SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

# Tecnológico Nacional de México

## Centro Nacional de Investigación y Desarrollo Tecnológico

# Tesis de Maestría

## Análisis de datos genómicos para el diagnóstico temprano de osteosarcoma


presentada por

**Ing. Carlos Alberto Moncada Vázquez**


como requisito para la obtención del grado de

**Maestría en Ciencias de la Computación**


Director de tesis

**Dr. Gerardo Reyes Salgado**

Codirector de tesis

**Dr. Heriberto Manuel Rivera**



**Cuernavaca, Morelos, México. Junio del 2019.**


cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLOGICO NACIONAL DE MEXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Morelos a 19 de junio del 2019
OFICIO No. DCC/063/2019

Asunto: **Aceptación de documento de tesis**

**DR. GERARDO V. GUERRERO RAMÍREZ**
**SUBDIRECTOR ACADÉMICO**
**PRESENTE**

Por este conducto, los integrantes de Comité Tutorial del **Ing. Carlos Alberto Moncada Vázquez**, con número de control M17CE096, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado **"Análisis de datos genómicos para el diagnóstico temprano de osteosarcoma"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

M.C. Gerardo Reyes Salgado
Maestro en Ciencias de la Computación
2493370

REVISOR 1

Dra. Andrea Magadán Salazar
Doctorado en Ciencias Computacionales
10654097

REVISOR 2

Dr. Manuel Mejía Lavalle
Doctor en Ciencias Computacionales
8342472

C.p. M.E. Guadalupe Garrido Rivera - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLOGICO NACIONAL DE MEXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Mor., 20 de junio de 2019
OFICIO No. SAC/227/2019

**Asunto:** Autorización de impresión de tesis

**ING. CARLOS ALBERTO MONCADA VÁZQUEZ**
**CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS**
**DE LA COMPUTACIÓN**
**P R E S E N T E**

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Análisis de datos genómicos para el diagnóstico temprano de osteosarcoma"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

**A T E N T A M E N T E**
***Excelencia en Educación Tecnológica®***
*"Conocimiento y tecnología al servicio de México"*

SEP TecNM
CENTRO NACIONAL
DE INVESTIGACIÓN
Y DESARROLLO
TECNOLÓGICO
SUBDIRECCIÓN
ACADÉMICA

**DR. GERARDO VICENTE GUERRERO RAMÍREZ**
**SUBDIRECTOR ACADÉMICO**

C.p. Mtra. Guadalupe Garrido Rivera .- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

# DEDICATORIA

Dedico esta tesis a mis amigos, quienes fueron un gran apoyo emocional durante el tiempo de desarrollo de este trabajo de investigación.

A mis padres quienes me apoyaron todo el tiempo y con su amor, paciencia y esfuerzo me han permitido llegar a cumplir hoy un sueño más, gracias por inculcar en mí el ejemplo de esfuerzo y valentía.

A mis profesores quienes nunca desistieron al enseñarme, a ellos que continuaron depositando su esperanza en mí.

A todos los que me apoyaron para escribir y concluir esta tesis.

Para ellos es esta dedicatoria de tesis, pues es a ellos a quienes se las debo por su apoyo incondicional.

# AGRADECIMIENTOS

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por su apoyo y patrocinio para la realización de este proyecto de tesis.

De igual manera agradezco al Laboratorio de Biología de Sistemas y Medicina Traslacional (BSMT) de la Universidad Autónoma del Estado de Morelos (UAEM), por su valiosa colaboración y aportación de información para este proyecto.

Agradezco al Dr. Gerardo Reyes Salgado por ser el guía y asesor de esta tesis y al Dr. Heriberto Manuel Rivera por sus valiosas observaciones y aportaciones.

Sinceras gracias a: Dra. Andrea Magadán Salazar y Dr. Manuel Mejía Lavalle, quienes me asesoraron y atendieron mis dudas a lo largo de la realización de esta tesis

# RESUMEN

En esta tesis, se define un procedimiento que permite utilizar las técnicas de Aprendizaje Automático al problema de clasificación de datos masivos. Usando para ello los siguientes elementos: preprocesamiento de los datos, análisis de componentes independientes, entrenamiento de los clasificadores y la evaluación de los clasificadores.

El objetivo fundamental es proveer de una herramienta de diagnóstico molecular (HDM) adecuada para la solución de problemas complejos en el genoma, concretamente en la secuenciación de ADN de Osteosarcoma, basada en modelos predictivos, los cuales son obtenidos por inferencia automática de conocimiento a partir del manejo de los datos.

La experimentación realizada se hizo mediante un banco de datos proporcionado por el Laboratorio de Biología de Sistemas y Medicina Traslacional (BSMT) de la Universidad Autónoma del Estado de Morelos (UAEM). Los resultados obtenidos muestran que los métodos de clasificación propuestos logran hasta un 93.60% de precisión en la identificación de aquellos genes relacionados al Osteosarcoma.

# ABSTRACT

In this thesis, it defines a procedure that allows to use the techniques of Machine Learning to the problem of Big Data classification. Using the following elements: preprocessing of the data, independent components analysis, training of the classifiers and evaluation of the classifiers.

The main objective is to provide a molecular diagnostic tool (MDT) suitable for the solution of complex problems in the genome, specifically in the DNA sequencing of Osteosarcocma, based on predictive models, which are obtained by automatic inference of knowledge from the data management.

The experimentation was done through a data bank provided by the Systems Biology and Translational Medicine Laboratory (BSTM) of the Autonomous University of the State of Morelos (UAEM). The results obtained show that the proposed classification methods achieve up to 93.60% accuracy in the identification of those genes related to Osteosarcoma.

# GLOSARIO

**ADN (ácido desoxirribonucleico) o *DNA* (por sus siglas en inglés):** es un ácido nucleico que contiene la información de las características hereditarias de cada ser vivo y las secuencias para la creación de aminoácidos que generarán las proteínas vitales para el funcionamiento de los organismos [Checa, 2017].

**Genoma**: es el conjunto de instrucciones genéticas que se encuentra en una célula. En los seres humanos, el genoma consiste en 23 pares de cromosomas, que se encuentran en el núcleo, así como un pequeño cromosoma que se encuentra en las mitocondrias de las células. Cada conjunto de 23 cromosomas contiene aproximadamente 3.1 mil millones de bases de la secuencia de ADN [NIH, 2019].

**Cromosoma:** es una estructura condensada de ADN presente en las células que aparece en número constante en cada especie vegetal o animal. En los cromosomas se almacena gran parte de la información genética [Tolosa, 2017].

**SNP:** tipo de cambio más común en el ADN (moléculas dentro de las células que contienen información genética). El polimorfismo de un nucleótido se presenta cuando un nucleótido (elemento fundamental del ADN) es reemplazado por otro. Estos cambios pueden causar enfermedad y pueden afectar la forma en que una persona reacciona ante las bacterias, los virus, los medicamentos y otras sustancias. También se llama PSN [NIH 2019].

**Epigenómica**: sistema de regulación que controla la expresión de los genes sin afectar a la composición de los genes en sí mismos [Oryzon 2019].

# GLOSARIO DE LIBRERÍAS DE R

**Dplyr:** proporciona las funciones para eliminar los ausentes de valor o *missing values* (Dplyr, 2018).

**Tidyr:** realiza transformaciones en un D*ata Frame* (base de datos) con el fin de poder transformarlo a la estructura que se desea (Tidyr, 2018).

**caret:** proporciona múltiples funciones, para este trabajo se utilizaron para calcular la partición de los datos, los clasificadores y la matriz de confusión (Caret, 2018).

**FSelector:** realiza la técnica de análisis de componentes principales y otras técnicas de selección de atributos (FSelector, 2018).

**fastICA:** realiza la técnica de análisis de componentes independientes (fastICA, 2018).

**e1071:** proporciona diferentes agrupadores y clasificadores como: Random Forest, redes neuronales, RPART, máquina de soporte vectorial entre otras (e1071, 2018).

**doParallel:** proporciona un mecanismo necesario para ejecutar los núcleos en paralelo (doParallel, 2019).

**Iterators:** proporciona funciones para iterar sobre varias estructuras de datos en R como: matrices, vectores, listas y archivos, este paquete se utilizó para iterar el clasificador en la paquetería de doParallel (Iterators, 2019)

**ggplot 2:** son funciones para generar las gráficas (barras) que muestran la distribución de los datos y resultados de este trabajo (ggplot2, 2019).

# CONTENIDO

# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# CAPÍTULO 1. INTRODUCCIÓN

En el presente capítulo muestra la problemática, objetivos, metodología de solución y organización de este trabajo de tesis.

## 1.1 Motivación

El cáncer se caracteriza por la presencia de células anormales, las cuales se multiplican de manera descontrolada e invaden otros órganos del cuerpo (Floor, 2012). Así mismo, se define como un grupo heterogéneo de enfermedades debido a que se pueden observar diferencias entre los distintos tipos de cáncer; estas diferencias son la causa de alteraciones genéticas y epigenéticas que tienen la capacidad de iniciar una transformación maligna, sobrepuesto con la inestabilidad genómica y la continua adaptación y/o selección durante la evolución del tumor (Burrell, 2013) (Horne, 2015). En este sentido, el proyecto del genoma humano ofrece nuevas estrategias y oportunidades para estudiar el cáncer a una escala genómica, debido a que es posible detectar genes y procesos biológicos relacionados a esta enfermedad (Tang, 2013).

El Osteosarcoma, es un cáncer primario que se origina en los huesos, se define como un tumor maligno el cual produce tejido óseo patológico que se forma alrededor de las articulaciones y se presenta como una matriz no mineralizada (Klein, 2006). Uno de los grandes misterios del Osteosarcoma, es la alta predisposición a desarrollar metástasis en pulmones, ya que más del 90% de los casos de metástasis se reportan en este órgano. La falta de conocimiento al respecto es debido a que aún no se entiende el mecanismo por el cual las células de este cáncer migran a los pulmones y qué propiedades del microambiente de los pulmones es ideal para el desarrollo y proliferación.

El genoma humano alberga información acumulada de más de 6 millones de años en 46 cromosomas, en los cuales se encuentran 3.2 mil millones de pares de bases. Se estima que cada 300 pares de bases se genera un polimorfismo de un nucleótido

(*Single nucleotide polymorphism*, *SNP*, por sus siglas en inglés). Los *SNPs* están presentes en al menos el 1% de la población. Se estima que existen alrededor de 1,803,563,957 variantes depositadas en la base de datos de *SNPs* (*dbSNPs*), de los cuales 335,215,764 variantes están clasificadas por su localización en el cromosoma, y se encuentran en genes alrededor de 381,785,470. Un dato interesante es que sólo alrededor del 16 % de estos (113,862,023) han sido validados (Sherry, 2001). Dada la cantidad de datos a integrar, se propone implementar estrategias de Inteligencia Artificial asociada a la información genética, específicamente polimorfismos y estudios de asociación de genoma completo, con la intención de encontrar información útil que contribuya al diagnóstico de enfermedades (Atkinson, 2001).

Debido a la problemática dada en fenómenos tan complejos de la genética, como la simulación del efecto de medicinas, la predicción de enfermedades, etc. La bioinformática es una ciencia que surge de la necesidad de interpretar la información contenida en las secuencias de ADN, ARN y proteínas (Martínez, 2007), a través de la implementación de técnicas computacionales como lo es la inteligencia artificial (IA). Todas estas situaciones manejan gran cantidad de información y variables, de allí emerge la necesidad de apoyarse con la IA. (Arias, 2016). La IA se define como la capacidad computacional para realizar diferentes procesos de aprendizaje de máquina o "*Machine Learning*" (Kavakiotis, 2017); rama de la inteligencia artificial que construye y estudia sistemas capaces de aprender a partir de un conjunto de datos de entrenamiento y de mejorar procesos de clasificación y predicción.

El conocimiento extraído en bancos de datos complejos, por medio de técnicas de la IA, tiene la intención de interpretar un fenómeno biológico que se realiza mediante un proceso de tratamiento de datos multipasos (selección, preprocesamiento, transformación, clasificación, interpretación y evaluación) (Méndez, 2017). El objetivo de este conocimiento es de resolver el análisis de grandes cantidades de datos provenientes de secuenciación de ADN, cuya finalidad es generar sistemas que puedan reconocer patrones para un diagnóstico temprano por medio del aprendizaje automático artificial. Esto permitirá un sistema de salud más eficiente y con mejor utilización de recursos (Gartner, 2015).

## 1.2 Planteamiento del problema

Las herramientas de diagnóstico molecular (HDM), permiten detectar estados tempranos de patologías crónico-degenerativas humanas mediante la identificación de biomarcadores (indicaciones de un estado biológico). Para desarrollar las HDM, se requiere de sistemas que ayuden a reconocer y seleccionar los cambios genéticos a través de *Big Data*. Es por lo que, en este trabajo de tesis, se inició de la hipótesis de que es posible ayudar en la detección temprana de este tipo de patologías, incorporando técnicas de la inteligencia artificial, específicamente el reconocimiento de patrones, como la clasificación. Por el cual, analizando datos de secuencias genómicas es posible identificar patrones de dichas patologías (Osteosarcoma) cuyo análisis y gestión de la información genómica se está convirtiendo en una necesidad cada vez más requerida en la bioinformática.

## 1.3 Complejidad del problema

El volumen actual de datos ha superado las capacidades de procesamiento de los sistemas clásicos. Se ha entrado en la era del *Big Data* o datos masivos, que es definida con la presencia de gran volumen, velocidad y variedad en los datos. La necesidad de procesar y extraer conocimiento valioso de tal inmensidad de datos se ha convertido en un desafío considerable para científicos de datos y expertos en la materia. El valor del conocimiento extraído es uno de los aspectos esenciales de *Big Data*.

## 1.4 Objetivo general

Investigar, aplicar y evaluar técnicas de aprendizaje automático para el manejo de datos genómicos relacionados con el diagnóstico temprano de cáncer de hueso en humanos.

## 1.5 Objetivos específicos

a) Análisis y estudio de al menos dos algoritmos de aprendizaje automático artificial aplicado a información relacionada a cáncer de hueso y su diagnóstico.
b) Ejecución de los algoritmos seleccionados considerando características de muestras genéticas o *SNPs.*
c) Evaluación de algoritmos de aprendizaje automático seleccionados.
d) Identificación de patrones en muestras genéticas de cáncer de hueso.

## 1.6 Alcances y limitaciones

### A. Alcances

- Aplicar al menos una técnica de aprendizaje automático para el análisis de datos genómicos.
- El sistema es capaz de hacer una predicción con nuevos casos de Osteosarcoma.
- Evaluar el clasificador con diferentes métricas (precisión, sensibilidad, especificidad, *F-measure*, AUC).

### B. Limitaciones

- El sistema sólo es específico para el reconocimiento de patrones de Osteosarcoma.
- Se utilizará el banco de datos es proporcionado por el Laboratorio de Biología de Sistemas y Medicina Traslacional (BSMT) de la Universidad Autónoma del Estado de Morelos (UAEM).
- No se utilizará *software* comercial.

## 1.7 Metodología de solución

En esta sección se describe la metodología que permite resolver el objetivo planteado en este trabajo de investigación (ver Figura 1). Se compone de cuatro fases; preprocesamiento de los datos, extracción de características, entrenamiento de los clasificadores y la evaluación de los mismos.

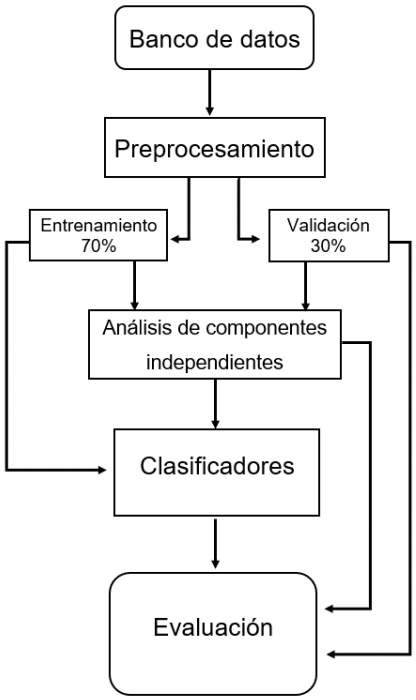



Figura 1.1. Esquema de método de solución.

### a) Preprocesamiento

La preparación de los datos se realizó de la siguiente manera:

- Transformación de los datos: de tipo cadena a tipo nominal.
- Eliminación de aquellos datos con ausencia de información (N/A).
- Partición de la base de datos, para entrenamiento (70%) y validación (30%).

Para esta etapa se utiliza la librería “dplyr”, “tidyr” del lenguaje R (ver glosario de librerías).

### b) Análisis de componentes independientes

Se decidió evaluar el conjunto de variables con el método de extracción de características; análisis de componentes independientes. Debido a que el vector de características se integra de pocos atributos, se decidió utilizar el análisis de componentes independientes para tener el mismo número de variables, eliminando el ruido presente en los datos, por lo que no se realizó una selección de variables para obtener las más representativas sino hay una independencia entre las variables para disminuir el ruido por lo cual se obtiene el mismo número de variables que se tenían originalmente.

En esta etapa se utiliza las *librerías "tidyr", "caret", "Fselector", "fastICA", "ggplot2*" de R (ver glosario de librerías).

### c) Clasificación

En esta fase se aplicaron y evaluaron dos clasificadores. Se seleccionaron los algoritmos de Random Forest y XGBoost, debido a las características de procesar múltiples variables de tipo cadena, cualitativos y cuantitativos.

Se utiliza las librerías *"caret", "e1071", "doParallel", "iterators"* en R (ver glosario de librerías de R).

### d) Análisis de resultados (evaluación de la Respuesta)

La evaluación de la respuesta de los clasificadores permite:

- Implementar las métricas de evaluación (precisión, especificidad, sensibilidad, *F-measure y AUC*) para determinar el rendimiento del sistema.
- Comparar los resultados (métricas) con los dos algoritmos de aprendizaje automático seleccionados.

Se utiliza las librerías "*caret", "ggplot*" en R (ver glosario de librerías).

## 1.8 Organización de la tesis

El presente documento se encuentra estructurado en 6 capítulos.

**Capítulo 2. Antecedentes y Estado del Arte**

El capítulo 2 muestra los antecedentes y trabajos relacionados con el preprocesamiento de los datos, análisis de componentes independientes y aprendizaje automático o *Machine Learning*; así como un análisis general de cada artículo.

**Capítulo 3. Marco Teórico**

El capítulo 3 muestra los conceptos y descripción de los métodos utilizados en este trabajo de investigación.

**Capítulo 4. Metodología de solución**

El capítulo 4 muestra la metodología utilizada para el desarrollo de este trabajo.

**Capítulo 5. Pruebas y resultados**

El capítulo 5 presenta las diferentes pruebas realizadas con los clasificadores. Así como los resultados obtenidos que demuestran la efectividad de los algoritmos propuestos.

**Capítulo 6. Conclusiones**

El capítulo 6 presenta información detallada de los resultados obtenidos en las pruebas. De igual forma se menciona el trabajo futuro que se pueden desarrollar a partir de este trabajo de investigación, así como las aportaciones y productos adicionales realizados en el transcurso del periodo de estudios de maestría.

# CAPÍTULO 2. ANTECEDENTES Y ESTADO DEL ARTE

En este capítulo muestra los trabajos realizados en el CENIDET y los trabajos externos relacionados a los temas de preprocesamiento de los datos, análisis de componentes independientes y aprendizaje automático o *Machine Learning*; así como un análisis general de cada artículo.

## 2.1 Antecedentes

En el CENIDET, en el área de ciencias computacionales se han desarrollado tesis y artículos de métodos computacionales para el área de la salud, específicamente el cáncer en datos poblacionales. En nuestro caso, se aborda el estudio de un problema de salud, como el Osteosarcoma, pero a partir de datos genómicos de una secuenciación de ADN de personas. A continuación, se mencionan los artículos y tesis revisadas para esta investigación.

### 2.1.1 A data preparation methodology in data mining applied to mortality population databases (Pérez, 2015)

Este artículo muestra una nueva metodología de preparación de datos orientada al dominio epidemiológico en la que se han identificado dos conjuntos de tareas a realizar: preparación de datos generales y preparación de datos específicos. Para ambos conjuntos, se opta una guía para el proceso estándar interindustrial para la minería de datos (CRISP-DM).

Para la validación de la metodología propuesta, se desarrolló un sistema de minería de datos y todo el proceso se aplicó a las bases de datos de mortalidad real. Los

resultados fueron buenos ya que se observó que el uso de la metodología redujo algunas de las tareas que consumían mucho tiempo y el sistema de minería de datos mostró hallazgos de patrones desconocidos y potencialmente útiles para los servicios de salud pública en México.

### 2.1.2 **Data mining system oriented to population databases for cancer** (Pérez, 2007)

Se aborda el problema de descubrir patrones de interés en las bases de datos poblacionales para el cáncer. Se implementó un sistema de minería de datos, que se desarrolló específicamente para este tipo de bases de datos. Se utilizó el algoritmo *K-means* para la generación de patrones, lo que permite expresar patrones como regiones o grupos de distritos con afinidad en sus parámetros de localización y tasa de mortalidad.

Como resultado de la aplicación del sistema, se generó un conjunto de patrones de agrupación, que define la distribución de la mortalidad por cáncer de estómago en los distritos mexicanos.

### 2.1.3 **Minería de datos orientada al big data en el área de salud** (Pérez, 2016)

En este documento de tesis, se muestra que es factible el desarrollo de un prototipo de un sistema de minería de datos orientado a manejar grandes instancias como las que se presentan en el paradigma de *Big Data* en el dominio de la salud. En particular, el objetivo del prototipo es encontrar regiones del territorio mexicano y estadounidense con altas tasas de incidencia de mortalidad por diabetes, a partir de bases de datos poblacionales.

En esta investigación propuso el uso del algoritmo n-means (Pérez, 2016) para realizar la tarea de agrupamiento en el proceso de Minería de Datos. Para realizar las tareas de visualización se propuso un módulo cartográfico que hace uso de los

mapas proporcionados por Google Maps, los cuales comprenden el territorio de México y de los Estados Unidos.

El prototipo se validó de manera sistemática con un conjunto de casos de prueba diseñado para tal fin. Como base para las pruebas se usaron los datos de mortalidad de los censos del año 2000 y 2010. Es destacable que el volumen de datos para la experimentación fue del orden de los tres gigabytes, con más de cuatro millones de registros.

Los resultados obtenidos para esta enfermedad hacen evidente la utilidad de las ciencias computacionales y en particular de la minería de datos en el área de salud, ya que proporcionan elementos de apoyo para la toma de decisiones de los funcionarios y autoridades encargadas de la salud de la población.

### 2.1.4 Desarrollo de un prototipo para la aplicación de técnicas de minería de datos sobre una base de datos real de base poblacional de cáncer (Barrón, 2008)

Este trabajo exploratorio tiene como objetivo determinar si es posible, mediante el uso de técnicas de minería de datos, explorar los datos de una base de datos poblacional real de mortalidad por cáncer en México, y encontrar patrones de interés dentro de los datos.

Partiendo de la naturaleza de los datos, se decidió generar grupos de municipios, de acuerdo con las similitudes de los valores de sus tasas de mortalidad y de la posición geográfica de sus cabeceras municipales.

