan con el tratamiento mediante bombardeo con material reabsorbible (RBT) [268], el cual proporciona una superficie compleja que mejora la estabilidad y la osteointegración.

[31] MICROFIT. Sistema de Implantes Dentales IN-Fit. Disponible en http://www.microfit.com.ar/interno.html. (Consultado el 14/10/2020).
[32] NEODENT. NeoPoros: Una superficie con un historial de más de 10 años. Disponible en https://www.straumann.com/neodent/mx/es/website/professionals/products/surfaces/neoporos.html. (Consultado el 14/10/2020).
[33] FEDERA. Superficie Oseomimética. Disponible en http://federa.com.ar/page/linea-fis. (Consultado el 14/10/2020).
[34] BIOHORIZONS. Tapered Internal Catálogo de Producto 2010. Disponible en https://issuu.com/biodental/docs/tapered_internal__espa_ol_/6. (Consultado el 14/10/2020).

Tratamiento de electrodeposición con óxido de Titanio enriquecida en calcio y fósforo

Las marcas que utilizan este tipo de técnica para el tratamiento de superficie de sus implantes son: SMILETECH[35], BIOUNITE[36] y NOBEL BIOCARE[37].

Los implantes de SMILETECH y BIOUNTE cuentan con el tratamiento BIO-CAP [269], [270], el cual consiste en un tratamiento anódico de electrodeposición que genera una matriz de óxido de Titanio (TiO_2) enriquecida con calcio y fósforo.

La empresa NOBEL BIOCARE aplica a sus implantes el tratamiento TiUnite [271]. Este se caracteriza por una capa de TiO_2 gruesa y moderadamente rugosa con un alto grado de cristalinidad y propiedades osteoconductoras que aceleran la formación del hueso.

10.3 Información Adicional

En este anexo se presenta información adicional relacionada al aval para la recolección de datos otorgada por el Colegio de Odontólogos de la provincia de Misiones. Así mismo, se exhibe la nota presentada al Comité de Ética de la Universidad Nacional de Misiones, donde se solicitó autorización para realizar esta investigación, resguardando los datos personales de los pacientes (como nombre, apellido, número de documento o dirección) de manera que no sea posible identificarlo, ni tampoco relacionar el paciente con el especialista implantólogo.

[35] SmileTech. Tratamiento de Superficie. Disponible en http://www.red-dental.com/O_N67901.HTM. (Consultado el 14/10/2020).

[36] Biounite. Biounite® Nanotecnología a su alcance. Disponible en http://www.red-dental.com/O_N72601.HTM. (Consultado el 14/10/2020).

[37] Nobel Biocare. TiUnite. Disponible en https://www.nobelbiocare.com/en-us/tiunite. (Consultado el 14/10/2020).

Aval *Colegio de Odontólogos de la provincia de Misiones*

COLEGIO DE ODONTOLOGOS DE LA PROVINCIA MISIONES

Ley Nº XVII Nº 1

Personería Jurídica A-74

Belgrano 2135 Tel/Fax 437102 Email:info@colodmis.arnetbiz.com.ar (C.P. 3300) Posadas-Misiones

Carta presentación Plan de Tesis de la Lic- Nancy Ganz

La Lic. Nancy Ganz presentó al Colegio de Odontólogos su Plan de Tesis en el marco de la carrera de Doctorado en Ciencias Aplicadas que se dicta en la Facultad de Ciencias Exactas, Químicas y Naturales de la Universidad Nacional de Misiones y solicitó la colaboración de nuestra institución porque algunos temas del proyecto de investigación están referidas a las propiedades de los biomateriales que se utilizan en la práctica de implantes dentales.

Ante esta solicitud, el Colegio consideró auspicioso colaborar con dicha profesional y le cedió el registro de los colegas que realizan dichas prácticas con el propósito de facilitarle el acceso a reuniones o entrevistas que llevará a cabo la tesista con dichos profesionales.

Esta decisión se tomó con el compromiso mutuo entre la investigadora y el Colegio de respetar los recaudos éticos de confidencialidad sobre la identidad de los colegas y el resguardo y/o codificación de los datos obtenidos con la finalidad exclusiva de ser utilizados en su Tesis Doctoral.