Para ello, se creó un almacén de datos con registros de tasas de mortalidad por cáncer de pulmón correspondiente al año 2000, y un prototipo con dos módulos: el primero, analiza los datos con la implementación del algoritmo k-means de Weka-3-4, y el segundo es un módulo de visualización de resultados que plasma

los modelos obtenidos en un mapa de la república mexicana con la finalidad de mostrar la distribución de los grupos de municipios.

El resultado es un prototipo que permite visualizar patrones expresados como regiones con altas tasas de mortalidad por cáncer de pulmón en el territorio nacional.

## 2.2 Estado del arte

En este apartado se muestra los trabajos relacionados de preprocesamiento, análisis de componentes independientes y aprendizaje automático o *Machine Learning*,

### 2.2.1 Preprocesamiento de los datos

Para la primera etapa, se estudia el siguiente artículo.

1. **Big data: preprocesamiento y calidad de datos** (García, 2016)

En este artículo se menciona que la eficacia de los algoritmos de extracción de conocimiento depende en gran medida de la calidad de los datos, la cual puede ser garantizada por los algoritmos de preprocesamiento. Sin embargo, en esta era de *Big Data*, los algoritmos de preprocesamiento tienen dificultades para trabajar con tal cantidad de datos, siendo necesario nuevos modelos que mejoren su capacidad de escalado. El objetivo de este trabajo es presentar la importancia del preprocesamiento de datos en *Big Data*, así como, estudiar las herramientas y técnicas de análisis de datos que dan soporte a la tarea del preprocesamiento de datos masivos.

Tabla 2.1. Artículo de preprocesamiento

| Artículo | Objetivo | Técnicas o áreas | Conclusión |
|---|---|---|---|
| *Big Data:* Preprocesamiento y calidad de datos (García, 2016) | Presentar la importancia del preprocesamiento de datos en *Big Data*, así como, estudiar las herramientas y técnicas de análisis de datos que dan soporte a la tarea del preprocesamiento de datos masivos. | *Big Data*, preprocesamiento y herramientas para el análisis de datos masivos. | En este trabajo se estudia la importancia del preprocesamiento de datos en *Big Data*. Se presenta una revisión de las tecnologías de *Big Data*, herramientas de análisis de datos y técnicas y algoritmos disponibles para el preprocesamiento de datos masivos. |

### 2.2.2 Análisis de componentes independientes

Para la segunda etapa se estudian los siguientes artículos.

#### 2.2.2.1 Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes (Biton, 2014)

La extracción de información relevante de datos a gran escala, da buenas oportunidades en cancerología. En este artículo se aplica el análisis de componentes independientes (ACI) a los conjuntos de datos del transcriptoma de cáncer de vejiga y se interpretan los componentes mediante el análisis de enriquecimiento de genes y la información de procesamiento molecular, clínico-patológica asociada al tumor. Se identificaron los componentes asociados con los procesos biológicos de las células tumorales o el microentorno del tumor, y otros componentes revelaron sesgos técnicos. La aplicación de ACI a diferentes tipos de cáncer identificó componentes compartidos del cáncer y componentes específicos del cáncer de vejiga. Donde se pudo lograr una independencia del cáncer de vejiga respecto a otros cánceres

#### 2.2.2.2 A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results (Khiabani, 2016)

El objetivo de este estudio fue mostrar la aplicación de algunos métodos de selección de variables, generalmente utilizados en la minería de datos, para un estudio epidemiológico, tales como Info Gain, Relief, OneR, Wrapper, donde fueron evaluados con el algoritmo de Random Forest y la métrica de AUC

Tabla 2.2. Artículos de análisis de componentes independientes

| Artículo | Objetivo | Técnicas o áreas | Conclusión |
|---|---|---|---|
| Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes (Biton, 2014) | Aplicar análisis de componentes independientes a diferentes tipos de cáncer para identificar componentes compartidos del cáncer y componentes específicos del cáncer de vejiga. | Análisis de componentes independientes | La aplicación de ACI para lograr la independencia de las variables del cáncer de vejiga con el fin de evitar el traslape de información con los otros tipos de cáncer. |
| A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results (Khiabani, 2016) | El objetivo de este estudio fue mostrar la aplicación de algunos métodos de selección de variables, generalmente utilizados en el procesamiento de datos. | Info Gain, Relief, OneR, Wrapper, Random Forest y AUC | Los experimentos mostraron que los métodos de selección de variables utilizados en el procesamiento de datos podrían mejorar el rendimiento de los modelos de predicción clínica. |

### 2.2.3 Aprendizaje automático artificial

Para esta etapa se presentan los siguientes artículos.

#### 2.2.3.1 Machine learning in genomic medicine: a review of computational problems and data sets (Leung, 2016)

En este documento se aborda la problemática de la medicina genómica, cuyo objetivo es determinar cómo las variaciones en el ADN de las personas pueden afectar el riesgo de diferentes enfermedades y cómo así encontrar explicaciones causales para diseñar terapias o tratamientos específicos, el enfoque de aprendizaje automático puede ayudar a modelar la relación del ADN y la cantidad de variables celulares pueden estar asociadas con riesgo de enfermedad (Figura 2.1).

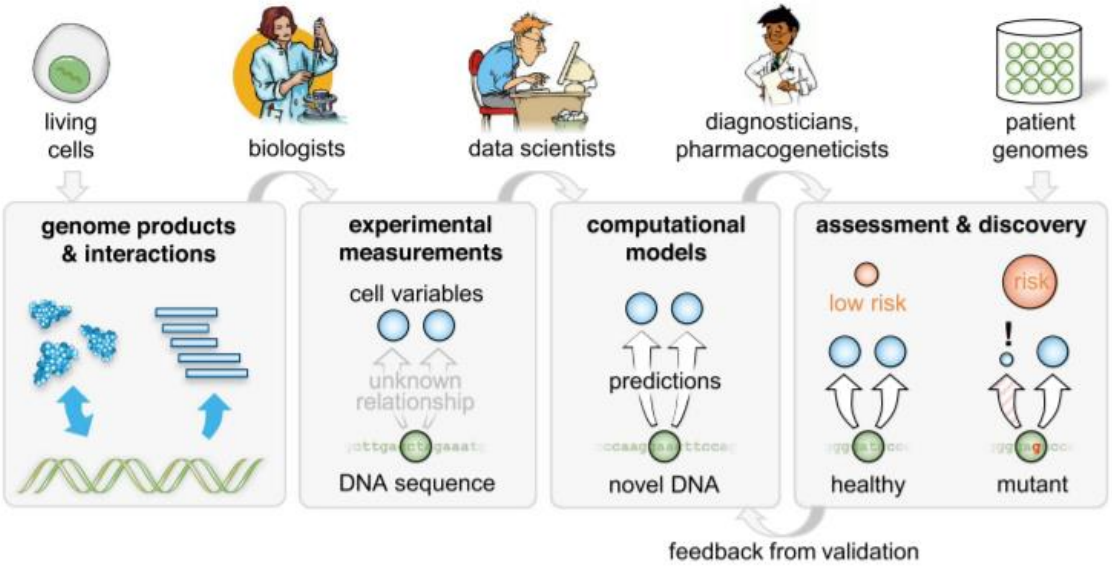



Figura 2.1. Bioinformática (Leung, 2016).

El algoritmo de aprendizaje automático estudiado para dicha relación son las redes neuronales recurrentes (RNN por sus siglas en inglés) y arboles de decisión. En este artículo sólo estudian cómo estos algoritmos podrían resolver el problema sujeto a estudio, pero no aplican dichos algoritmos.

#### 2.2.3.2 Classification of human cancer diseases by gene expression profiles (Salem, 2017)

Se realiza una revisión de otros artículos cuyas bases de datos son: datos de tumores de próstata, conjunto de datos de tumores pulmonares, conjunto de datos de leucemia y cáncer de pulmón de células no pequeñas (CPCNP), donde unos trabajos realizan selección de características con algoritmos genéticos, ganancia de información y en otros trabajos realizan extracción de características con el método de análisis de componente principales. Posteriormente a la reducción de la dimensionalidad, se aplica un clasificador ya sea vecino más cercano o una máquina de soporte vectorial, y los resultados son evaluados con las métricas de precisión, *recall*, exactitud y *F-measure* (Figura 2.2).

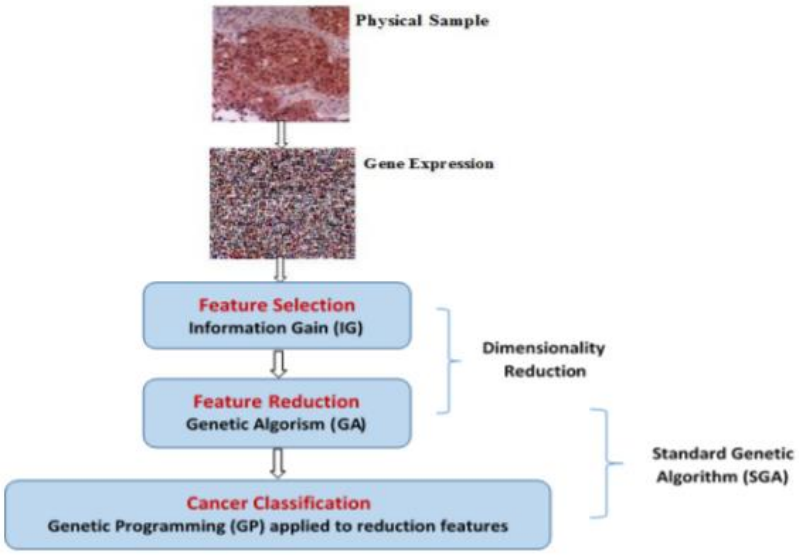



Figura 2.2. Metodología empleada (Salem, 2017).

#### 2.2.3.3 Gene expression data classification using support vector machine and mutual information-based gene selection (Arockia, 2015)

En este artículo utilizó una base de datos de cáncer colon de 62 tejidos y 2000 valores de expresión genética y la base de datos del linfoma que es un término

amplio que abarca una variedad de cánceres del sistema linfático. El conjunto de datos de linfoma incluye 45 tejidos x 4026 genes. Se utilizó una Máquina de vector soporte con diferentes *kernels* como: *kernel* lineal, *kernel* cuadrático, *kernel* polinomial y *kernel* función de base radial. Además, se realizó una comparación con otros métodos, cómo, vecino más cercano (knn por sus siglas en inglés) (Figura 2.3) y redes neuronales artificiales, debido a su capacidad para mapear los datos de entrada-salida.

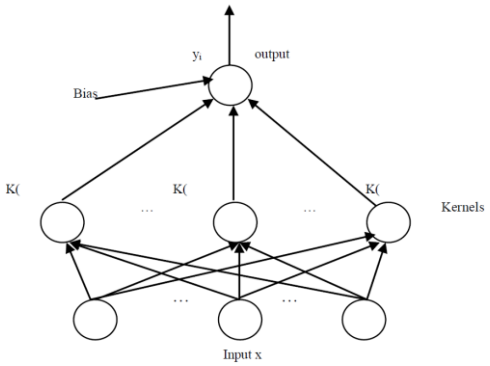



Figura 2.3 Máquina de vector soporte (Arockia, 2015).

Se concluyó que las Máquinas de Vector Soporte con *kernel* lineal ofrece un mejor índice con un 74% en la base de datos de colon y en el segundo caso con la base de datos de linfoma se tiene un empate con 100% de correctamente clasificadas, entre los métodos de redes neuronales y Máquinas de Vector Soporte con *kernel* lineal.

#### 2.2.3.4 **Gene expression based cancer classification** (Tarek, 2017)

En este artículo se evalúa el método de vecino más cercano (K-NN) con diferentes bases de datos, tales como: leucemia, cáncer de mama y colon, y a través del valor singular de descomposición que es una selección de características. Obtuvieron resultados que fueron graficados en un modelo de evaluación que fue la curva ROC (Figura 2.4).

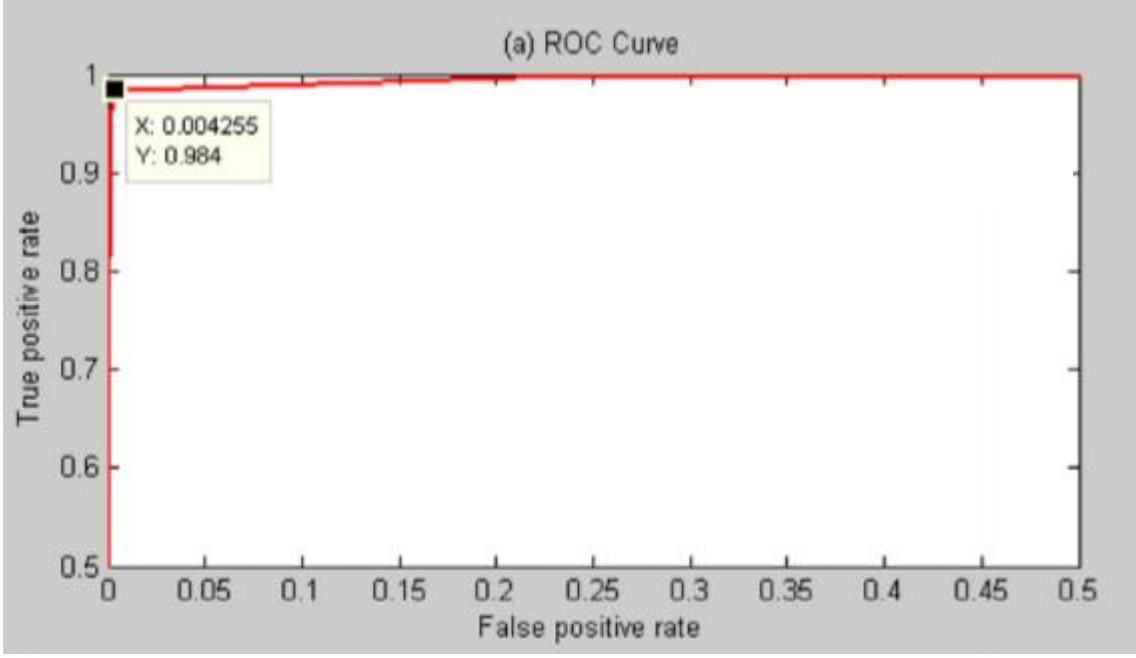



Figura 2.4. Resultado curva ROC (Tarek, 2017).

#### 2.2.3.5 **Machine learning based approaches for cancer classification using gene expression data** (Bhola, 2015)

Se realiza una comparación de diferentes algoritmos de clasificación como: Naive Bayes, K-NN (vecino más cercano), Random Forest, Máquina de Vector Soporte, Bagging, Adaboost. Se aplicaron a base de datos de leucemia, próstata, cáncer de mama, cáncer de pulmón y el cáncer de linfoma. Además, utilizaron métodos de selección de características tales como: ganancia de información (IG), máquina de vector de soporte con eliminación de características recursivas (SVMRFE), optimización de enjambre de partículas (PSO), selección de características basadas en la correlación rápida (FCBF). En donde cada uno de los clasificadores fue evaluado con cada uno de los selectores de características (Tabla 2.3).

Tabla 2.3. Resultados de la evaluación de clasificadores (Bhola, 2014).

| Classifier | IG | RelifF | SVMRFE | PSO | FCFB |
|---|---|---|---|---|---|
| NB | 94.1 | 92.1 | 97.0 | 96.1 | 95.1 |
| k-NN | 96.1 | 93.1 | 97.5 | 89.7 | 94.6 |
| RF | 91.6 | 90.1 | 93.6 | 88.2 | 92.6 |
| SVM | 68.5 | 68.5 | 94.6 | 92.6 | 94.1 |
| Bagging | 94.1 | 92.6 | 92.1 | 88.7 | 93.6 |
| AdaBoost | 78.3 | 78.3 | 77.3 | 72.4 | 75.4 |

#### 2.2.3.6 Classifying osteosarcoma patients using machine learning approaches (Li, 2017)

En este artículo evaluan los algoritmos de Random Forest, Regresión logística, y Máquina de Vector Soporte en cáncer de hueso, utilizando una base de datos creada a través de pacientes con estudios sistemáticos de los procesos químicos, donde comparan los clasificadores con tres casos, control sano, casos de tumor benigno y Osteosarcoma (Figura 2.5) . Se realizó una partición de los datos de forma aleatoria con 60% para entrenamiento y un 40% para prueba, en el cual Random Forest fue el algoritmo con mejor desempeño con un 97% de exactitud en comparación de los otros dos (Figura 2.5).

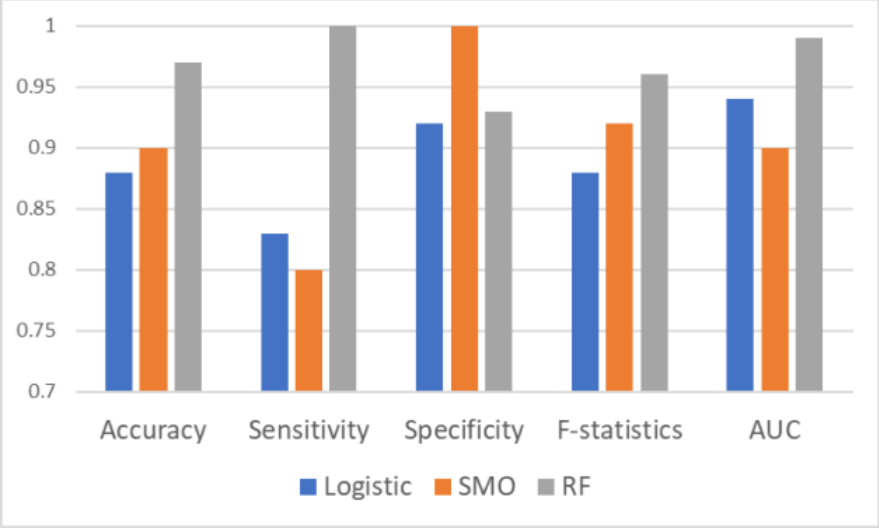



Figura 2.5. Resultados con los tres casos (Li, 2017).

Para la clasificación más delicada entre pacientes con tumores benignos y pacientes con osteosarcoma (Figura 2.6) , ninguno de los tres métodos tuvo éxito, la Regresión Logística fue el insatisfactorio.

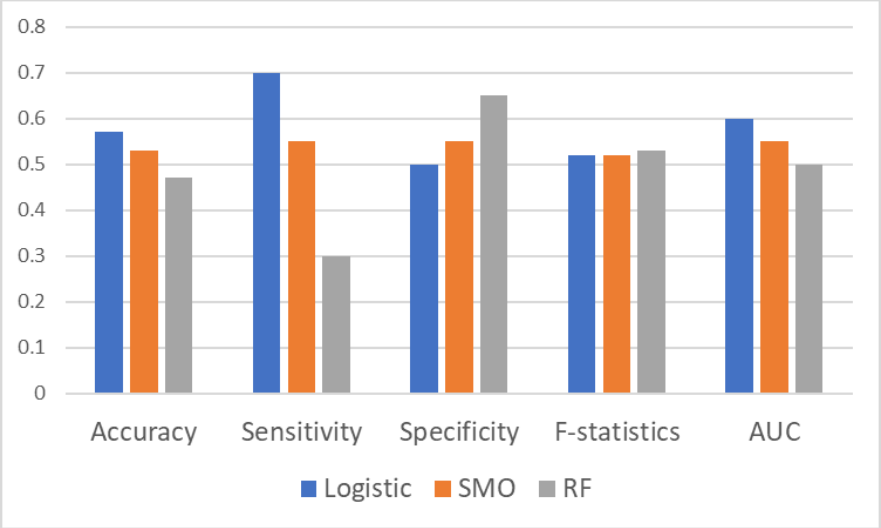



Figura 2.6. Resultados con 2 casos (Li, 2017).

Se demuestra que Random Forest es un clasificador exitoso mientras se realiza en datos de controles sanos y pacientes con tumor benigno y osteosarcoma.

#### 2.2.3.7 **Random Forests for big data** (Genuer 2017)

El *Big Data* siempre involucra datos masivos, y recientemente, algunos métodos estadísticos se han adaptado para procesarlos, como los modelos de regresión lineal, los métodos de agrupamiento y técnicas de aprendizaje automático. Basados en árboles de decisión combinados con ideas de agregación y *Bootstrap*, Random Forest fue desarrollado por Breiman en 2001. Son un método estadístico poderoso que permite considerar problemas de regresión, así como una clasificación de dos clases y clases múltiples. Enfocándose en los problemas de clasificación, este documento propone una revisión selectiva de las propuestas disponibles que tratan sobre escalar bosques aleatorios a problemas de *Big Data*.

#### 2.2.3.8 **Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia** (Pan, 2017)

La predicción de recaída en la leucemia linfoblástica aguda (LLA) infantil es un factor crítico para el éxito del tratamiento y la planificación del seguimiento. El objetivo de este trabajo es construir un modelo de predicción de recaída de LLA basado en algoritmos de aprendizaje automático. En el conjunto de datos con LLA recién diagnosticados, se empleó un algoritmo de selección de características para encontrar las variables más discriminatorias. Para permitir una estimación imparcial del modelo de predicción para nuevos pacientes, además de los conjuntos de pruebas divididas de 150 pacientes, se introdujo otro conjunto de datos independientes de 84 pacientes para evaluar el modelo. El modelo de Random Forest con 14 características logró una precisión con validación cruzada k=10, de 0.827 ± 0.031, con un área bajo la curva de 0.902 ± 0.027. Por lo que se sabe, este es el primer estudio que utiliza modelos de aprendizaje automático para predecir la recaída de la LLA en la infancia basándose en los datos médicos del registro médico electrónico, que facilitará aún más los tratamientos de estratificación.

#### 2.2.3.9 **Using machine learning algorithms for breast cancer risk prediction and diagnosis** (Asri, 2016)

Este artículo comenta que los métodos de clasificación y extracción de datos son una forma efectiva de analizar los datos. Especialmente en el campo de la medicina, donde esos métodos son ampliamente utilizados en el diagnóstico y análisis para tomar decisiones. En este documento, se hizo una comparación del rendimiento entre diferentes algoritmos de aprendizaje automático tales como: Máquina de Vector Soporte (SVM), árbol de decisión (C4.5), Naive Bayes (NB) y vecinos más cercanos (K-NN) con la base de datos del Wisconsin Breast Cancer (original).

Los resultados se evaluaron en términos de exactitud, precisión, sensibilidad y especificidad. Los resultados experimentales muestran que SVM obtuvo la más alta precisión alcanzando un 97.13% (Figura 2.7) con la tasa de error más baja.

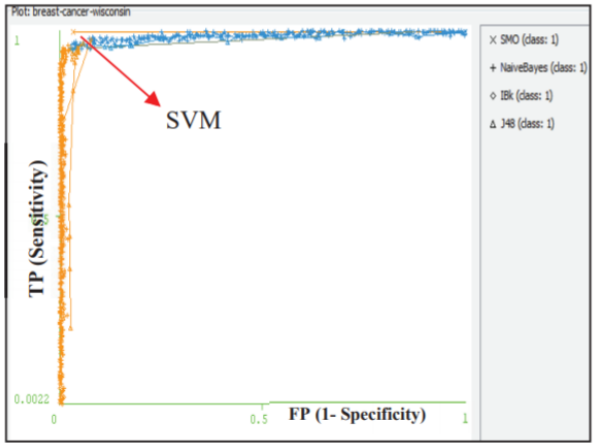



Figura 2.7. Curva Roc, Máquina de Vector Soporte (Asri, 2016).

### 2.2.3.10 **A support vector machine classifier for the prediction of osteosarcoma metastasis with high accuracy** (He, 2017)

En este estudio, se analizaron perfiles de expresión génica de osteosarcoma (OS) (ver Figura 2.8) para identificar genes críticos asociados con metástasis por medio del algoritmo de máquina de soporte vectorial, se experimentó con 4 diferentes bases de datos privadas.

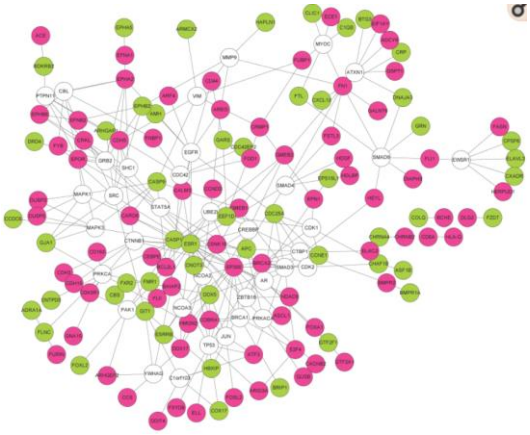


Figura 2.8. Perfiles de expresión genética (He, 2017).

El clasificador SVM mostró una alta precisión de predicción en los 4 conjuntos de datos, con una precisión de 100%, 100%, 92.6% y 100%, respectivamente. En general, se construyó y validó un clasificador SVM con alta precisión de predicción, en el que también se revelaron genes clave asociados con metástasis en OS.

#### 2.2.3.11 Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls (Bheravan, 2018)

En este artículo se propone un enfoque de aprendizaje automático eficaz para identificar un grupo de polimorfismos de un nucleótido (*SNP*) que interactúan, y que son los que más contribuyen al riesgo de cáncer de mama.

Se adopta la técnica de XGBoost para obtener los *SNP* más significativos para capturar interacciones no lineales complejas *SNP-SNP*, y, en consecuencia, se obtiene un grupo de *SNP* interactivos con alto potencial predictivo de riesgo de cáncer de mama (Figura 2.9).

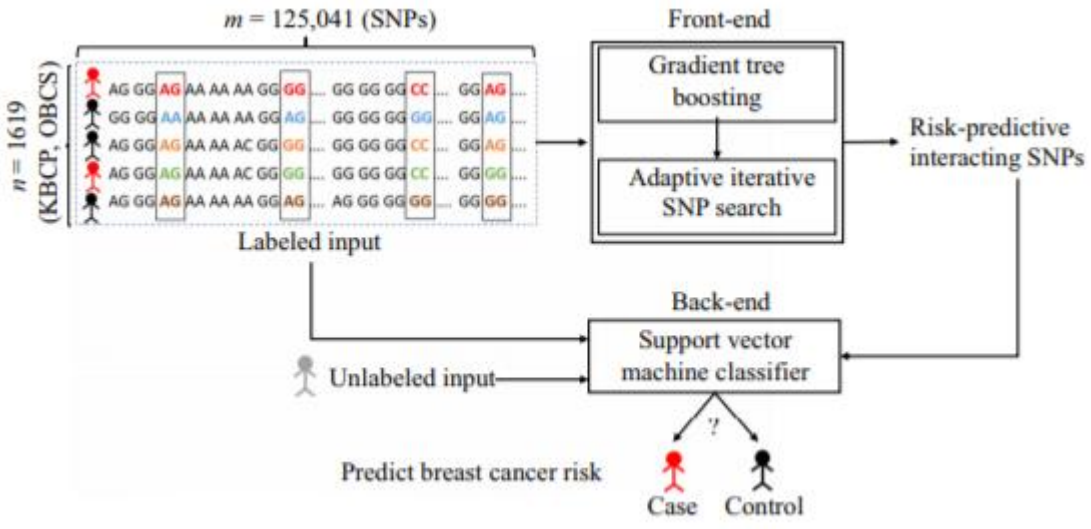



Figura 2.9 Metodología utilizada con *SNP* (Bheravan, 2018).

#### 2.2.3.12 **Pathway analysis using XGBoost classification in biomedical data** (Dimitrakopoulos, 2018)

En este trabajo proponen un esquema de clasificación basado en XGBoost, un algoritmo de clasificación reciente, basado en árboles, con el fin de detectar las vías más discriminatorias relacionadas a una enfermedad. Posteriormente, se extraen las subvías con respecto a su capacidad para clasificar correctamente las muestras de diferentes condiciones experimentales (Figura 2.10). Además, se compararon con otros clasificadores cómo, máquina de soporte vectorial, regresión lineal, K vecinos cercanos, Random Forest.

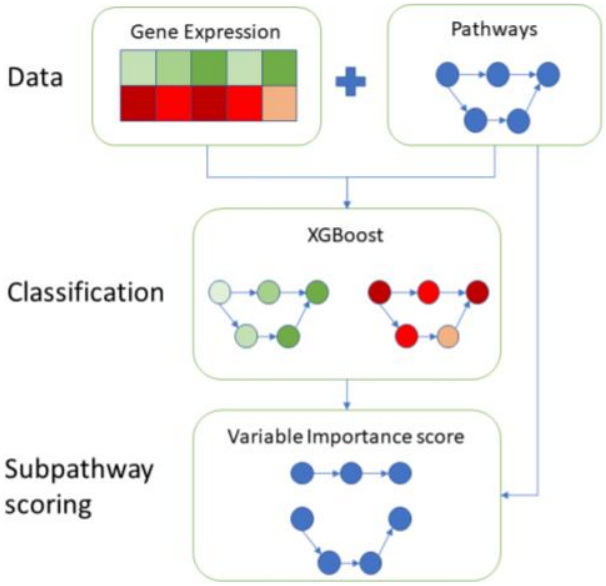



Figura 2.10. Esquema de clasificación para subvías (Dimitrakopoulos, 2018).

#### 2.2.3.13 **An empirical study of machine learning algorithms for cancer identification** (Turki, 2018)

En este artículo, se utilizan algoritmos de aprendizaje automático cómo, AdaBoost, deepboost, xgboost y máquina de soporte vectorial, y se evaluaron utilizando el área media bajo la curva en datos clínicos reales relacionados con el cáncer de tiroides, cáncer de colon y cáncer de hígado.

La base de dato de cáncer de tiroides cuenta con 25 ejemplos y 1146 características, cáncer de colon, cuenta con 21 ejemplos 5607 características y el cáncer de hígado consta de 21 ejemplos y 5607 características, las tres bases de datos únicamente cuentan con dos clases, tumor benigno y tumor maligno.

## 2.3 Discusión

Los trabajos realizados en el CENIDET muestran algunas problemáticas que se han intentado resolver por medio del aprendizaje automático. Sin embargo, ninguno de estos trabajos aborda el área de la genómica, por lo que este trabajo de investigación se orienta para abordar esta área. En la Tabla 2.4 se muestra el objetivo de cada trabajo del Estado del Arte.

Tabla 2.4. Artículos de aprendizaje automático

| Articulo | Objetivo | Técnicas o áreas | Conclusión |
|---|---|---|---|
| Machine learning in genomic medicine: a review of computational problems and data sets (Leung, 2016) | Determinar cómo las variaciones en el ADN de las personas pueden afectar el riesgo de diferentes enfermedades y cómo así encontrar explicaciones causales para diseñar terapias o tratamientos específicos. | Redes neuronales recurrentes y árboles de decisión | El enfoque de aprendizaje automático puede ayudar a modelar la relación del ADN y la cantidad de variables celulares, pueden estar asociadas con riesgo de enfermedad. |
| Classification of human cancer diseases by gene expression profiles (Salem et al., 2017) | Se realiza una revisión de otros artículos cuyas bases de datos son: datos de tumores de próstata, conjunto de datos de tumores pulmonares, conjunto de datos de leucemia y cáncer de pulmón de células no pequeñas (CPCNP), | Info Gain, análisis de componentes principales, K vecinos cercanos y máquina de soporte vectorial | Muestra que la mejor combinación fue Info Gain con máquina de soporte vectorial en las diferentes bases de datos. |

| Artículo | Objetivo | Técnicas o áreas | Conclusión |
|---|---|---|---|
| Gene expression data classification using support vector machine and mutual information-based gene selection (Arockia, 2015) | Clasificar el cáncer de colon y linfoma a través de la expresión genética de la enfermedad. | K vecinos cercanos y máquina de soporte vectorial con diferentes *kernels* | Máquinas de vector soporte con *kernel* lineal ofrece un mejor índice de clasificación. |
| Gene expression based cancer classification (Tarek, 2017) | Evaluar el método de vecino más cercano (KNN) con diferentes bases de datos, tales como: leucemia, cáncer de mama y colon | K vecinos cercanos | Se presenta un nuevo sistema para la clasificación del cáncer basado en los perfiles de expresión génica. |
| Machine learning based approaches for cancer classification using gene expression data(bhola, 2015) | Comparar diferentes algoritmos como: N*aive Bayes*, K-NN (vecino K cercano), Random Forest, máquina de vector soporte, y *Adaboost*. En base de datos de leucemia, próstata, cáncer de mama, cáncer de pulmón y el cáncer de linfoma. Utilización de selección de características | N*aive Bayes*, K-nn, Random Forest, máquina de vector soporte, y *Adaboost,* y selección de características | A través de este trabajo de investigación, se espera comprender mejor el problema de la clasificación del cáncer por medio de datos de expresión genética. |

| Artículo | Objetivo | Técnicas o áreas | Conclusión |
|---|---|---|---|
| Classifying osteosarcoma patients using machine learning approaches (Zhi Li, 2017). | Evaluar y comparar los algoritmos de Random Forest, regresión ligistica, y máquina de soporte vectorial en cáncer de hueso | Random Forest, regresión ligistica, y máquina de soporte vectorial | Se demuestra que Random Forest es un clasificador exitoso mientras se realiza en datos de controles sanos y pacientes con tumor benigno y Osteosarcoma. |
| Random Forest for big data (Genuer, 2017) | Random Forest para procesar datos masivos. | Random Forest y su paralelización | Random Forest es un ejemplo interesante entre los métodos ampliamente utilizados en el aprendizaje automático, ya que ofrece varias formas de tratar con datos masivos de diferentes contextos. |
| Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia (Pan, 2017) | Random Forest para la predicción de recaída en la leucemia linfoblástica aguda (LLA). | Random Forest, aprendizaje automático. | Primer estudio que utiliza modelos de aprendizaje automático para predecir la recaída de la LLA en la infancia basándose en los datos del registro médico electrónico |

| Artículo | Objetivo | Técnicas o áreas | Conclusión |
|---|---|---|---|
| Using machine learning algorithms for breast cancer risk prediction and diagnosis (Asri, 2016) | Comparación de diferentes algoritmos de aprendizaje automático: máquina de soporte vectorial (MVS), árbol de decisión (C4.5), Naive Bayes (NB) y k vecinos cercanos (k-NN) con la base de datos del Wisconsin Breast Cancer | Máquina de soporte vectorial, árbol de decisión (C4.5), Naive Bayes y k vecinos cercanos | Los resultados experimentales muestran que MVS obtiene la más alta precisión (97.13%) con la tasa de error más baja. |
| A support vector machine classifier for the prediction of osteosarcoma metastasis with high accuracy (He, 2017) | Analizar perfiles de expresión génica de Osteosarcoma (OS) para identificar genes críticos asociados con metástasis por medio del algoritmo de máquina de soporte vectorial (MVS). | Máquina de soporte vectorial. | El clasificador MVS mostró una alta precisión de predicción en todos los 4 conjuntos de datos, con una precisión de 100%, 100%, 92.6% y 100%, respectivamente. |
| Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls (Bheravan, 2018) | Propone un enfoque de aprendizaje automático eficaz para identificar un grupo de polimorfismos de un nucleótido (*SNP*) que interactúan, y que son los que más contribuyen al riesgo de cáncer de mama. | XGBoost, polimorfismo de un nucleótido, aprendizaje automático | Logran identificar aquellos *SNP* con mayor peso en la clasificación con XGBoost por lo que se puede precisar que tienen un potencial de riesgo. Con esto se puede lograr una alta precisión de predicción de riesgo de cáncer de mama para diferentes poblaciones. |

| Artículo | Objetivo | Técnicas o áreas | Conclusión |
|---|---|---|---|
| Pathway analysis using XGBoost classification in Biomedical Data (Dimitrakopoulos, 2018) | Detectar las vías más discriminatorias relacionadas a una enfermedad a través de XGBoost. | XGBoost, regresión lineal, máquina de soporte vectorial, K vecinos cercanos y Random Forest. | XGBoost supera a otros métodos de clasificación bien conocidos en datos biológicos. |
| An Empirical Study of Machine Learning Algorithms for Cancer Identification (Turki, 2018) | Comparar algoritmos de aprendizaje automático, cómo, AdaBoost, Deepboost, XGBoost y Máquina de soporte vectorial | AdaBoost, Deepboost, XGBoost y Máquina de soporte vectorial | Máquina de soporte vectorial, supera a los otros algoritmos, cuya base de datos fue de dos clases |

# CAPÍTULO 3. MARCO TEÓRICO

En este capítulo, se proporciona los conceptos básicos en los dominios que serán involucrados en el estudio, a saber: los aspectos biológicos del Osteosarcoma, *Big Data*, aprendizaje automático, clasificadores y métricas de evaluación.

## 3.1 Osteosarcoma

El Osteosarcoma es una neoplasia (tumor) maligna, que se caracteriza por la formación de tejido óseo y tejido osteoide con presencia de células tumorales. Esta patología tiene predominancia en adolescentes y adultos jóvenes en incidencia de entre los 15 a 19 años. Cifras de diagnóstico en Estados Unidos abarcan cerca de 900 casos al año, de los cuales entre el 15% y 20% presentan diagnóstico de metástasis(diseminación) (Rosales, 2014).

La incidencia reportada en México por el Instituto Mexicano del Seguro Social (IMSS) es de 9 acontecimientos por cada millón entre las edades de 10 y 14 años. A nivel mundial el Osteosarcoma representa el 15% de todas las biopsias realizadas y analizadas de tejidos de hueso. La presencia es de tres casos por cada millón de habitantes al año con lo cual representa el 0.2% de los tumores malignos (Rosales, 2014).

El esqueleto axial se ve afectado en raras ocasiones, y si lo está, es más frecuente en adultos. El fémur, la tibia y el húmero son los huesos con mayor frecuencia de daño (85%) mientras que menos del 1% se encuentra en manos o pies, mientras que, en los huesos largos, el Osteosarcoma origina la metástasis.

En la Figura 3.1 se puede observar la localización con más incidencia en la población por edad y por sitio anatómico del Osteosarcoma.

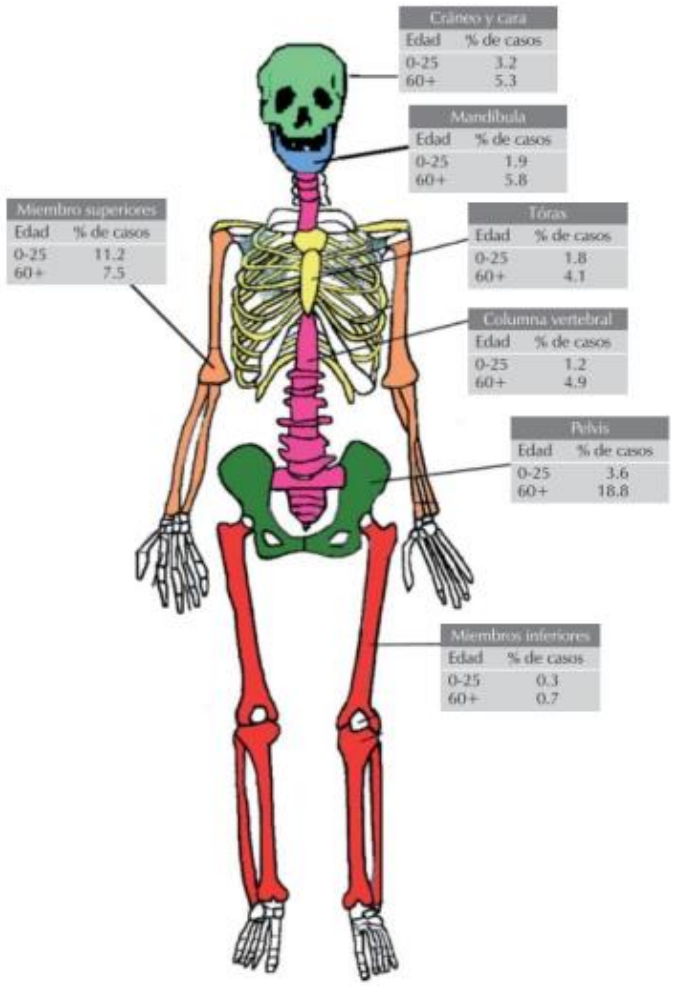



Figura 3.1. Localización del osteosarcoma por sitio anatómico (Rosales, 2014).

### 3.1.1 Etiología

El origen del Osteosarcoma aún es desconocido. Se han sugerido un origen viral al inducir sarcomas en animales por virus en extractos libres de células. El único agente ambiental conocido que causa el Osteosarcoma en humanos es la radiación ionizante, la cual está implicada en 2% de los Osteosarcomas, observándose un intervalo de 10 a 20 años entre la exposición y la formación del tumor (Rosales, 2014).

### 3.1.2 Moléculas asociadas con el Osteosarcoma

Se sabe que el Osteosarcoma tiene un complejo cariotipo y con alta frecuencia de amplificaciones genéticas. Hoy en día sólo la respuesta histológica y/o el momento de una cirugía, representaría una referencia en la predicción de resultados en pacientes con Osteosarcoma no metastásico.

El uso del análisis del método de hibridación genómica comparativa (CGH) y la tecnología de micro-arreglos de cDNA (es un ADN de cadena sencilla), sirven para estudiar y comprender las alteraciones que han sido identificadas por medio de secuencias de DNA y mRNA (ver Figura 3.2), esto sirve para identificar los genes relevantes para el pronto diagnóstico y de esta manera establecer las características de estos tumores. Dentro de los descubrimientos de una secuenciación, se encuentra la asociación del gen RB1 con el Osteosarcoma, localizado en el cromosoma 13q14, así como el gen TP53, y entre otros más (Rosales, 2014).

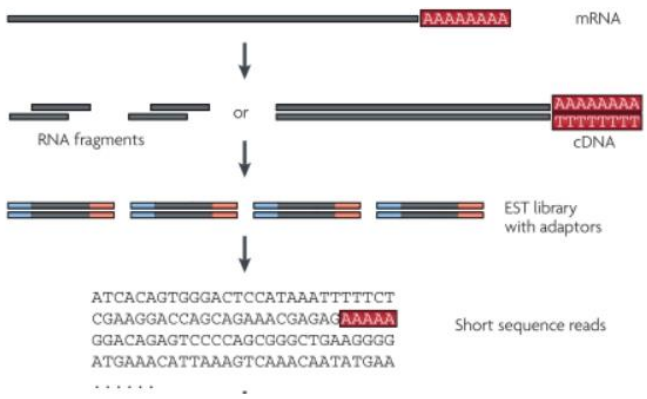



Figura 3.2 secuenciación de mRNA. El ARN largo se convierte en una biblioteca de fragmentos de cADN mediante la fragmentación del ARN o la fragmentación del ADN. Los adaptadores de secuenciación (barra azul) se añaden, posteriormente, a cada fragmento de cDNA y se obtiene una secuencia corta de cada cDNA utilizando la tecnología de secuenciación (Wang, 2009).

## 3.2 Big Data

Según (Gartner, 2018), *Big Data* es la información que se caracteriza por ser de gran volumen, generarse a gran velocidad y/o ser de gran variedad, lo cual demanda formas innovadoras y rentables de procesamiento de la información que permite una visión mejorada, la toma de decisiones y la automatización de procesos.

Es por lo que se necesitan técnicas para analizar, procesar estas grandes cantidades de datos (*Big Data*) para encontrar patrones o reglas que expliquen el comportamiento de dichos datos. Para eso se recurre al aprendizaje automático artificial.

En este caso, se tiene el banco de datos genómicos que fueron adquiridos a través de secuenciación de ADN en personas, la cuál es proporcionada por el Dr. Heriberto Manuel Rivera del Laboratorio de Biología de Sistemas y Medicina Traslacional (BSMT) de la Universidad Autónoma del Estado de Morelos (UAEM).

### 3.2.1 Preprocesamiento

El preprocesamiento de datos es una etapa esencial del proceso de descubrimiento de información o *KDD* (*Knowledge Discovery in Databases*, en inglés). Esta etapa se encarga de la preparación de los datos.

La preparación de datos está formada por una serie de técnicas que tienen el objetivo de inicializar correctamente los datos que servirán de entrada para los algoritmos de clasificación. Este tipo de técnicas pueden considerarse como de uso obligado, ya que sin ellas los algoritmos de extracción de conocimiento no podrían ejecutarse u ofrecerían resultados erróneos. En esta área se incluye la transformación de datos, integración, limpieza de ruido y de valores perdidos (ver Figura 3.3 (García, 2016).

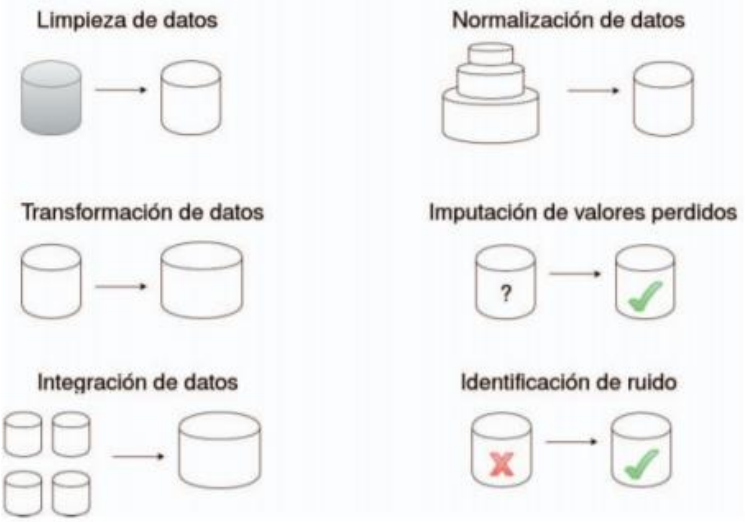



Figura 3.3. Preprocesamiento (García, 2016)

### 3.2.2 Análisis de componentes independientes

El análisis de componentes independientes (ACI) es una de las herramientas más populares en el procesamiento de datos que se utiliza para la separación de señales multivariadas en subcomponentes aditivos (Bouzalmat, 2014). Surge de la técnica conocida por su sigla BSS, o *Blind Sepparation Source*, que intenta obtener las fuentes independientes a partir de combinaciones de las mismas.

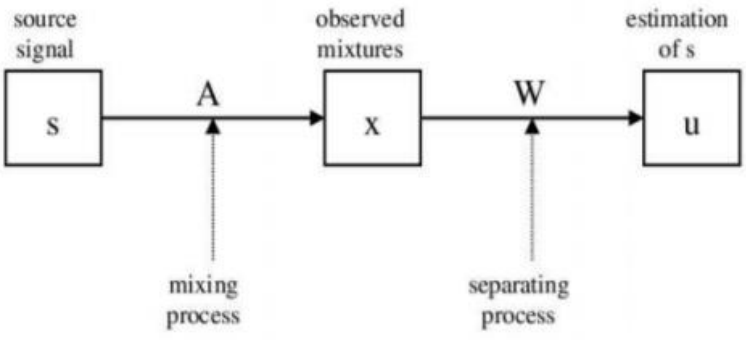



Figura 3.4. Análisis de componentes independientes (Draper, 2003)

En la Figura 3.4, sea $S$ la matriz de señales independientes y $X$ la matriz de observación.