MESA DIRECTIVA

Julio de 2016

COLEGIO DE ODONTOLOGOS PCIA. DE MISIONES Ley 169/57

Dr. ORLANDO ABEL BUS...
PRESIDENTE
COLEGIO DE ODONTOLOG...
DE LA PROV. DE MISIONES

Consentimiento Informado

Posadas, Misiones, 24 de Mayo de 2016.-

Señores Comité de Ética e Integridad de la Investigación

Secretaría de Investigación y Postgrado

Facultad de Ciencias Exactas Químicas y Naturales

Universidad Nacional de Misiones

Su Despacho.-

De mi mayor consideración:

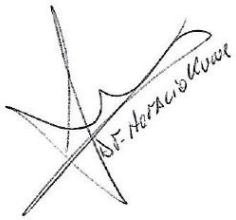
Por medio de la presente, quisiera comentarles que me desempeño como Becaria del CONICET y me encuentro cursando el Doctorado en Ciencias Aplicadas que se dicta en esta facultad y mi lugar de trabajo es el Instituto de Materiales de Misiones. Actualmente me encuentro realizando mi Tesis Doctoral sobre el tema "APLICACIÓN DE LA MINERIA DE DATOS PARA LA SELECCIÓN DE BIOMATERIALES", bajo la Dirección del Dr. Horacio D.KUNA y la Codirección de la Dra. Alicia E. Ares.

Esta investigación tiene por objetivo el mejoramiento de algoritmos de la minería de datos para obtener conocimientos previos de las propiedades de los biomateriales utilizados como implantes, para ello necesito contar con información suficientemente calificada sobre los pacientes y los implantes utilizados, como ser:

Datos de los Pacientes	Datos de la Intervención	Datos del Implante
- Fecha de nacimiento - Sexo - Ocupación - Enfermedades crónicas - Toma medicamento - Alergia - Fuma	- Estudios previos - Fecha de intervención - Lugar de intervención	- Tipo de material - Tipo de conexión - Longitud - Diámetro - Etapas - Sitio de Colocación - Vida útil - Causas de fallo - Procedencia - Proveedor - País de procedencia

Dada la gran cantidad de datos con el cual se trabajará para la realización de este trabajo, se hace imposible solicitar a cada paciente la autorización para el uso de sus datos de las historias clínicas. Como se puede apreciar no se solicitarán datos que identifiquen al paciente como nombre, apellido, número de documento o dirección, por esta razón solicito autorización a este Comité de Ética para llevar a cabo la investigación.

Sin más y agradecida por su tiempo, le saluda atentamente:



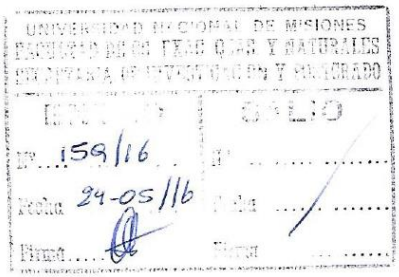
Lic. Nancy GANZ

D.N.I.: 36.458.838

Correo Electrónico: nan_bg1@hotmail.com

Nro. Telefónico: 3755-553213

Dra. ALICIA ESTHER ARES
PROF. TIT. CIENCIA DE MATERIALES
INSTITUTO DE MATERIALES DE MISIONES
CIC - UNaM - CONICET



ACTA Nº1

Posadas, 16 de junio de 2016

Señora
Secretaria de Investigación y Posgrado
Mgter Celina Vedoya
S / D

De nuestra mayor consideración:

Nos dirigimos a Ud. con referencia a la solicitud realizada por la Lic. Nancy Ganz respecto al manëjo de datos de pacientes para su tesis denominada " Aplicación de la minería de datos para la selección de biomateriales" dirigida por los Dres. Horacio Kuna y Alicia Ares.

Al respecto este Comité de Ética e Integridad de la Investigación considera que en la medida que se conserve la privacidad y autonomía moral del paciente, no existen inconvenientes para que se lleve a cabo ésta tesis.

Sin otro particular, saludamos a Ud. atentamente.

Esp. Silvia Caronia Dr. Juan A. Gallippo Dr. Marcelo A. Mierez Dr. Miguel E. Schmalko

Comité de Ética e Integridad de la Investigación

ES COPIA

A.S.C. VICTOR J. NARVAEZ
FIRMA AUTORIZADA
Secretaría Investigación y Postgrado
Fac. Cs. Ex. Qcas. y Nat. - UNaM