Si $A$ es la matriz de combinación desconocida, el modelo de combinación se puede escribir como:

$$X = A.S \qquad \text{Ec. 3.1}$$

Asumiendo que las señales fuente son independientes unas de las otras y que la matriz $A$ es invertible, el algoritmo ACI tratará de encontrar la matriz de separación $W$, tal que (Spurek, 2018):

$$U = W*X = W*A*S \qquad \text{Ec. 3.2}$$

donde $U$: estimación de las componentes independientes.

### 3.2.3 Reconocimiento de patrones

En reconocimiento de patrones, un patrón se representa por un vector numérico de dimensión '*n*'. De esta forma, un patrón es un punto en un espacio n-dimensional (de características). El reconocimiento de patrones funciona en dos modos diferentes: entrenamiento y reconocimiento.

En el entrenamiento, se diseña el extractor o selector de características para representar los patrones de entrada y se entrena al clasificador con un conjunto de datos de entrenamiento de forma que el número de patrones mal identificados se minimice. En el modo de reconocimiento, el clasificador ya entrenado toma como entrada el vector de características de un objeto desconocido y lo asigna a una de las clases o categorías (Cano, 2014).

### 3.2.4 Aprendizaje automático artificial

- Los sistemas basados en inteligencia artificial (IA) se caracterizan porque contienen una representación del conocimiento, que les permite tomar decisiones de forma autónoma,
- En el aprendizaje automático, se adquiere el conocimiento a través de analizar datos y manipularlos, usando estrategias basadas en teorías matemáticas (Gil, 2016).

Con el aumento de la cantidad de datos (*Big Data*), el aprendizaje automático se ha convertido en una técnica clave para resolver problemas en áreas tales como:

- Finanzas computacionales: para la calificación crediticia y operaciones de mercados financieros.
- Procesamiento de imágenes y visión artificial: para el reconocimiento facial, la detección de movimiento y la detección de objetos
- Biología computacional: para la detección de tumores, el descubrimiento de fármacos y la secuenciación del ADN
- Producción de energía: para la previsión de la carga y el precio
- Automoción, sector aeroespacial y fabricación: para el mantenimiento predictivo
- Procesamiento del lenguaje natural: para aplicaciones de reconocimiento de voz

El aprendizaje automático se divide en dos ramas como se puede observar en la Figura 3.5 de acuerdo a (MathWorks 2018).

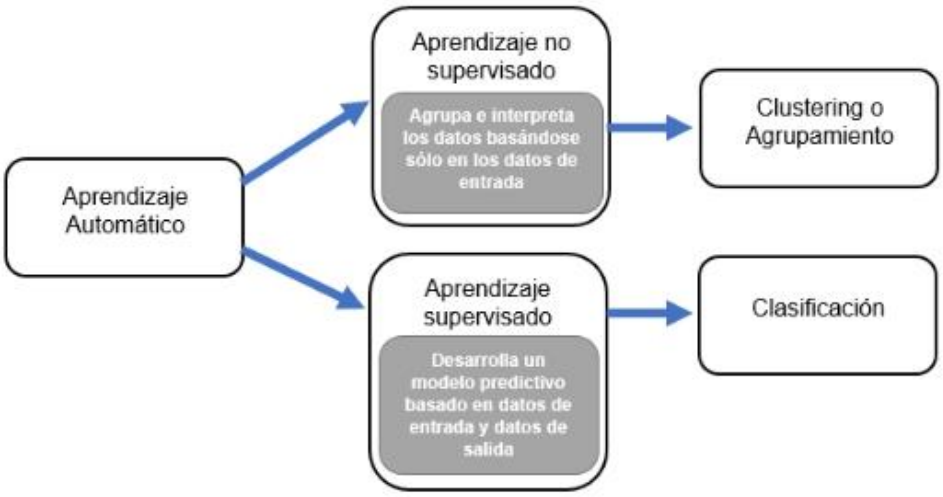



Figura 3.5. Ramas de aprendizaje automático (MathWorks, 2018).

### a) Aprendizaje supervisado

Toma un conjunto conocido de datos de entrada y respuestas conocidas para estos datos (salidas) y entrena un modelo con objeto de generar predicciones razonables como respuesta a datos nuevos.

El objetivo del aprendizaje supervisado es lograr que la computadora aprenda el modelo de clasificación. Implica en proporcionar información para entrenar al modelo a reaccionar con las clasificaciones que se le den, así aplicará lo que aprenda para dar respuestas a situaciones completamente nuevas que no ha visto antes (Mathworks, 2018).

Existen múltiples técnicas de aprendizaje supervisado, a continuación, se muestran algunas de las más comunes:

- Máquinas de vectores de soporte (SVM)
- Redes neuronales
- Clasificador Naïve Bayes
- Árboles de decisión
- Vecinos más cercanos (kNN)

### 3.2.5 Clasificación

El problema de clasificación ha sido ampliamente estudiado por investigadores en el área de bases de datos, estadísticas y aprendizaje automático. En el pasado, se propusieron muchos algoritmos de clasificación, como el análisis de discriminación lineal, los métodos de árbol de decisión, la red bayesiana, KNN, etc. En los últimos años se han aplicado estos algoritmos en el área de la bioinformática para conocer cómo los cambios en la expresión genética están relacionados con diferentes tipos de cáncer (Bhola, 2015).

#### 3.2.5.1 Random Forest

El concepto de Bosques Aleatorios (Random Forest) fue introducido por (Breiman, 2001). Random Forest es una combinación de árboles predictivos (clasificadores débiles); es decir, una modificación del *Bagging Bootstrap Aggregating*, el cual crea ejemplos separados del conjunto de entrenamiento y genera un clasificador para cada ejemplo. El resultado de estos clasificadores es combinado (como un promedio o por mayoría de voto) (Figura 3.6). La estrategia es que cada ejemplo del conjunto de entrenamiento es diferente, entonces, cada clasificador o árbol entrenado tiene un diferente enfoque y perspectiva del problema (Duval-Poo, 2012).

En el algoritmo, cada árbol depende de las variables aleatorias en un vector de la muestra de manera independiente y con la misma distribución en todos los demás árboles en el bosque. El uso de una selección aleatoria de características para dividir cada nodo produce tasas de error que se comparan favorablemente al algoritmo AdaBoost, pero son más robustos con respecto al ruido (clasificador fuerte) (Medina, 2017).

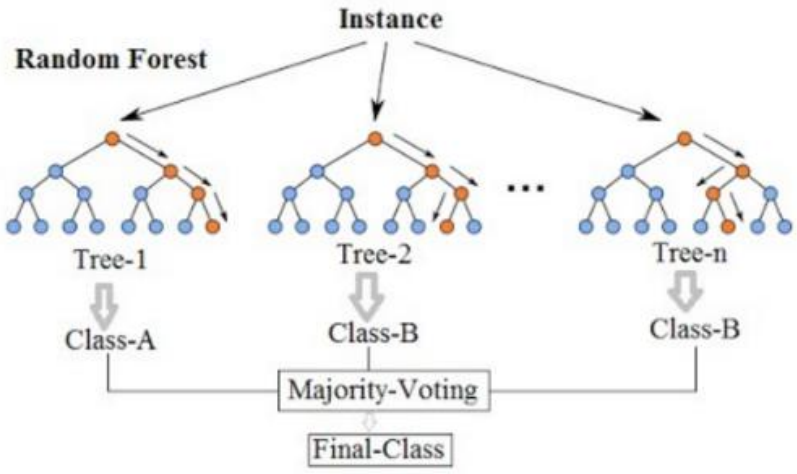



Figura 3.6. Funcionamiento de Random Forest (Duval-Poo, 2012).

## Ventajas de Random Forest

a) Trabaja sobre grandes bases de datos con múltiples variables.
b) Poder manejar cientos de variables entrantes sin excluir ninguna.
c) Dar estimados de qué variables son importantes en la clasificación.
d) Trabaja con variables categóricas.
e) Admite *missing values,* o valores perdidos en medio del procesamiento del algoritmo.

## Desventajas de Random Forest

a) Utiliza grandes cantidades de memoria RAM.
b) A diferencia de los árboles de decisión, la clasificación hecha por Random Forest es difícil de interpretar.
c) Necesita pocos parámetros para evitar un sobre-entrenamiento, como lo son; número de árboles (número de predictores a entrenar), profundidad (variables al azar) y el peso (ponderación o importancia de instancias).

### 3.2.5.2 XGBoost

XGBoost significa *eXtreme Gradient Boosting*. Es una implementación de árboles de decisión con *Gradient boosting* diseñada para minimizar la velocidad de ejecución y maximizar el rendimiento. XGBoost pertenece a una familia de algoritmos *Boosting* que convierten al aprendizaje débil en aprendizaje fuerte. Un aprendiz débil es uno que es ligeramente mejor que adivinar al azar. *Boosting* es un proceso secuencial; es decir, los árboles se cultivan utilizando la información de un árbol previamente crecido uno tras otro (Wyner, 2017). Este proceso aprende lentamente de los datos e intenta mejorar su predicción en iteraciones posteriores para reducir el error de clasificación (ver Figura 3.7) (Chen, 2016).

*Box*= clasificador

*D*=división o split

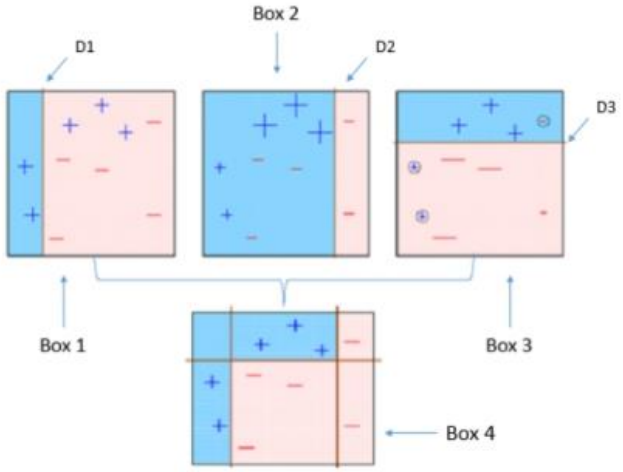



Figura 3.7. Funcionamiento XGBoost (Serrano, 2017)

**Ventajas de XGBoost**

a) Paralelización: debido a que el algoritmo central XGBoost es paralelizable, puede aprovechar la potencia de las computadoras de múltiples núcleos.

b) Constantemente supera a otros métodos de algoritmo: ha demostrado un mejor rendimiento en una variedad de conjuntos de datos de referencia de aprendizaje automático.
c) Amplia variedad de parámetros de ajuste: XGBoost tiene parámetros internos para validación cruzada, regularización, parámetros de árbol, API compatible con *scikit-learn*, reciente librería en R.
d) Trabaja sobre grandes bases de datos con múltiples variables, tanto categóricas como numéricas.
e) Admite *missing values*, o valores perdidos en medio del procesamiento del algoritmo.
f) Continua en el modelo existente: permite a los usuarios retomar o actualizar un proceso de entrenamiento que se había dejado a medias.
g) Da estimaciones de la importancia de las variables e instancias en la clasificación.

**Desventajas de XGBoost**

a) Utiliza grandes cantidades de memoria RAM.
b) En efecto, se recomienda equipos de más de 8GB para correr bases de datos extensas si se quiere trabajar con todas las variables.
c) Se debe ajustar correctamente los parámetros para minimizar el error en la precisión (mostrado en la sección de experimentación y resultados), como lo son; número de árboles, profundidad (número máximo de árboles), tasa de aprendizaje (grado en que a cada árbol se le permite corregir los errores de los árboles anteriores para una rápida convergencia en el entrenamiento) y el peso.

### 3.2.6 Métricas de evaluación

El objetivo más importante de cualquier clasificador es hacer predicciones precisas para casos nuevos. El resultado de la clasificación se resume en una matriz de confusión (también llamada Tabla de contingencia o matriz de clasificación). La matriz de confusión (ver Tabla 3.1) se puede usar para calcular varias métricas de rendimiento (Lodziensis, 2014).

Tabla 3.1. Matriz de confusión (Salem, 2017).

| Clase predicha | Clase verdadera | | ∑ |
|---|---|---|---|
| | Intancias positivas (P) | Intancias negativas (N) | |
| Intancias positivas (P) | **TP** *(True Positive* ) instancias positivas clasificadas como positivas | **FP** *(False Positive* ) instancias negativas clasificadas como positivas | **TP+FP** |
| Intancias negativas (N) | **FN** (*False Negative* ) instancias positivas clasificadas como negativas | **TN** (*True Negative* ) instancias negativas clasificadas como negativas | **FN+TN** |
| ∑ | **TP+FN** | **FP+TN** | **TP+FP+FN+TN** |

La Tabla 3.1 describe las diversas medidas de rendimiento de un clasificador. Instancias positivas (P) e instancias negativas (N), *True Positive* (TP): la cantidad de instancias positivas diagnosticadas correctamente. *True Negative* (TN): el número de instancias negativas diagnosticadas correctamente. *False Positive* (FP): el número de instancias negativas detectadas como positivas. *False Negative* (FN): el número de instancias positivas detectadas como negativas (Salem, 2017). A partir de la matriz de confusión se calculan otras métricas de evaluación como lo son:

#### 3.2.6.1 Precisión (Accuracy)

Se calcula la precisión de clasificación del algoritmo cuya formula se muestra en la Ecuación 3.3.

$$Precisión = \frac{Número de muestras correctamente clasificadas (VP + VN)}{Número total de muestras (VP + FN + FP + VN)} \quad \text{Ec. 3.3}$$

### 3.2.6.2 Sensibilidad (Sensitivity)

Es el parámetro de validación que define la probabilidad de que un individuo enfermo tenga un resultado positivo en la prueba. Es decir, detecta a los verdaderamente enfermos (ver Ecuación 3.4).

$$sensibilidad = \frac{VP}{VP + FN} \quad \text{Ec. 3.4}$$

### 3.2.6.3 Especificidad (Especificity)

Es el parámetro de validación que define la probabilidad de que un individuo sano tenga un resultado negativo en la prueba. Es decir, excluye a los verdaderamente sanos (ver Ecuación 3.5) (Infante, 2017).

$$Especificidad = \frac{VN}{VN + FP} \quad \text{Ec. 3.5}$$

### 3.2.6.4 F-measure

Es el promedio ponderado de Precisión y la relación entre las observaciones positivas pronosticadas correctamente y todas las observaciones en la clase real (ver Ecuación 3.6).

$$F-measure=\frac{2}{\frac{1}{\frac{VP}{VP+FN}}+\frac{1}{precisión}}$$ Ec. 3.6

#### 3.2.6.5 Área bajo la curva (AUC)

El AUC proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles (ver Ecuación 3.7).

$$AUC=0.5*(sensibilidad+especificidad)$$ Ec. 3.7

## 3.3 Discusión

El aprendizaje automático ofrece la posibilidad de comparar y relacionar la información genética con fines deductivos, así surgen respuestas que no parecen obvias a la vista de los resultados de los experimentos. En los últimos años, se aprecia el crecimiento de una corriente de investigación y desarrollo de nuevas técnicas para la extracción del conocimiento, la minería de datos o clasificación y la visualización, cuyo objetivo es acelerar los descubrimientos científicos. Estas nuevas técnicas son útiles para la investigación de distintas enfermedades y el diagnóstico clínico.

# CAPÍTULO 4. METODOLOGÍA DE SOLUCIÓN

En el presenta capitulo se presenta las librerías utilizadas para el diseño del sistema, preprocesamiento, análisis de componentes independientes, clasificadores y la paralelización.

## 4.1 Diseño del sistema

Para cumplir los objetivos propuestos, se propuso seguir la metodología que se muestra en la Figura 4.1 (ver Anexo A). Cada una de las etapas se describirán en detalle en las siguientes subsecciones de este capítulo.

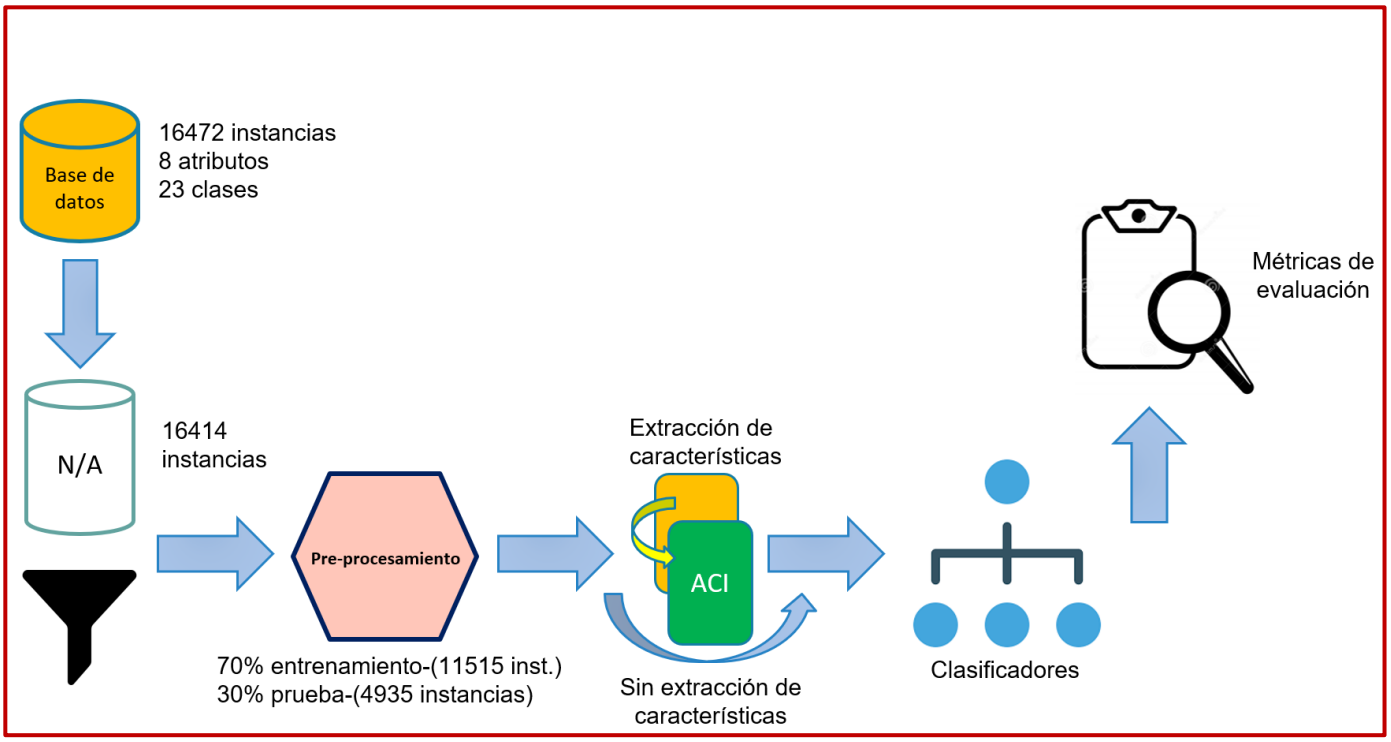



Figura 4.1. Metodología de solución

Para el desarrollo del proyecto se utilizó R versión 1.1.463; R es un entorno libre y lenguaje de programación con un enfoque al análisis estadístico. Se trata de uno de los lenguajes de programación más utilizados en investigación científica, siendo popular en el campo de la minería de datos, la investigación biomédica,

la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo y graficación (Reyes 2017). Como además contar con el equipo de cómputo con las siguientes especificaciones:

Computadora HP OMEN, Intel Core i7-8700k 3.7GHz con 8 núcleos, 24 GB de memoria RAM, Sistema operativo Windows 10, 1 Tb de disco duro, Tarjeta gráfica NVIDIA GeForce GTX 1050ti de 4 GB.

## 4.2 Banco de datos

La base de datos (Figura 4.3) fue proporcionada por el Laboratorio de Biología de Sistemas y Medicina Traslacional (BSMT) de la Universidad Autónoma del Estado de Morelos (UAEM) proveniente de la secuenciación de ADN, la cual consta de 8 atributos:

1. *Gene symbol:* característica de tipo cadena, ejemplo: *ABCA5, ABCB1 ABCC4.*
2. *Function/ phenotype:* característica de tipo cadena, ejemplo: *Actins are highly conserved proteins that are involved in various types of cell motility and are ubiquitously expressed in all eukaryotic cells.* Esta característica puede llegar a tener 4942 caracteres como se muestra en la Figura 11.
3. *Clinical significance:* esta característica es de tipo cualitativo, ejemplo: *Benign drug response, Pathogenic.*
4. *Chromosome*; esta característica es de tipo numérico, ejemplo: 17, 7, 13
5. *dbSNP*: característica de tipo nominal, ejemplo: 199753304, 2032582, 1045642, 1128501, 1751034.
6. *Association:*, esta característica es de tipo cadena, ejemplo: *Familial adenomatous polyposis 1; Hereditary cancer-predisposing síndrome.* de igual forma que la segunda característica, esta puede contener más de 1000 caracteres.
7. *Ancestral allele*: característica de tipo cualitativo, ejemplo: C, A, T, G.
8. *Alternate allele:* característica de tipo cualitativo, ejemplo: A, C, T, G.

Estas variables son de tipo cadena, cualitativos y cuantitativos, además cuenta con 16,472 instancias y 23 clases (PMID). El número de instancias por cada clase se describe en la Tabla 4.1 y Figura 4.3.

| GeneSymbol | Function/Phenotype | ClinicalSignificance | Chromosome | dbSNP | Association | Ancestral Allele | Alternate Allele | Clase |
|---|---|---|---|---|---|---|---|---|
| ABCA5 | May play a role in the processing of autolysosomes. | Pathogenic/Likely pathogenic | 17 | 199753304 | Gingival fibromatosis with hypertrichosis;Inborn genet | C | G | 1 |
| ABCB1 | Energy-dependent efflux pump responsible for decreased drug accumu | Benign | 7 | 2032582 | Ovarian Neoplasms;not specified | A | T | 1 |
| ABCB1 | Energy-dependent efflux pump responsible for decreased drug accumu | drug response | 7 | 1045642 | Non-small cell lung cancer;digoxin response - Other;fe | A | G | 1 |
| ABCB1 | Energy-dependent efflux pump responsible for decreased drug accumu | Pathogenic | 7 | 1128501 | Colchicine resistance | C | A | 2 |
| ABCC4 | May be an organic anion pump relevant to cellular detoxification. | drug response | 13 | 1751034 | tenofovir response - | C | T | 3 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Benign | 10 | 3816245 | Familial aortopathy;not specified | C | T | 3 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely benign | 10 | 372824072 | Aortic aneurysm, familial thoracic 6 | T | C | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely benign | 10 | 750005327 | Aortic aneurysm, familial thoracic 6 | C | A | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely benign | 10 | 1060503847 | Aortic aneurysm, familial thoracic 6 | G | A | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely benign | 10 | 200213764 | Aortic aneurysm, familial thoracic 6;Thoracic aortic an | G | A | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely benign | 10 | 367977687 | Moyamoya disease;Multisystemic smooth muscle dysf | C | T | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely benign | 10 | 111265233 | Moyamoya disease;Multisystemic smooth muscle dysf | C | T | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely benign | 10 | 141538225 | Thoracic aortic aneurysm and aortic dissection | G | A | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely pathogenic | 10 | 746972765 | Thoracic aortic aneurysm and aortic dissection | C | T | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Likely pathogenic | 10 | 727502878 | Thoracic aortic aneurysm and aortic dissection | C | G | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic | 10 | 397515325 | Aortic aneurysm, familial thoracic 6 | T | C | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic | 10 | 387906592 | Aortic aneurysm, familial thoracic 6;Connective tissue | C | T | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic | 10 | 121434527 | Aortic aneurysm, familial thoracic 6;Moyamoya diseas | C | T | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic | 10 | 112602953 | Aortic aneurysm, familial thoracic 6;not provided | C | T | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic | 10 | 121434526 | Aortic aneurysm, familial thoracic 6;Thoracic aortic an | G | A | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic | 10 | 886038978 | Thoracic aortic aneurysm and aortic dissection | G | A | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic/Likely pathogenic | 10 | 121434528 | Aortic aneurysm, familial thoracic 6;Moyamoya diseas | G | A | 4 |
| ACTA2 | Actins are highly conserved proteins that are involved in various types | Pathogenic/Likely pathogenic | 10 | 112901682 | Aortic aneurysm, familial thoracic 6;Thoracic aortic an | G | A | 4 |

Figura 4.2. Banco de datos

Se dio a la tarea de conocer el número de ejemplos por clase (ver Tabla 4.1 y Figura 4.3). Debido a que se quiere clasificar el osteosarcoma como además su tipo de metástasis (diseminación), se cuentan con 23 clases.

Tabla 4.1. Número de objetos por cada clase.

| objetos por clase | |
|---|---|
| Clase 1 | 283 |
| Clase 2 | 2179 |
| Clase 3 | 415 |
| Clase 4 | 892 |
| Clase 5 | 575 |
| Clase 6 | 514 |
| Clase 7 | 766 |
| Clase 8 | 928 |
| Clase 9 | 192 |
| Clase 10 | 333 |
| Clase 11 | 787 |
| Clase 12 | 1022 |
| Clase 13 | 286 |
| Clase 14 | 267 |
| Clase 15 | 2442 |
| Clase 16 | 1571 |
| Clase 17 | 349 |
| Clase 18 | 311 |
| Clase 19 | 363 |
| Clase 20 | 238 |
| Clase 21 | 451 |
| Clase 22 | 661 |
| Clase 23 | 647 |

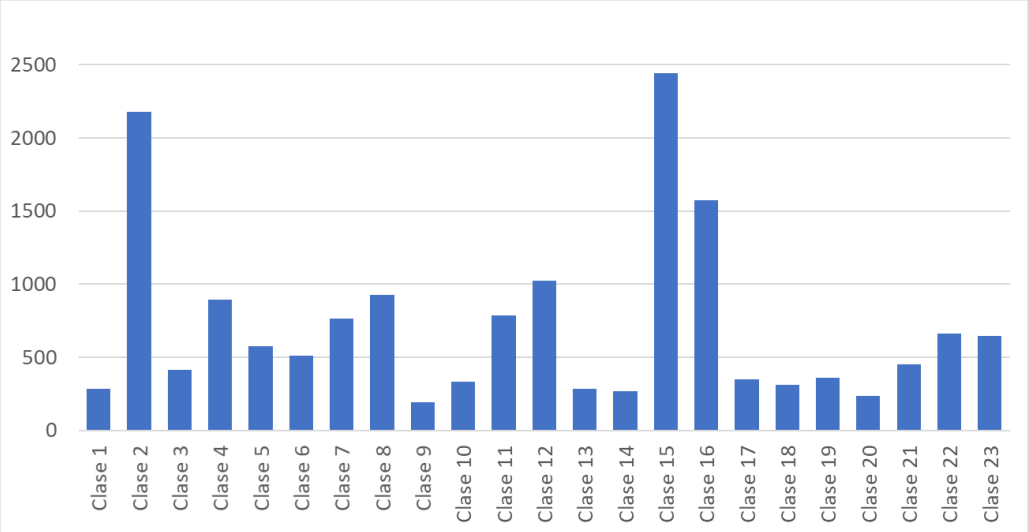



Figura 4.3. Número de objetos por clase

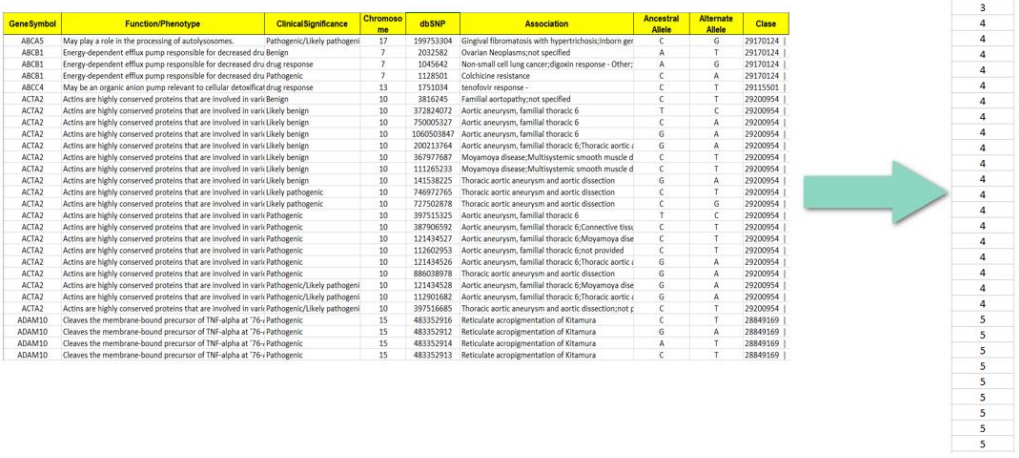
## 4.3 Preprocesamiento de los datos

Se analizó el banco de datos y se encontraron con datos muy grandes como se observa en la Figura 4.4. La cual muestra que la variable se integra de 4,942 caracteres, esto afecta el tiempo de entrenamiento y también imposibilita el utilizar algoritmos de clasificación numéricos y además se observó que hay celdas sin información (NA).


Figura 4.4. Tamaño del dato

A partir del análisis de los datos, se realizaron las siguientes adecuaciones para su limpieza y mejor manejo:

- Se encontró que algunas variables presentaban valores de N/A (ausencia de valor). Estos valores se eliminaron ya que no aportan información. Esta actividad se llevó a cabo con la librería *dplyr* (ver glosario de librerías de R). Dicha librería tiene la función *na.omit( )* donde solamente se especifica cual banco de datos se desea limpiar. Que resultó en una reducción del tamaño de la base de datos pasando de 16,472 a 16,414.
- Posteriormente se pasó a la transformación del atributo clase (23 clases). Como se mencionó, inicialmente esta variable era de tipo cadena con un tamaño de hasta 4,942 caracteres; esta información se transformó a tipo nominal con el fin de reducir el tiempo de entrenamiento por medio de una codificación, esta actividad se llevó a cabo con la librería *tidyr* (ver glosario de librerías de R), el resultado se observa en la Figura 4.5.



| GeneSymbol | Function/Phenotype | ClinicalSignificance | Chromosome | dbSNP | Association | Ancestral Allele | Alternate Allele | Clase |
|---|---|---|---|---|---|---|---|---|
| ABCA5 | May play a role in the processing of autolysosomes. | Pathogenic/Likely pathogeni | 17 | 199753304 | Gingival fibromatosis with hypertrichosis;Inborn ger | C | G | 29170124 |
| ABCB1 | Energy-dependent efflux pump responsible for decreased dru | Benign | 7 | 2032582 | Ovarian Neoplasms;not specified | A | T | 29170124 |
| ABCB1 | Energy-dependent efflux pump responsible for decreased dru | drug response | 7 | 1045642 | Non-small cell lung cancer;digoxin response - Other; | A | G | 29170124 |
| ABCB1 | Energy-dependent efflux pump responsible for decreased dru | Pathogenic | 7 | 1128501 | Colchicine resistance | C | A | 29170124 |
| ABCC4 | May be an organic anion pump relevant to cellular detoxificat | drug response | 13 | 1751034 | tenofovir response - | C | T | 29115501 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Benign | 10 | 3816245 | Familial aortopathy;not specified | C | T | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely benign | 10 | 372824072 | Aortic aneurysm, familial thoracic 6 | T | C | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely benign | 10 | 750005327 | Aortic aneurysm, familial thoracic 6 | C | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely benign | 10 | 1060503847 | Aortic aneurysm, familial thoracic 6 | G | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely benign | 10 | 200213764 | Aortic aneurysm, familial thoracic 6;Thoracic aortic a | G | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely benign | 10 | 367977687 | Moyamoya disease;Multisystemic smooth muscle d | C | T | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely benign | 10 | 111265233 | Moyamoya disease;Multisystemic smooth muscle d | C | T | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely benign | 10 | 141538225 | Thoracic aortic aneurysm and aortic dissection | G | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely pathogenic | 10 | 746972765 | Thoracic aortic aneurysm and aortic dissection | C | T | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Likely pathogenic | 10 | 727502878 | Thoracic aortic aneurysm and aortic dissection | C | G | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic | 10 | 397515325 | Aortic aneurysm, familial thoracic 6 | T | C | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic | 10 | 387906592 | Aortic aneurysm, familial thoracic 6;Connective tissu | C | T | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic | 10 | 121434527 | Aortic aneurysm, familial thoracic 6;Moyamoya dise | C | T | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic | 10 | 112602953 | Aortic aneurysm, familial thoracic 6;not provided | C | T | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic | 10 | 121434526 | Aortic aneurysm, familial thoracic 6;Thoracic aortic a | G | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic | 10 | 886038978 | Thoracic aortic aneurysm and aortic dissection | G | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic/Likely pathogeni | 10 | 121434528 | Aortic aneurysm, familial thoracic 6;Moyamoya dise | G | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic/Likely pathogeni | 10 | 112901682 | Aortic aneurysm, familial thoracic 6;Thoracic aortic a | G | A | 29200954 |
| ACTA2 | Actins are highly conserved proteins that are involved in vari | Pathogenic/Likely pathogeni | 10 | 397516685 | Thoracic aortic aneurysm and aortic dissection;not p | C | T | 29200954 |
| ADAM10 | Cleaves the membrane-bound precursor of TNF-alpha at "76- | Pathogenic | 15 | 483352916 | Reticulate acropigmentation of Kitamura | C | T | 28849169 |
| ADAM10 | Cleaves the membrane-bound precursor of TNF-alpha at "76- | Pathogenic | 15 | 483352912 | Reticulate acropigmentation of Kitamura | G | A | 28849169 |
| ADAM10 | Cleaves the membrane-bound precursor of TNF-alpha at "76- | Pathogenic | 15 | 483352914 | Reticulate acropigmentation of Kitamura | A | T | 28849169 |
| ADAM10 | Cleaves the membrane-bound precursor of TNF-alpha at "76- | Pathogenic | 15 | 483352913 | Reticulate acropigmentation of Kitamura | C | T | 28849169 |

| (Clase) |
|---|
| 1 |
| 1 |
| 1 |
| 2 |
| 3 |
| 3 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 4 |
| 5 |
| 5 |
| 5 |
| 5 |
| 5 |
| 5 |
| 5 |
| 5 |
| 5 |

Figura 4.5. Codificación de la clase

## 4.4 Extracción de características

Debido a que el vector de características se integra de pocos atributos, se decidió utilizar el análisis de componentes independientes para tener el mismo número de variables, pero eliminando el posible ruido presente en los datos. En esta etapa se utilizó las librerías "*caret", "FSelector", "fastica", "ggplot2*" de R (ver glosario de librerías de R).

### 4.4.1 Análisis de componentes independientes

El análisis de componentes independientes (ACI) es una de las herramientas más populares en el procesamiento de datos que se utiliza para la separación de señales multivariadas en subcomponentes aditivos (Bouzalmat, 2014). Surge de la técnica conocida por su sigla *BSS*, o *Blind Separation Source*, que intenta obtener las fuentes independientes a partir de combinaciones de las mismas. Por lo que no hay una selección de variables para obtener las más representativas sino hay una independencia entre las variables para disminuir el ruido por lo cual se obtiene el mismo número de variables que se tenían originalmente.

### 4.4.2 Aplicación del Análisis de componentes independientes

Es recomendable que para la implementación de la técnica de ACI se visualice antes la distribución de los datos (ver Figura 4.6) para observar la diferencia en la dispersión de los datos, y comparar posteriormente con los resultados de la aplicación de ACI.

La técnica de análisis de componentes independientes no arroja un valor para las variables más representativas, la técnica de ACI sólo reduce o elimina el ruido y da como resultado, variables más independientes.

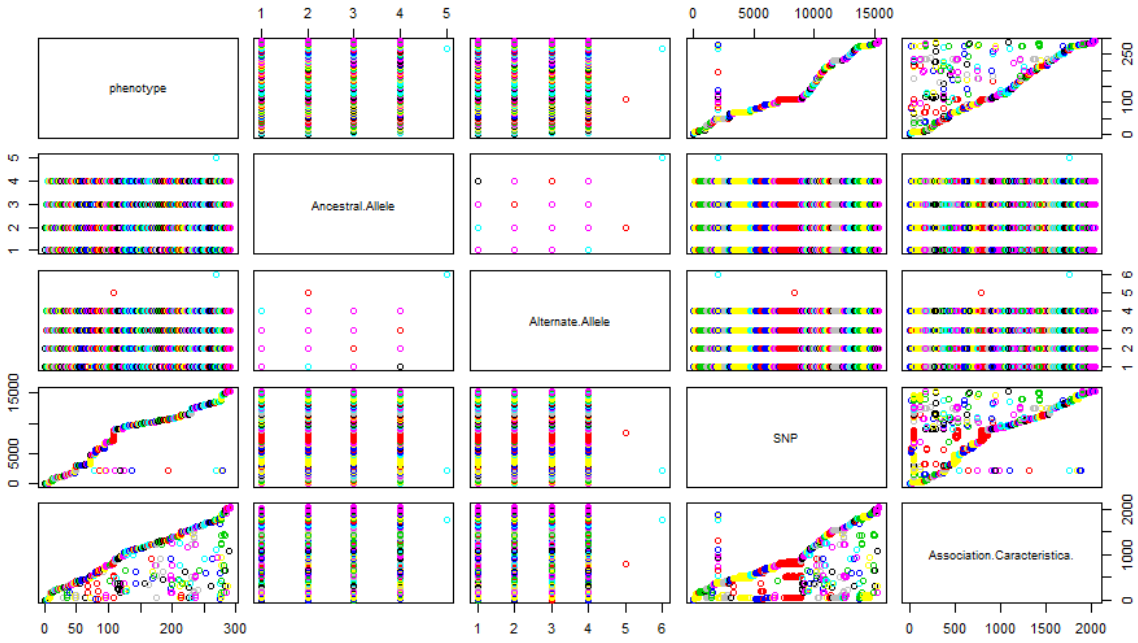



Figura 4.6. Distribución de los datos

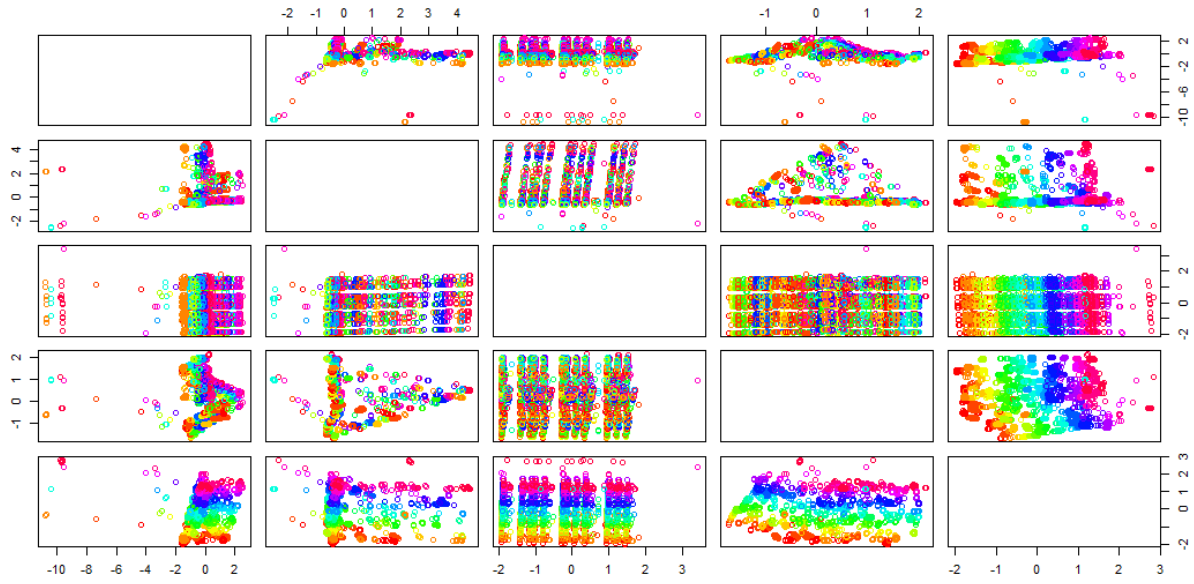


Figura 4.7. Análisis de componentes independientes

En la Figura 4.7 se puede observar la aplicación de ACI en todo el banco de datos, la cual muestra la independencia de las variables de la base de datos de osteosarcoma, en comparación de la Figura 4.6 que muestra una dependencia en los datos.

Un ejemplo específico es la Figura 4.8, la cual muestra las variables: *SNP* y *Association*. En la cual se observa antes de aplicar ACI y después de la aplicación del algoritmo ACI.

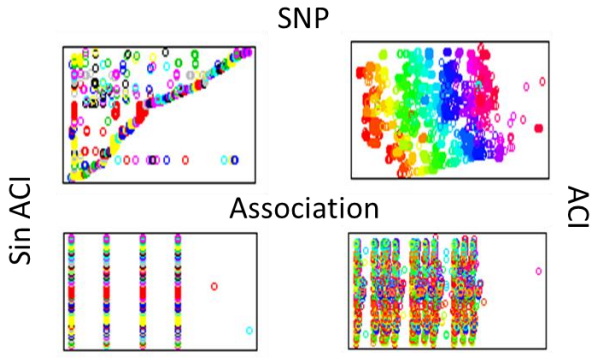



Figura 4.8. Comparación de dos variables

En estas dos variables hay una dependencia antes de aplicar ACI, una vez aplicado ACI, muestra una separación o independencia de las variables, esto permite una rápida convergencia en el tiempo de entrenamiento, como explica y menciona en el capítulo 6.

## 4.5 Clasificación

En esta fase se evaluaron dos clasificadores y se aplicaron en el banco de datos en cuestión. Los algoritmos de Random Forest y XGBoost son los utilizados para este trabajo, debido a las características de procesar múltiples variables de tipo cadena, cualitativos y cuantitativos. A diferencia de otros clasificadores como, Naive Bayes, máquina de soporte vectorial, K vecinos cercanos, Redes neuronales, regresión logística, que son algoritmos de clasificación más utilizados, no son capaces de procesar la naturaleza de las variables ya mencionadas.

Se utilizó las librerías *"caret", "e1071", "doParallel", "iterators"* en R (ver glosario de librerías de R).

## 4.6 Paralelización de los clasificadores

Una de las ventajas que comparten Random Forest y XGBoost es la capacidad de paralelizar el procesamiento de cómputo. Debido a esta característica se realizó y se modificaron los paquetes de R-studio: caret, dplyr, iterators, doParallel (ver glosario de librerías de R) con sus respectivas librerías con la intención de detectar y establecer conjuntos de núcleos del *hardware* a utilizar en este proceso. Adicionalmente, esta paralelización da la pauta para conocer el tiempo de convergencia del clasificador (Anexo B).

## 4.7 Métricas de evaluación

Para conocer el desempeño de un clasificador se utilizan diferentes métricas de evaluación. Para calcular cada métrica es necesario calcular la matriz de confusión (Díaz, 2015). Una matriz de confusión contiene información sobre lo real y lo predicho por un sistema de clasificación.

El desempeño de tales sistemas es comúnmente evaluado utilizando los datos en la matriz. La Tabla 4.2 muestra la matriz de confusión para un clasificador de dos clases (Visa, 2011). En esta etapa se utilizó la librería *caret*. (ver Glosario de librerías de R)

Tabla 4.2. Matriz de confusión (Díaz, 2015)

| *Matriz de Confusión* | | *Clase verdadera* | |
|---|---|---|---|
| | | *Pos* | *Neg* |
| *Clase Predicha* | *pos* | *VP* | *FP* |
| | *neg* | *FN* | *VN* |
| *Total columna* | | *P* | *N* |

En la Tabla 4.2 las siglas VP y VN representan los elementos bien clasificados de la clase positiva y negativa respectivamente y FP y FN identifican los elementos negativos y positivos mal clasificados. Basados en estas medidas, se calcula el error, la exactitud, sensibilidad, la tasa de FP (falsos positivos), la especificidad los falsos negativos (Navin, 2016).

La evaluación de la respuesta de los clasificadores permitió:

- Implementar las métricas de evaluación (precisión, especificidad, sensibilidad, F-measure y AUC) para determinar el rendimiento del sistema.
- Comparar los resultados (métricas) con los algoritmos de aprendizaje automático seleccionados.

Se utilizó las librerías "*caret", "ggplot*" en R (ver glosario de librerías de R).

## 4.8 Discusión

Para poder aplicar dos algoritmos de clasificación para datos genómicos, se hizo de la siguiente manera.

Primeramente, conocer los datos, tamaño y magnitud y naturaleza, conociendo esta información se eligieron los clasificadores acordes a las necesidades de poder procesar numerosas cantidades de instancias (16,472) y características tipo cuantitativas, cualitativas y tipo cadena. Por ello los algoritmos de clasificación seleccionados fueron Random Forest y XGBoost.

# CAPÍTULO 5. PRUEBAS Y RESULTADOS

En este capítulo, se describen a detalle las actividades de prueba y los resultados obtenidos al utilizar los algoritmos de clasificación y las métricas de evaluación explicadas en el capítulo anterior.

Se comenzaron con las pruebas de variar diferentes parámetros para encontrar los óptimos, los experimentos son, número de árboles, profundidad, tasa de aprendizaje, parámetros sobre ajustados y validación cruzada. Para ello se particionó de manera aleatoria con el 70% de los datos para entrenamiento (11,515) y el 30% para la etapa de validación (4,935). Esta actividad se llevó a cabo con la librería *caret* (ver glosario de librerías de R).

## 5.1 Experimentación 1: número árboles

Posterior a la aplicación de ACI, se comenzó con la ejecución de los clasificadores, con él 70% del *dataset* para entrenar a los clasificadores y el 30% para su validación y evaluación de los clasificadores.

Random Forest y XGBoost son clasificadores basados en árboles, por lo que cuentan con variables o parámetros a modificar para el entrenamiento tales como:

- Número de árboles.
- Profundidad.
- Tasa de aprendizaje (en el único caso de XGBoost).
- Pesos.

El primer parámetro que se llevó a cabo en la experimentación fue “número de árboles”, y los demás parámetros no cambiaron. Donde se realizaron y evaluaron cinco pruebas como se ve en la Tabla 5.1, cuyos parámetros fueron:

Prueba 1: número de árboles=100, Profundidad=3, Tasa aprendizaje= 0.2, Peso=1.

Prueba 2: número de árboles=200, Profundidad=3, Tasa aprendizaje= 0.2, Peso=1.

Prueba 3: número de árboles=500, Profundidad=3, Tasa aprendizaje= 0.2, Peso=1.

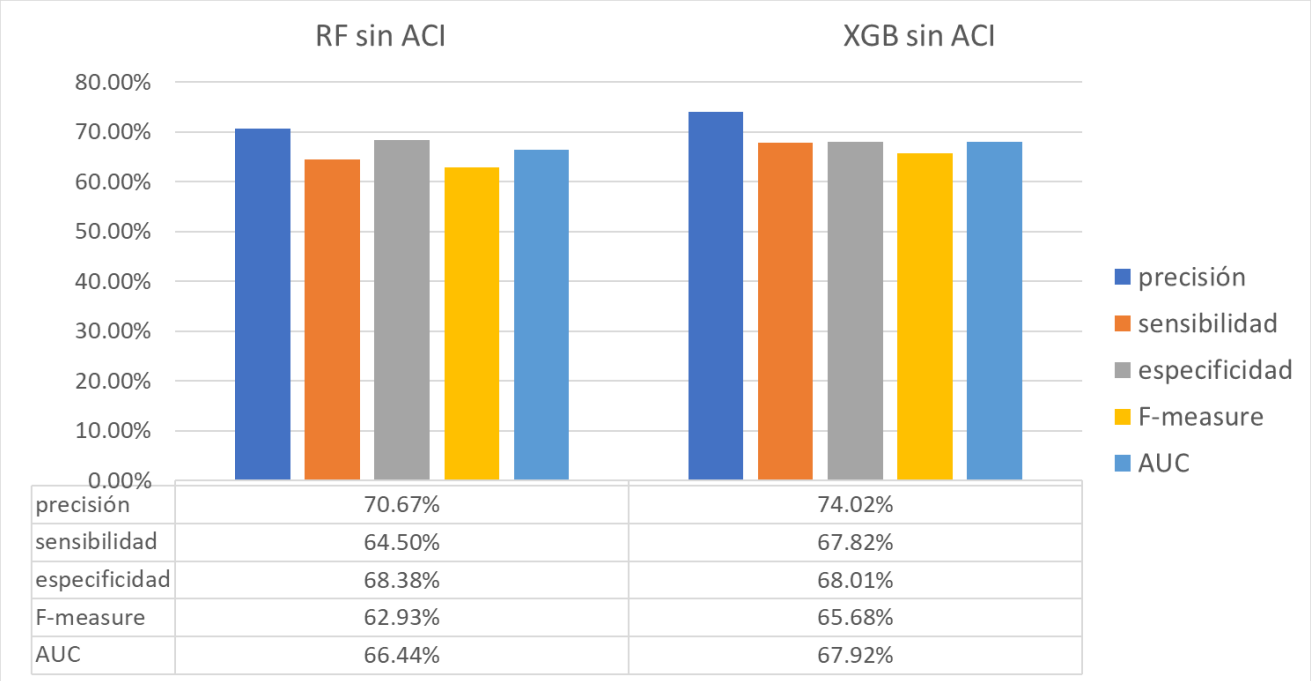
Prueba 4: número de árboles=1000, Profundidad=3, Tasa aprendizaje= 0.2, Peso=1.

Prueba 5: número de árboles=2000, Profundidad=3, Tasa aprendizaje= 0.2, Peso=1.

Tabla 5.1. Resultados del experimento 1

| Algoritmo | Técnica | Precisión | Sensibilidad | Especificidad | F-measure | AUC | No. Arboles | Profundidad | Tasa de aprendizaje | Peso | No. Prueba |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Random Forest | sin ACI | 69.09% | 56.02% | 64.12% | 61.69% | 60.07% | 100 | 3 | 0.2 | 1 | 1 |
| | ACI | 68.17% | 62.20% | 66.13% | 58.96% | 64.17% | | | | | |
| XGBoost | sin ACI | 71.01% | 61.06% | 63.40% | 62.33% | 62.23% | | | | | |
| | ACI | 70.43% | 63.10% | 63.75% | 62.07% | 63.43% | | | | | |
| Random Forest | sin ACI | 69.36% | 63.10% | 66.20% | 58.59% | 64.65% | 200 | 3 | 0.2 | 1 | 2 |
| | ACI | 69.80% | 64.59% | 67.35% | 58.43% | 65.97% | | | | | |
| XGBoost | sin ACI | 72.60% | 62.48% | 68.85% | 57.15% | 65.67% | | | | | |
| | ACI | 73.20% | 65.35% | 67.41% | 65.23% | 66.38% | | | | | |
| Random Forest | sin ACI | 69.36% | 63.10% | 66.20% | 64.29% | 64.65% | 500 | 3 | 0.2 | 1 | 3 |
| | ACI | 69.80% | 64.59% | 67.35% | 66.99% | 65.97% | | | | | |
| XGBoost | sin ACI | 72.60% | 62.48% | 67.85% | 64.19% | 65.17% | | | | | |
| | ACI | 73.20% | 65.35% | 67.41% | 57.00% | 66.38% | | | | | |
| Random Forest | sin ACI | 72.10% | 69.18% | 72.40% | 65.64% | 70.79% | 1000 | 3 | 0.2 | 1 | 4 |
| | ACI | 74.56% | 61.25% | 71.43% | 60.97% | 66.34% | | | | | |
| XGBoost | sin ACI | 74.00% | 73.80% | 68.13% | 73.09% | 70.97% | | | | | |
| | ACI | 75.10% | 70.47% | 73.20% | 68.02% | 71.84% | | | | | |
| Random Forest | sin ACI | 73.43% | 71.10% | 73.00% | 64.46% | 72.05% | 2000 | 3 | 0.2 | 1 | 5 |
| | ACI | 76.40% | 65.37% | 74.16% | 61.16% | 69.77% | | | | | |
| XGBoost | sin ACI | 79.87% | 79.30% | 71.80% | 71.61% | 75.55% | | | | | |
| | ACI | 80.98% | 68.90% | 79.10% | 59.23% | 74.00% | | | | | |

En la Tabla 5.1 se evaluó el experimento 1 con cinco métricas (precisión, sensibilidad, especificidad, F-measure y AUC), debido a los múltiples resultados se optó por realizar un promedio y se compararon (Figura 5.1).

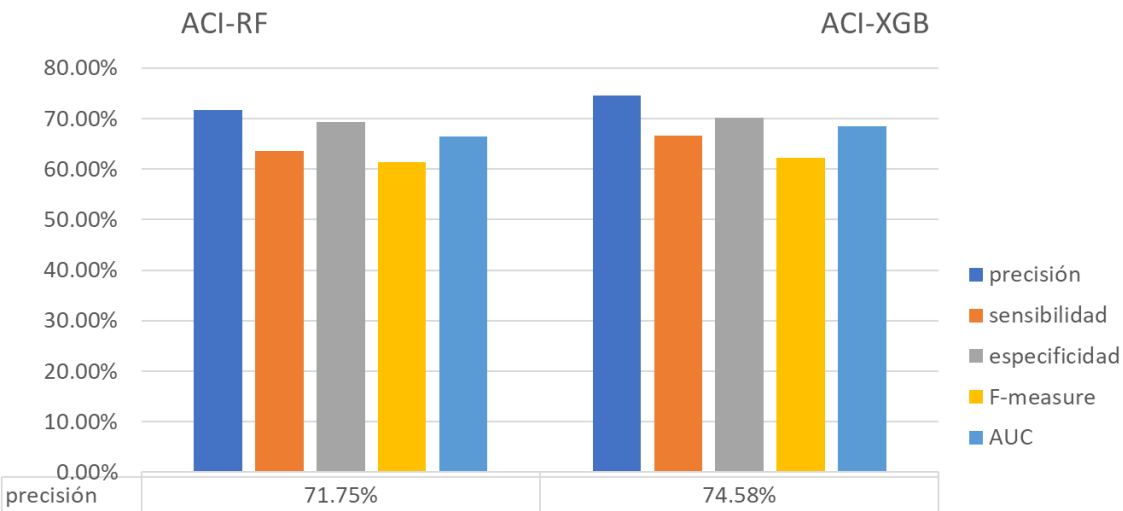



| | | |
|---|---|---|
| precisión | 70.67% | 74.02% |
| sensibilidad | 64.50% | 67.82% |
| especificidad | 68.38% | 68.01% |
| F-measure | 62.93% | 65.68% |
| AUC | 66.44% | 67.92% |

Figura 5.1. Promedio de experimento 1 sin ACI.

El promedio de la precisión de Random Forest (RF) sin ACI fue de 70.67% donde el menor valor fue de 69.09% (100 árboles) y su máximo valor fue de 73.43% (2000 árboles), la sensibilidad promedio RF sin ACI fue de 64.50%, un 68.38% en especificidad, un 62.93% de F-measure y un 66.44 de AUC.

En el caso de XGBoost (XGB) sin ACI, tuvo un promedio de precisión de 74.02% donde su valor mínimo fue de 71.01% (100 árboles) y su máximo de 79.87% (2000 árboles), un 64.82% en sensibilidad promedio, 65.81% en especificidad promedio, 65.68% de F-measure y un 67.92% de AUC.Después se realizó una comparación de los promedios de las métricas con la técnica de ACI (ver Figura 5.2).




| | ACI-RF | ACI-XGB |
|---|---|---|
| precisión | 71.75% | 74.58% |
| sensibilidad | 63.60% | 66.63% |
| especificidad | 69.28% | 70.17% |
| F-measure | 61.30% | 62.31% |
| AUC | 66.44% | 68.40% |

Figura 5.2. Promedio de experimento 1 con ACI.

El promedio de la precisión con RF con ACI fue de 71.75% donde el menor valor fue de 68.17% (100 árboles) y su máximo valor fue de 76.40% (2000 árboles), la sensibilidad fue de 63.60%, especificidad 69.28%, F-measure 61.30%, y un AUC de 66.44%.

XGBoost (XGB) con ACI tuvo un promedio de la precisión de 74.58% pasando de una precisión mínima de 70.43% (con 100 árboles) a un 80.98% (2000 árboles), una sensibilidad de 66.63%, especificidad de 70.17%, F-measure de 62.31% y un AUC de 68.40%

El objetivo de este experimento fue evaluar el aumento de los valores de las métricas respecto al parámetro de número de árboles, donde XGBoost con ACI obtiene los valores más altos con un total de 2000 árboles.

## 5.2 Experimentación 2: profundidad

En este segundo experimento, el parámetro que se puso a prueba fue el de "profundidad", donde de igual forma, se evaluaron con las cinco métricas y se compararon.

Los parámetros utilizados para ambos clasificadores fueron:

- Parámetros: número de árboles=1000, Profundidad=5, Tasa aprendizaje= 0.2, Peso=1
- Parámetros: número de árboles=1000, Profundidad=10, Tasa aprendizaje= 0.2, Peso=1
- Parámetros: número de árboles=1000, Profundidad=15, Tasa aprendizaje= 0.2, Peso=1
- Parámetros: número de árboles=2000, Profundidad=10, Tasa aprendizaje= 0.2, Peso=1
- Parámetros: número de árboles=2000, Profundidad=13, Tasa aprendizaje= 0.2, Peso=1
- Parámetros: número de árboles=2000, Profundidad=11, Tasa aprendizaje= 0.2, Peso=1

Se inicia con 1000 árboles ya que fue donde comienza a dar mejores resultados de acuerdo con el experimento 1, a partir de ahí, se da comienzo a variar la profundidad.

Se alcanzó un valor máximo de 13 pero al detectar que no entrega una mayor precisión, se opta por disminuir, hasta alcanzar el óptimo.

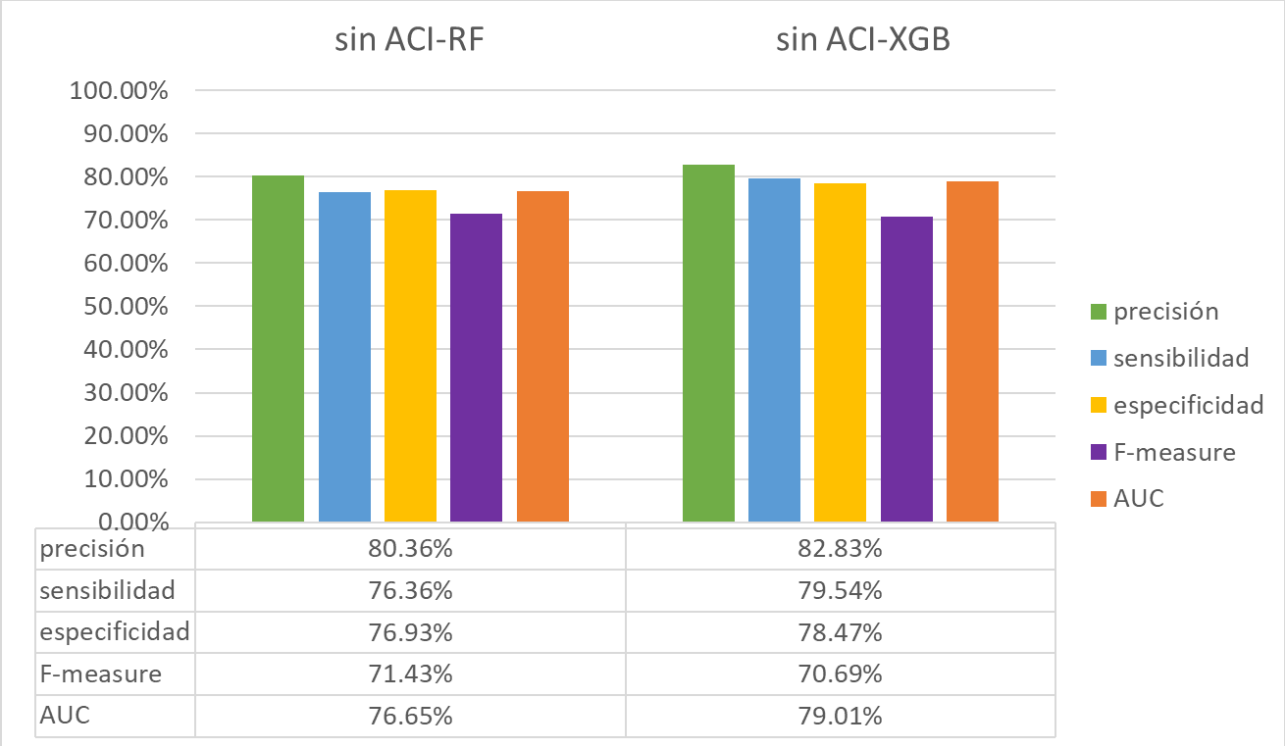
Una vez conocida la profundidad, se empezó a variar el número de árboles, cuyos resultados se pueden ver en la Tabla 5.2.

Tabla 5.2. Resultados del experimento 2

| Algoritmo | Técnica | Precisión | Sensibilidad | Especificidad | F-measure | AUC | No. Arboles | Profundidad | Tasa de aprendizaje | Peso | No. Prueba |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Random Forest | sin ACI | 74.20% | 68.16% | 68.09% | 65.42% | 68.13% | 1000 | 5 | 0.2 | 1 | 6 |
| | ACI | 76.07% | 68.00% | 69.00% | 63.39% | 68.50% | | | | | |
| XGBoost | sin ACI | 74.24% | 69.47% | 68.45% | 62.45% | 68.96% | | | | | |
| | ACI | 76.00% | 65.00% | 72.30% | 66.85% | 68.65% | | | | | |
| Random Forest | sin ACI | 79.85% | 74.55% | 82.00% | 69.80% | 78.28% | 1000 | 10 | 0.2 | 1 | 7 |
| | ACI | 81.25% | 76.50% | 81.14% | 71.22% | 78.82% | | | | | |
| XGBoost | sin ACI | 82.60% | 72.00% | 81.00% | 75.39% | 76.50% | | | | | |
| | ACI | 82.00% | 78.00% | 84.00% | 71.58% | 81.00% | | | | | |
| Random Forest | sin ACI | 77.98% | 71.65% | 75.68% | 70.59% | 73.67% | 1000 | 15 | 0.2 | 1 | 8 |
| | ACI | 78.50% | 78.00% | 78.00% | 61.35% | 78.00% | | | | | |
| XGBoost | sin ACI | 80.50% | 82.00% | 74.35% | 65.73% | 78.18% | | | | | |
| | ACI | 79.65% | 81.00% | 76.00% | 79.46% | 78.50% | | | | | |
| Random Forest | sin ACI | 83.15% | 82.46% | 79.80% | 71.00% | 81.13% | 2000 | 10 | 0.2 | 1 | 9 |
| | ACI | 84.00% | 83.00% | 86.00% | 74.87% | 84.50% | | | | | |
| XGBoost | sin ACI | 86.00% | 85.00% | 84.00% | 73.63% | 84.50% | | | | | |
| | ACI | 88.50% | 85.35% | 83.90% | 79.73% | 84.63% | | | | | |
| Random Forest | sin ACI | 82.30% | 82.40% | 76.50% | 79.91% | 79.45% | 2000 | 13 | 0.2 | 1 | 10 |
| | ACI | 84.25% | 80.56% | 77.98% | 78.56% | 79.27% | | | | | |
| XGBoost | sin ACI | 84.65% | 80.88% | 77.60% | 71.24% | 79.24% | | | | | |
| | ACI | 87.65% | 84.00% | 80.60% | 79.74% | 82.30% | | | | | |
| Random Forest | sin ACI | 82.65% | 80.95% | 79.50% | 71.83% | 80.23% | 2000 | 11 | 0.2 | 1 | 11 |
| | ACI | 82.92% | 82.76% | 80.30% | 74.70% | 81.53% | | | | | |
| XGBoost | sin ACI | 88.96% | 87.91% | 85.41% | 75.73% | 86.66% | | | | | |
| | ACI | 90.20% | 86.40% | 88.00% | 77.77% | 87.20% | | | | | |

La Tabla 5.2 muestra las seis pruebas realizadas en este segundo experimento, donde se evalúa con las cinco métricas, en las primeras tres pruebas (6, 7 y 8) se mantuvo constante el número de árboles y se fue aumentando la profundidad, donde se encuentra que la profundidad de 10 en Random Forest y 11 en XGBoost fueron los mejores parámetros.

En la Figura 5.3 se muestran los resultados del promedio de las métricas de evaluación de las 6 pruebas realizadas.



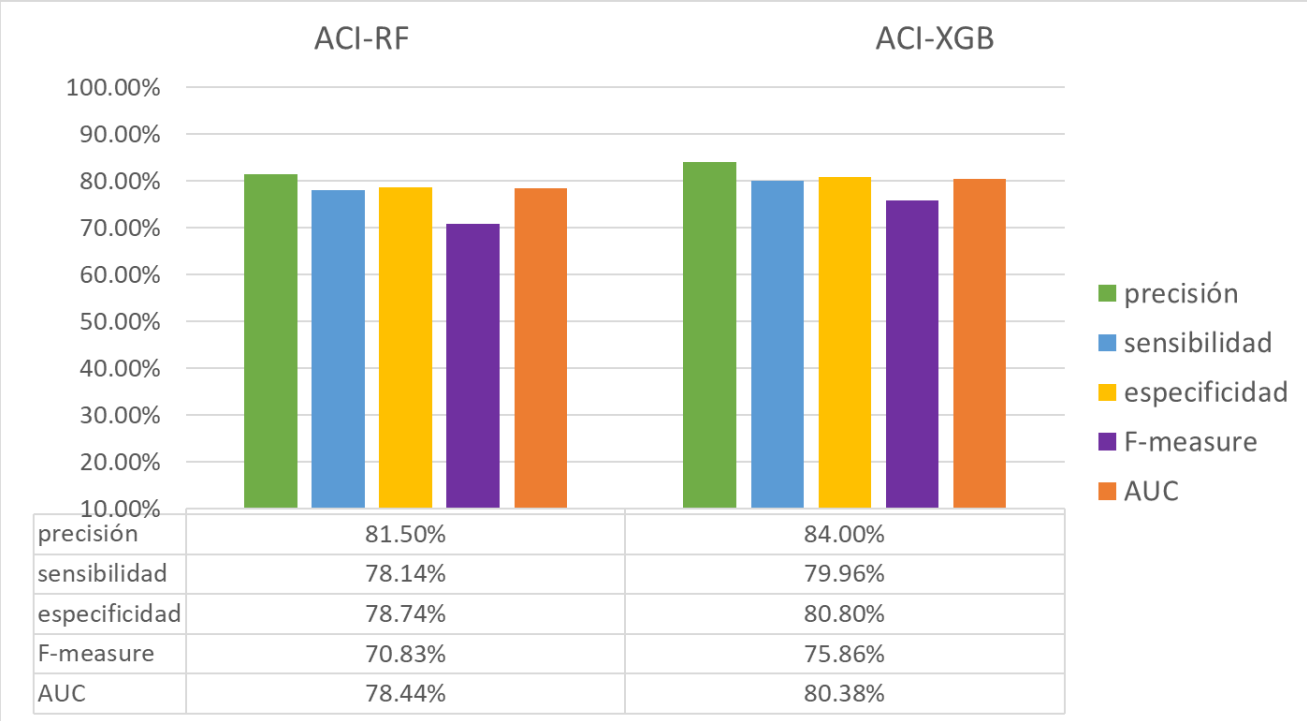

| | | |
|---|---|---|
| precisión | 80.36% | 82.83% |
| sensibilidad | 76.36% | 79.54% |
| especificidad | 76.93% | 78.47% |
| F-measure | 71.43% | 70.69% |
| AUC | 76.65% | 79.01% |

Figura 5.3. Promedio de experimento 2 sin ACI.

El promedio de la precisión de RF sin ACI fue de 80.36% donde el valor mínimo fue de 74.20% (profundidad de 5 y 1000 árboles) y un valor máximo de 83.15.65% (profundidad de 10 y 2000 árboles), una sensibilidad promedio de 76.36%, especificidad promedio de 76.93%, F-measure promedio de 71.43% y un AUC de 76.65%.

XGBoost sin ACI obtuvo una precisión promedio de 82.83%, con un valor mínimo de 74.24% (profundidad 5 y 1000 árboles) y un valor máximo de 88.96% (profundidad 11 y 2000 árboles), una sensibilidad de promedio de 79.54%, una especificidad promedio de 78.47%, F-measure promedio de 70.69% y un AUC promedio de 79.01%.

Posteriormente se hizo una comparación de los promedios de las métricas con los datos resultantes del ACI (ver Figura 5.4).




| | | |
|---|---|---|
| precisión | 81.50% | 84.00% |
| sensibilidad | 78.14% | 79.96% |
| especificidad | 78.74% | 80.80% |
| F-measure | 70.83% | 75.86% |
| AUC | 78.44% | 80.38% |

Figura 5.4. Promedio de experimento 2 con ACI.

El promedio de la precisión de RF con ACI fue de 81.50% donde el valor mínimo fue de 76.07% (profundidad de 5 y 1000 árboles) y un valor máximo de 84.92% (profundidad de 11 y 2000 árboles), una sensibilidad promedio de 78.14%, especificidad promedio de 78.74%, F-measure promedio de 70.83% y un AUC de 78.44%,

XGBoost con ACI obtuvo una precisión promedio de 84%, con un valor mínimo de 76% (profundidad 5 y 1000 árboles) y un valor máximo de 90.20% (profundidad 11 y 2000 árboles), una sensibilidad promedio de 79.96%, una especificidad promedio de 80.80%, F-measure promedio de 75.86% y un AUC promedio de 80.38%.

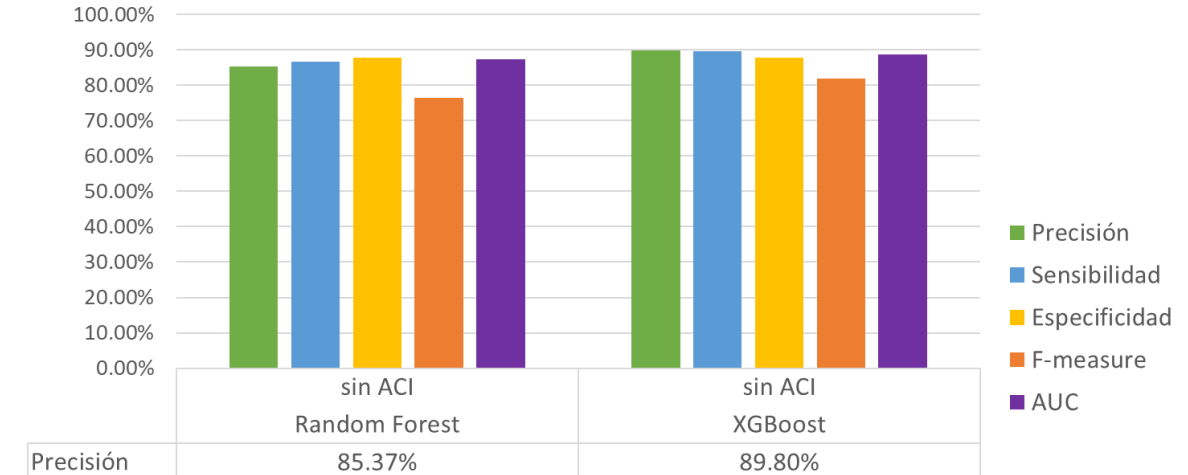
En este segundo experimento el valor más alto de las métricas de evaluación fue XGBoost con ACI, con una profundidad de 11 y 2000 árboles.

## 5.3 Experimentación 3: número de árboles y profundidad

En este tercer experimento, tras la comparación de las diferentes pruebas se dedujo que con una profundidad de 10 para Random Forest obtiene los valores más altos en las métricas en comparación; por otra parte, XGBoost con una profundidad de 11 alcanza los valores más altos respecto a las demás pruebas, por lo que se concluyó que esos eran los valores finales del parámetro "profundidad". Una vez ya determinado el valor de la profundidad, se experimentó con el aumento del número de árboles, donde de igual forma, se verificó su rendimiento con las cinco métricas ver Figura 5.5 y Figura 5.6.

Los parámetros utilizados en este experimento fueron:

número de árboles = 2,100, Profundidad = 10 (para RF) y 11 (para XGB), Tasa de aprendizaje = 0.2, Peso = 1.

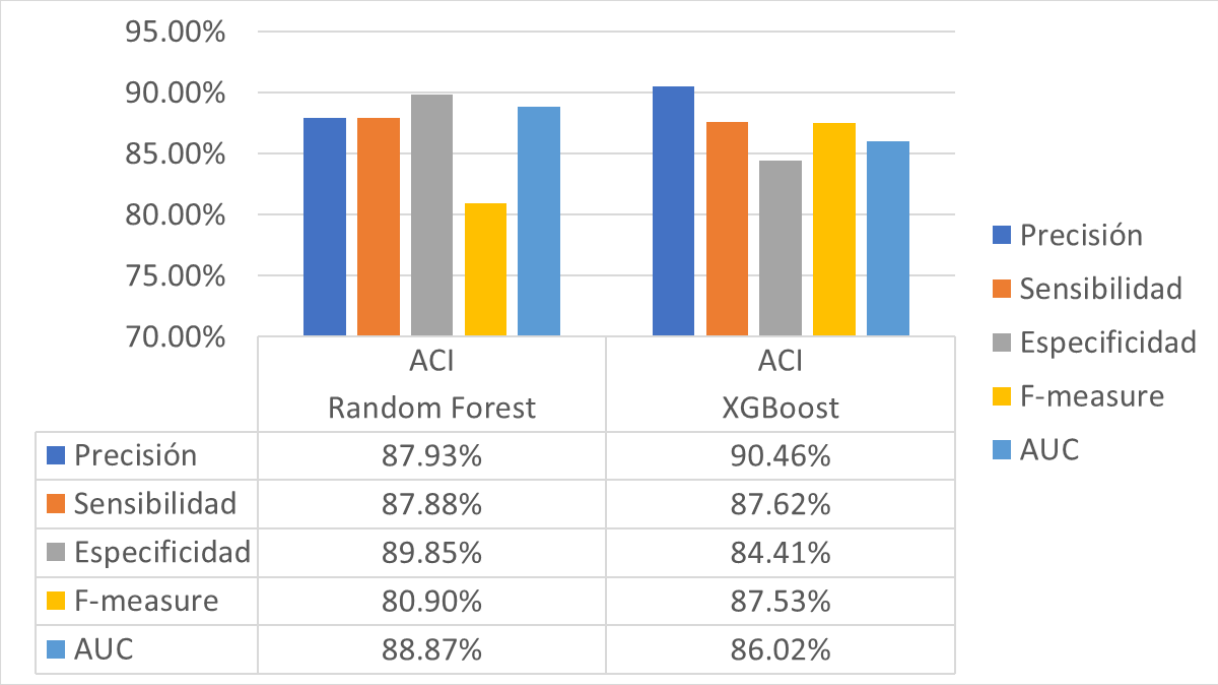



| | sin ACI Random Forest | sin ACI XGBoost |
|---|---|---|
| Precisión | 85.37% | 89.80% |
| Sensibilidad | 86.70% | 89.64% |
| Especificidad | 87.81% | 87.78% |
| F-measure | 76.43% | 81.76% |
| AUC | 87.26% | 88.71% |

Figura 5.5. Comparación de experimento 3 sin ACI.

La precisión de RF sin ACI fue de 85.37%, una sensibilidad de 86.70%, especificidad de 87.81%, F-measure de 76.43% y un AUC de 87.26%.

XGBoost sin ACI obtuvo una precisión de 89.80%, una sensibilidad de 89.64%, una especificidad de 87.81%, F-measure de 81.76% y un AUC de 88.71%.




| | ACI Random Forest | ACI XGBoost |
|---|---|---|
| Precisión | 87.93% | 90.46% |
| Sensibilidad | 87.88% | 87.62% |
| Especificidad | 89.85% | 84.41% |
| F-measure | 80.90% | 87.53% |
| AUC | 88.87% | 86.02% |

Figura 5.6. Comparación de experimento 3 con ACI.

La precisión de RF con ACI fue de 87.93%, una sensibilidad de 87.88%, especificidad de 89.85%, F-measure de 80.90% y un AUC de 88.87%.

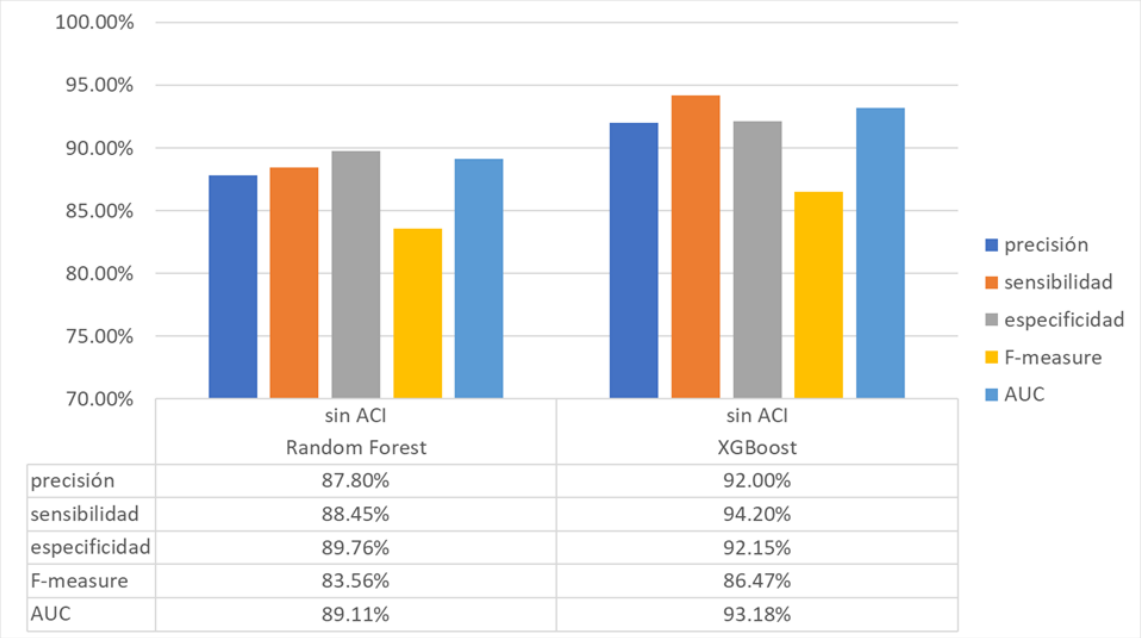
XGBoost con ACI obtuvo una precisión de 90.46%, una sensibilidad de 87.62%, una especificidad de 87.53%, F-measure de 87.53% y un AUC de 86.02%.

En este tercer experimento, se desea conocer los parámetros adecuados para cada algoritmo donde Random Forest obtuvo los valores más altos con 2800 árboles y una profundidad de 10, XGBoost obtuvo los valores más altos con 2200 árboles y una profundidad de 11.

## 5.4 Experimentación 4: tasa de aprendizaje

En este cuarto experimento se fue aumentando gradualmente el número de árboles partiendo de los parámetros del tercer experimento, con la finalidad de conocer el óptimo número de árboles para cada clasificador, también se modificó la tasa de aprendizaje para una rápida conversión en el entrenamiento, dando como resultado los siguientes parámetros y resultados (ver Figura 5.7 y Figura 5.8).

Parámetros: número de árboles = 2,800 RF y 2,200 XGB, Profundidad = 10 RF y 11 XGB, Tasa aprendizaje = 0.75, Peso = 1.

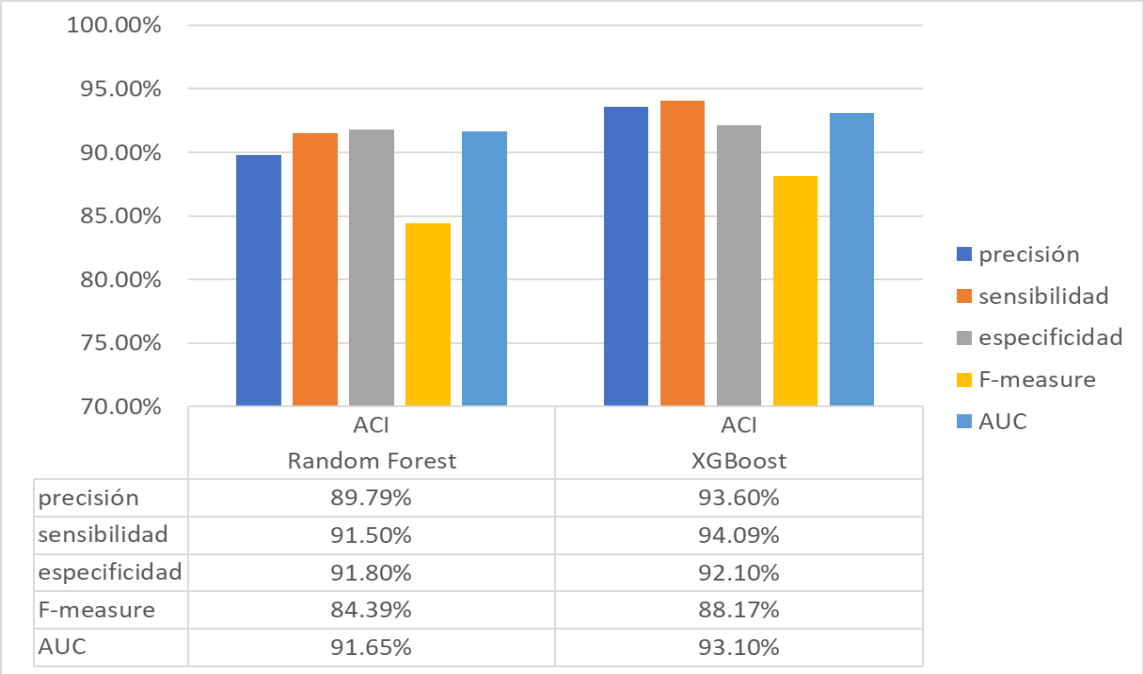



| | sin ACI Random Forest | sin ACI XGBoost |
|---|---|---|
| precisión | 87.80% | 92.00% |
| sensibilidad | 88.45% | 94.20% |
| especificidad | 89.76% | 92.15% |
| F-measure | 83.56% | 86.47% |
| AUC | 89.11% | 93.18% |

Figura 5.7. Comparación de resultados experimento 4 sin ACI.

La precisión de RF sin ACI fue de 87.80%, una sensibilidad de 88.45%, especificidad de 89.76%, F-measure de 83.56% y un AUC de 89.11%.

XGBoost sin ACI obtuvo una precisión de 92%, una sensibilidad de 94.20%, una especificidad de 92.15%, F-measure de 86.47% y un AUC de 93.18%.




| | ACI Random Forest | ACI XGBoost |
|---|---|---|
| precisión | 89.79% | 93.60% |
| sensibilidad | 91.50% | 94.09% |
| especificidad | 91.80% | 92.10% |
| F-measure | 84.39% | 88.17% |
| AUC | 91.65% | 93.10% |

Figura 5.8. Comparación de resultados experimento 4 con ACI.

La precisión de RF con ACI fue de 89.79%, una sensibilidad de 91.50%, especificidad de 91.80%, F-measure de 84.39% y un AUC de 91.65%.

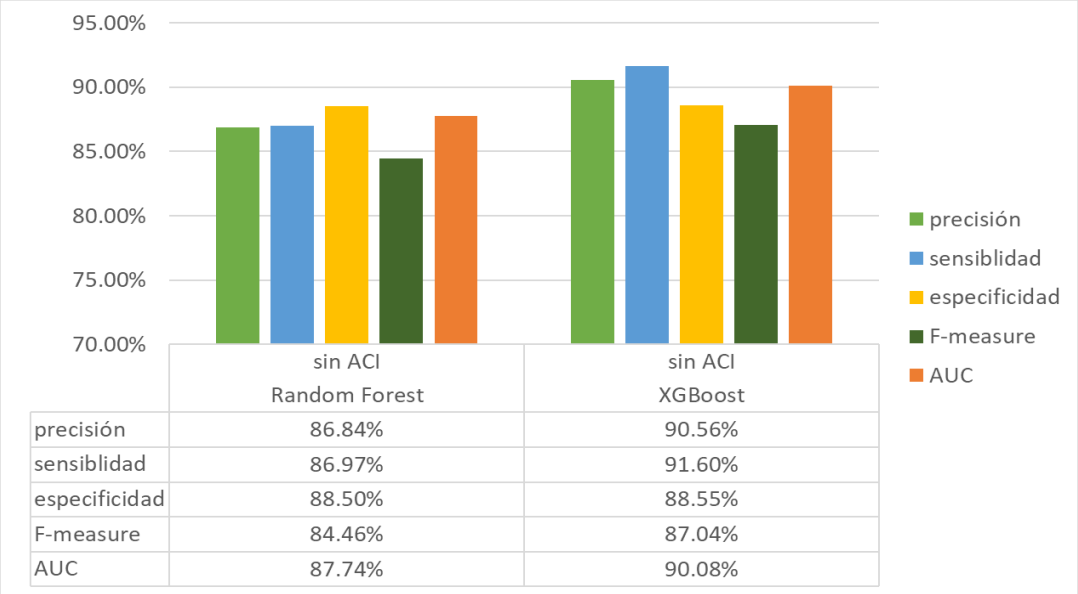
XGBoost con ACI obtuvo una precisión de 93.60%, una sensibilidad de 94.09%, una especificidad de 92.10%, F-measure de 88.17% y un AUC de 93.10%.

Con la tasa de aprendizaje de 0.75 se obtuvo una mejora en el tiempo de entrenamiento de los clasificadores, finalizando en 7 minutos menos respecto a la tasa de aprendizaje de los experimentos anteriores.

## 5.5 Experimentación 5: parámetro sobre ajustado

En este experimento se sobre ajustaron los parámetros con el aumento del número de árboles respecto al experimento anterior con el fin de conocer si, los parámetros del experimento 4 eran los adecuados o todavía se podía mejorar los clasificadores, cuyos parámetros y resultados (ver Figura 5.9 y Figura 5.10) fueron los siguientes:

número de árboles = 2,900 y 2,300, Profundidad = 10 y 11, Tasa aprendizaje = 0.75, Peso = 1.

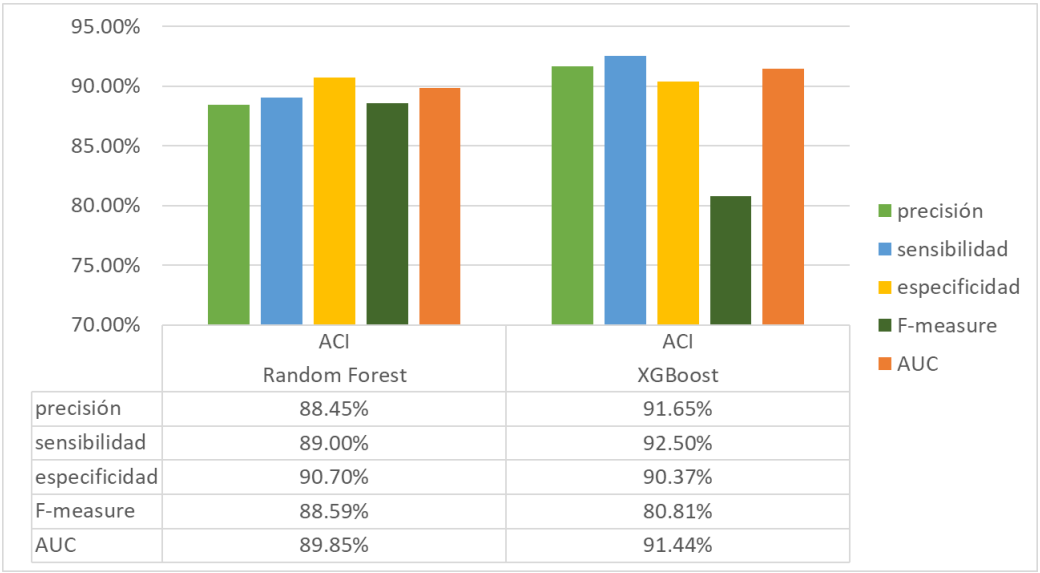



| | sin ACI Random Forest | sin ACI XGBoost |
|---|---|---|
| precisión | 86.84% | 90.56% |
| sensiblidad | 86.97% | 91.60% |
| especificidad | 88.50% | 88.55% |
| F-measure | 84.46% | 87.04% |
| AUC | 87.74% | 90.08% |

Figura 5.9. Comparación de resultados experimento 5 sin ACI.

La precisión de RF sin ACI fue de 86.84%, una sensibilidad de 86.97%, especificidad de 88.50%, F-measure de 84.46% y un AUC de 87.74%.

XGBoost sin ACI obtuvo una precisión de 90.56%, una sensibilidad de 91.60%, una especificidad de 88.55%, F-measure de 87.04% y un AUC de 90.08%.




| | ACI Random Forest | ACI XGBoost |
|---|---|---|
| precisión | 88.45% | 91.65% |
| sensibilidad | 89.00% | 92.50% |
| especificidad | 90.70% | 90.37% |
| F-measure | 88.59% | 80.81% |
| AUC | 89.85% | 91.44% |

Figura 5.10. Comparación de resultados experimento 5 con ACI.

La precisión de RF con ACI fue de 88.45%, una sensibilidad de 89%, especificidad de 90.70%, F-measure de 88.59% y un AUC de 89.85%.

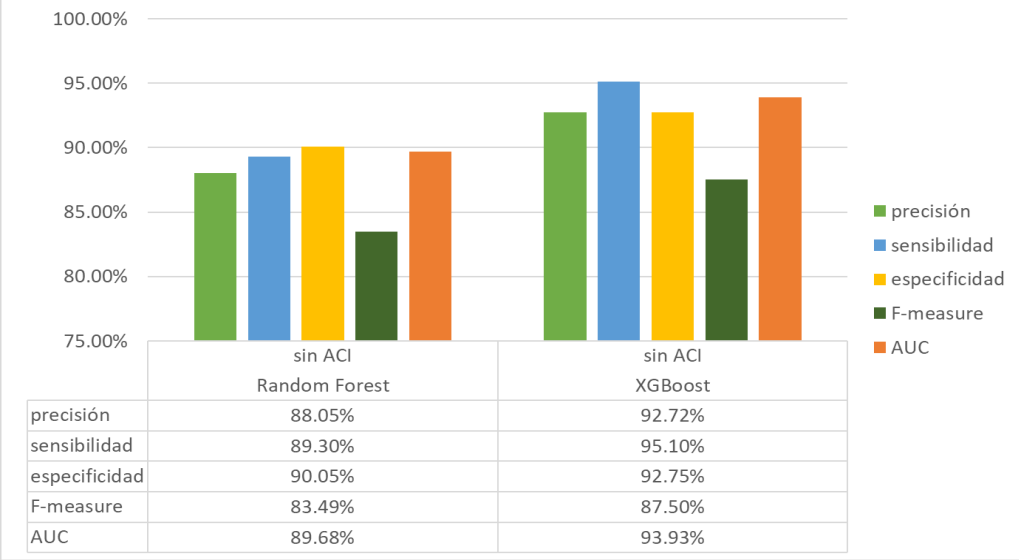
XGBoost con ACI obtuvo una precisión de 91.65%, una sensibilidad de 92.50%, una especificidad de 90.37%, F-measure de 80.81% y un AUC de 91.44%.

El objetivo de este experimento es probar si los parámetros son los adecuados o si, aumentando aún más, las métricas de igual forma aumentan, por lo que no fue así. Se concluyó que, de acuerdo con los experimentos realizados, los parámetros obtenidos con el experimento 4 son los óptimos.

## 5.6 Experimentación 6: validación cruzada

En este experimento se utilizó la validación cruzada *(K Cross validation*) respecto a los parámetros finales del experimento 4, con la finalidad de conocer si aumentaba el valor de las métricas, cuyos parámetros y resultados (ver Figura 5.11 y Figura 5.12) fueron los siguientes:

Parámetros: número de árboles = 2,800 y 2,200, Profundidad = 10 (Random Forest) y 11 (XGBoost), Tasa aprendizaje = 0.75, Peso = 1, *k* = 10 fold.

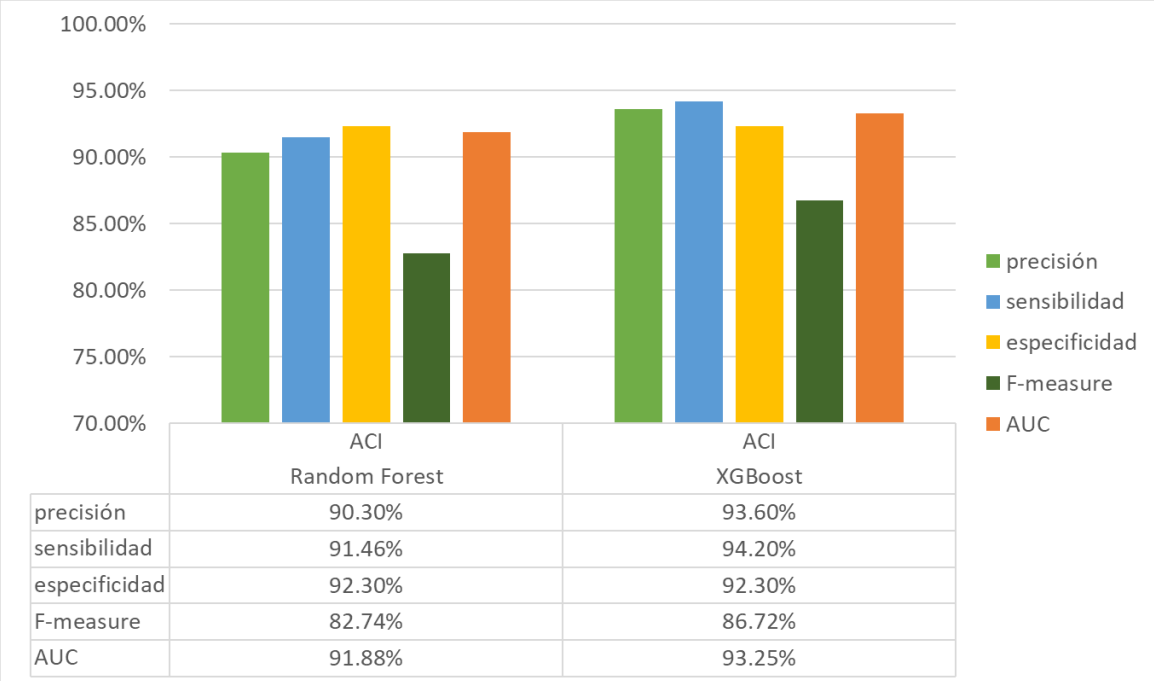



| | sin ACI Random Forest | sin ACI XGBoost |
|---|---|---|
| precisión | 88.05% | 92.72% |
| sensibilidad | 89.30% | 95.10% |
| especificidad | 90.05% | 92.75% |
| F-measure | 83.49% | 87.50% |
| AUC | 89.68% | 93.93% |

Figura 5.11. Comparación de resultados experimento 6 sin ACI.

La precisión de RF con ACI fue de 88.05%, una sensibilidad de 89.30%, especificidad de 90.05%, F-measure de 83.49% y un AUC de 89.68%.

XGBoost sin ACI obtuvo una precisión de 92.72%, una sensibilidad de 95.10%, una especificidad de 92.75%, F-measure de 87.50% y un AUC de 93.93%.




| | ACI Random Forest | ACI XGBoost |
|---|---|---|
| precisión | 90.30% | 93.60% |
| sensibilidad | 91.46% | 94.20% |
| especificidad | 92.30% | 92.30% |
| F-measure | 82.74% | 86.72% |
| AUC | 91.88% | 93.25% |

Figura 5.12. Comparación de resultados experimento 6 con ACI.

La precisión de RF con ACI fue de 90.30%, una sensibilidad de 91.46%, especificidad de 92.30%, F-measure de 82.74% y un AUC de 91.88%.

XGBoost con ACI obtuvo una precisión de 93.60%, una sensibilidad de 94.20%, una especificidad de 92.30%, F-measure de 86.72% y un AUC de 93.25%.

El objetivo de este experimento se desea conocer si con una validación cruzada de *k*=10, aumenta nuestras métricas por lo que se confirmó, pero el aumento no fue significativo, donde si fue significativo el aumento es en el tiempo de entrenamiento. Como se muestra a continuación.

## 5.7 Tiempo de entrenamiento en los clasificadores

En la Tabla 5.3 se observa los tiempos de entrenamiento que se emplean en los clasificadores Random Forest y XGBoost, en combinación con/sin ACI con el propósito de combinar la transformación de las variables y su clasificación, lo que permite observar la aceleración en la convergencia del clasificador.

Se utilizaron los parámetros del experimento 4.

Tabla 5.3. Tiempo en entrenamiento experimento 4

| Tiempo de entrenamiento | | |
|---|---|---|
| | XGBoost | Random Forest |
| ACI | 11:52 h | 16:39 h |
| Sin ACI | 12:38 h | 19:56 h |

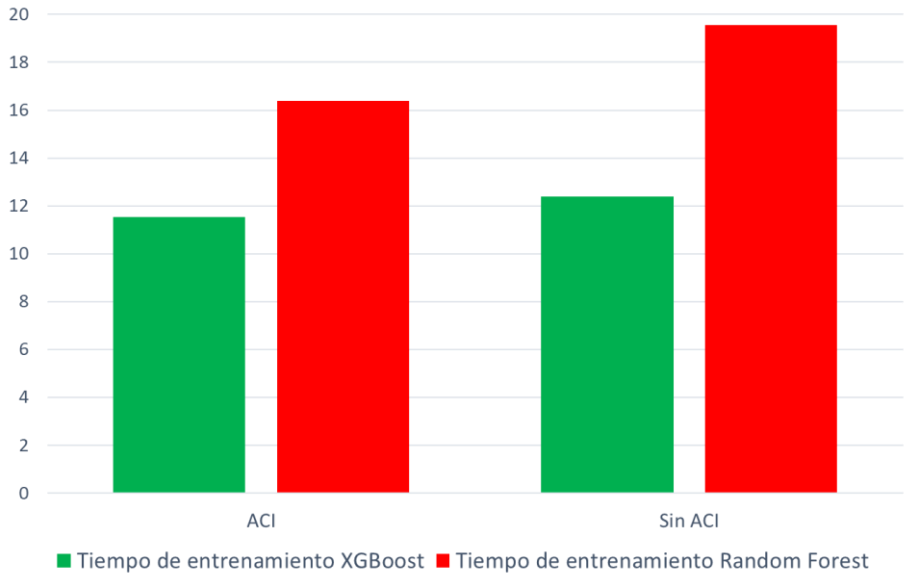



Figura 5.13. Tiempos de entrenamiento experimento 4

En la Figura 5.13 muestra el tiempo en horas que requiere Random Forest y XGBoost para completar la etapa de entrenamiento con ACI y sin ACI.

En los experimentos realizados, se aprecian dos experimentos (Experimento 4 y Experimento 6) con valores similares en las métricas de evaluación, por lo que la variable determinante es el tiempo de entrenamiento, donde el Experimento 4 (Figura 5.13) toma menos tiempo en comparación del experimento 6 (ver Figura 5.14).

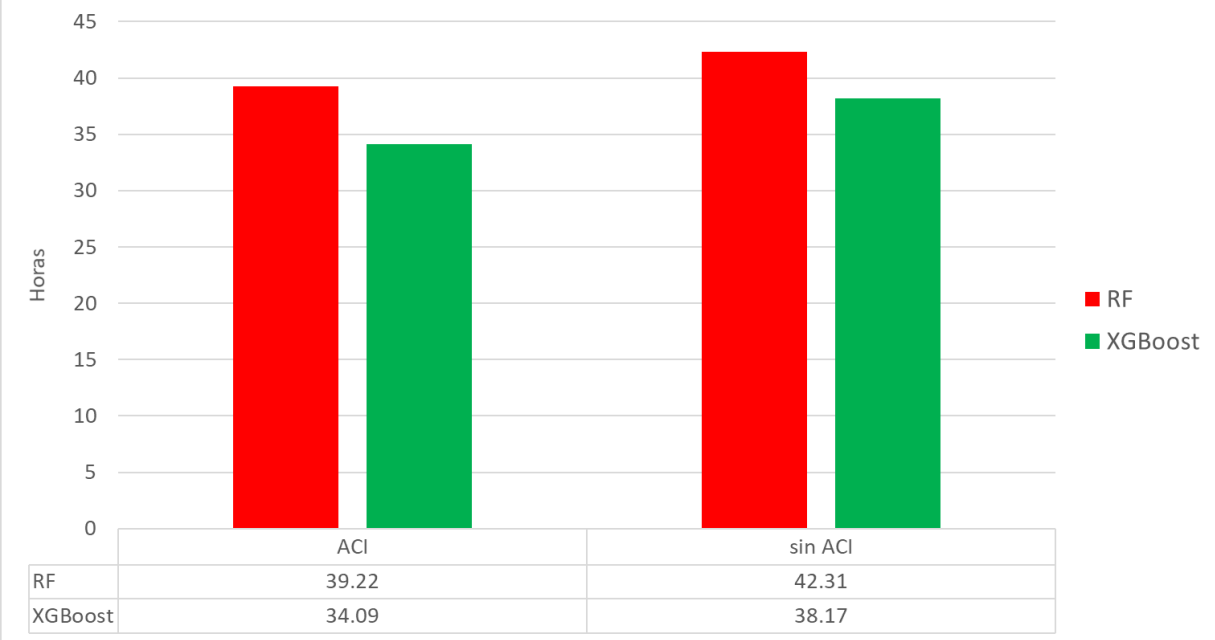



Figura 5.14. Tiempos de entrenamiento experimento 6.

En la Figura 5.16 muestra 39.22 hrs. en el entrenamiento de Random Forest con ACI y 42.31 hrs. sin ACI. Mientras que XGBoost, toma 34.09 hrs en terminar el entrenamiento con ACI y 38.17 hrs sin ACI.

## 5.8 Paralelización de los clasificadores

La paralelización se ejecutó exclusivamente para el entrenamiento y se compararon los clasificadores (ver Tabla 5.4)

Tabla 5.4. Comparación de tiempos experimento 4

| Tiempo de entrenamiento | | |
|---|---|---|
| | XGBoost | Random Forest |
| ACI | 11:52h | 16:39h |
| Sin ACI | 12:38h | 19:56h |
| | | |
| | XGBoost paralelizado | Random Forest paralelizado |
| ACI | 7:49h | 11:56h |
| Sin ACI | 9:20h | 13:43h |

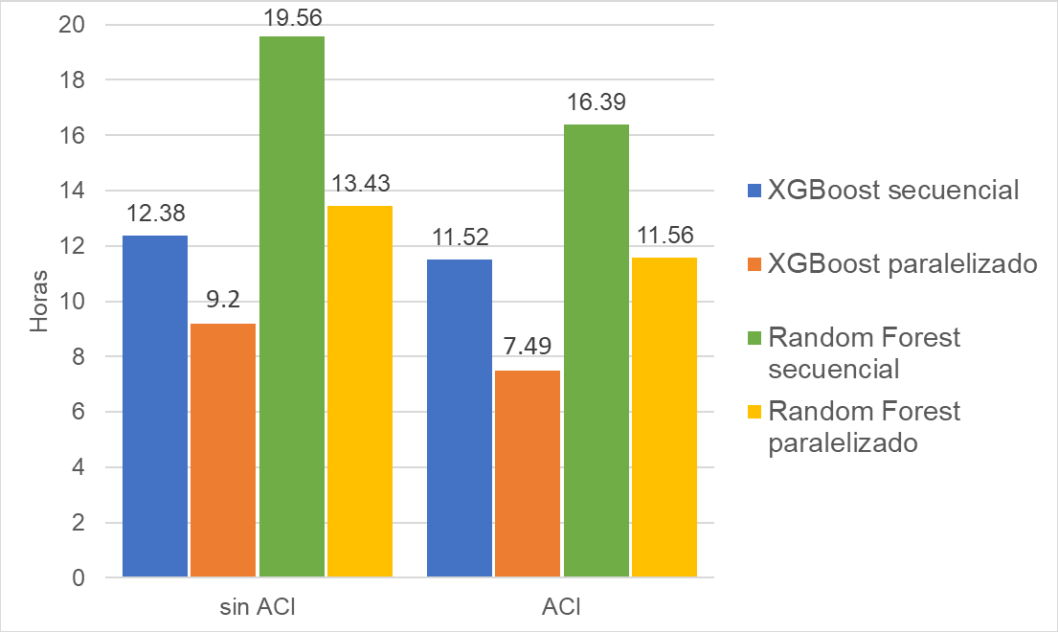



Figura 5.15. Comparación secuencial y paralelizado experimento 4.

En la Figura 5.15 se observa la diferencia de 4.03 horas en el proceso secuencial respecto al paralelizado de XGBoost con ACI. En tanto que sin ACI, XGBoost obtiene una diferencia de 3.18 horas. Con Random Forest la diferencia fue de 5.23 horas con ACI, sin ACI Random Forest obtuvo una diferencia de 6.13 horas.

Tabla 5.5. Comparación de tiempos experimento 6

| secuencial | | | paralelizado | | |
|---|---|---|---|---|---|
| | RF | XGBoost | | RF | XGBoost |
| ACI | 39.22 | 34.09 | ACI | 24.36 | 20.22 |
| sin ACI | 42.31 | 38.17 | sin ACI | 36.14 | 24.51 |

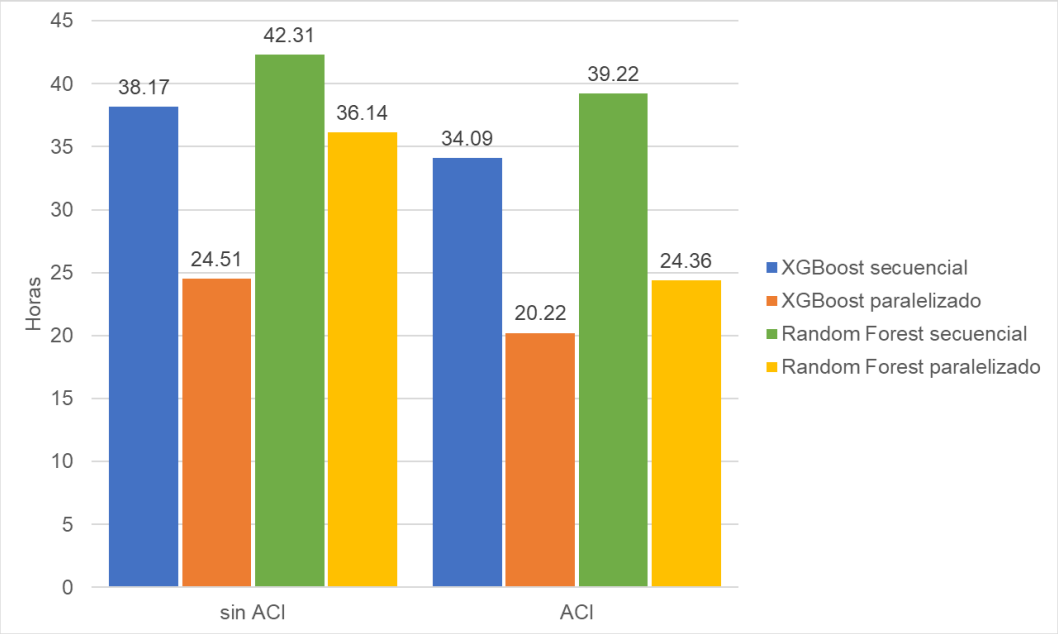



Figura 5.16. Comparación secuencial y paralelizado experimento 6.

En la Figura 5.16 y Tabla 5.5 se observa la diferencia de 15.36 horas en el proceso secuencial respecto al paralelizado de Random Forest con ACI. En tanto que sin ACI, Random Forest obtiene una diferencia de 6.17 horas. Para la XGBoost la diferencia fue de 14.27 horas en el proceso secuencial respecto al paralelizado con ACI, sin ACI, XGBoost obtuvo una diferencia de 14.06 horas.

En la Figura 5.17, se observa la utilización de todos los núcleos del CPU para el proceso de la paralelización, el cual permite la rápida convergencia en el entrenamiento de los algoritmos de clasificación.

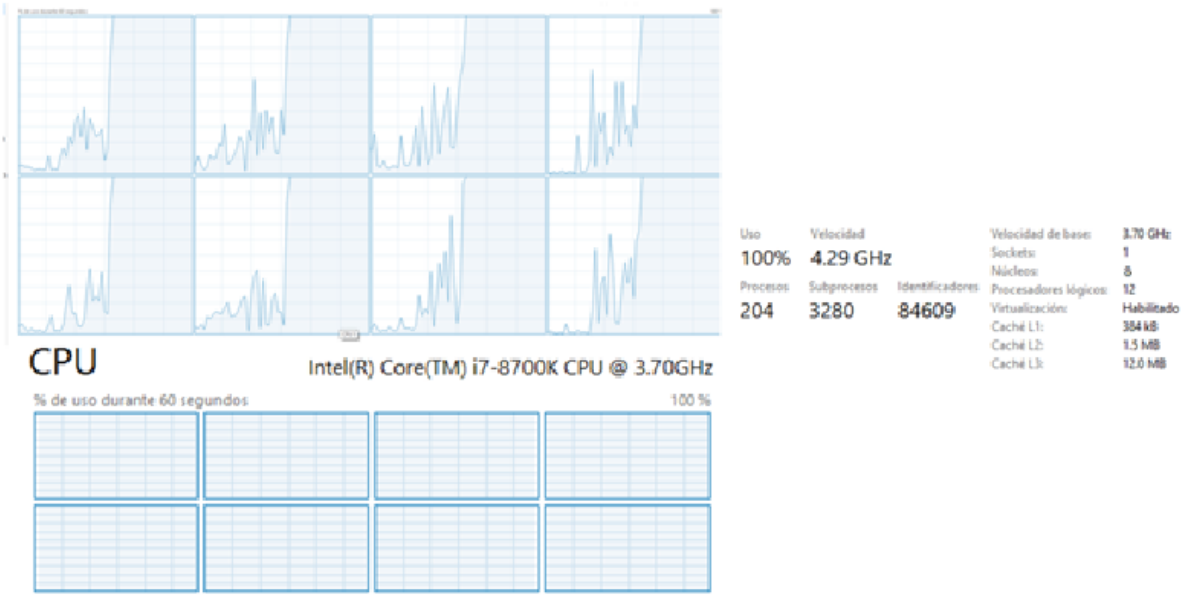



Figura 5.17. Paralelización.

## 5.9 Análisis de resultados

Durante los experimentos, se fueron modificando los parámetros para tener los mejores valores en las métricas de evaluación, por lo que se dio a la tarea de hacer el cálculo de la precisión por clase (ver Figura 5.18 y Tabla 5.6), se utilizó el experimento 4 para dicho cálculo ya que es el experimento donde mejores resultados se obtuvo.

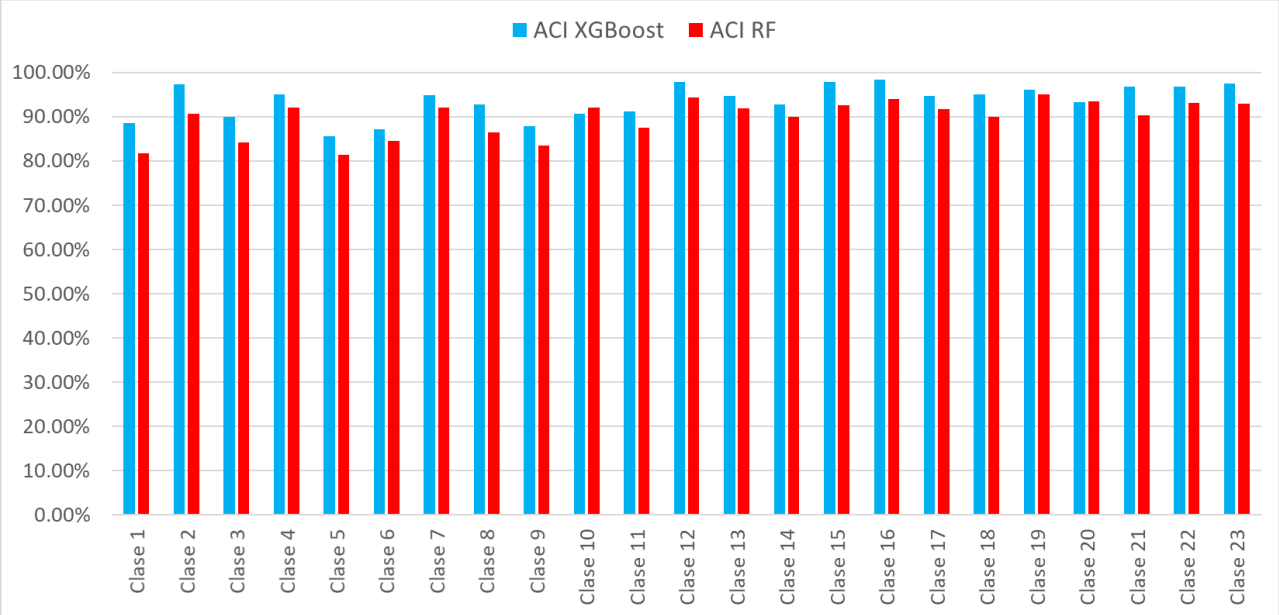



Figura 5.18. Precisión por clase

Tabla 5.6. Precisión por cada clase.

| | ACI XGBoost | ACI RF |
|---|---|---|
| Clase 1 | 88.62% | 81.66% |
| Clase 2 | 97.39% | 90.60% |
| Clase 3 | 89.91% | 84.21% |
| Clase 4 | 95.05% | 92.13% |
| Clase 5 | 85.60% | 81.46% |
| Clase 6 | 87.12% | 84.51% |
| Clase 7 | 94.87% | 92.02% |
| Clase 8 | 92.71% | 86.43% |
| Clase 9 | 87.94% | 83.45% |
| Clase 10 | 90.70% | 91.98% |
| Clase 11 | 91.25% | 87.51% |
| Clase 12 | 97.79% | 94.36% |
| Clase 13 | 94.67% | 91.82% |
| Clase 14 | 92.82% | 89.96% |
| Clase 15 | 97.88% | 92.65% |
| Clase 16 | 98.40% | 94.05% |
| Clase 17 | 94.61% | 91.77% |
| Clase 18 | 95.00% | 89.90% |
| Clase 19 | 96.13% | 95.00% |
| Clase 20 | 93.23% | 93.47% |
| Clase 21 | 96.78% | 90.26% |
| Clase 22 | 96.84% | 93.15% |
| Clase 23 | 97.57% | 92.90% |
| Precisión final | 93.60% | 89.79% |

En la Tabla 4.1 y Figura 4.3 se muestra la distribución del número de objetos o ejemplos por clase, donde la clase 2, 15 y 16 muestran el mayor número de ejemplos y la clase 9 muestra el menor número de ejemplos. Dónde al hacer el análisis se puede determinar que la clase 9 es la que cuenta con el valor más bajo de precisión en XGBoost obtuvo 87.94% y en Random Forest un 83.45%, por otra parte, la clase 15 que es la cuenta con la mayor cantidad de ejemplos (2442) alcanza un 97.88% en XGBoost y un 92.65% en Random Forest, pero otro ejemplo es la clase 19 que alcanza el 96.13% en XGBoost y un 95% en Random Forest y cuenta con 363 ejemplos, por lo que se puede concluir que el valor de la precisión no es proporcional a la cantidad de ejemplos u observaciones por clase.

Las evaluaciones efectuadas a los clasificadores en este trabajo permiten destacar que la combinación ACI-XGBoost del experimento 4 (ver Figura 5.19) tuvo mejor desempeño respecto al tiempo, precisión, sensibilidad, especificidad, F-measure y AUC.

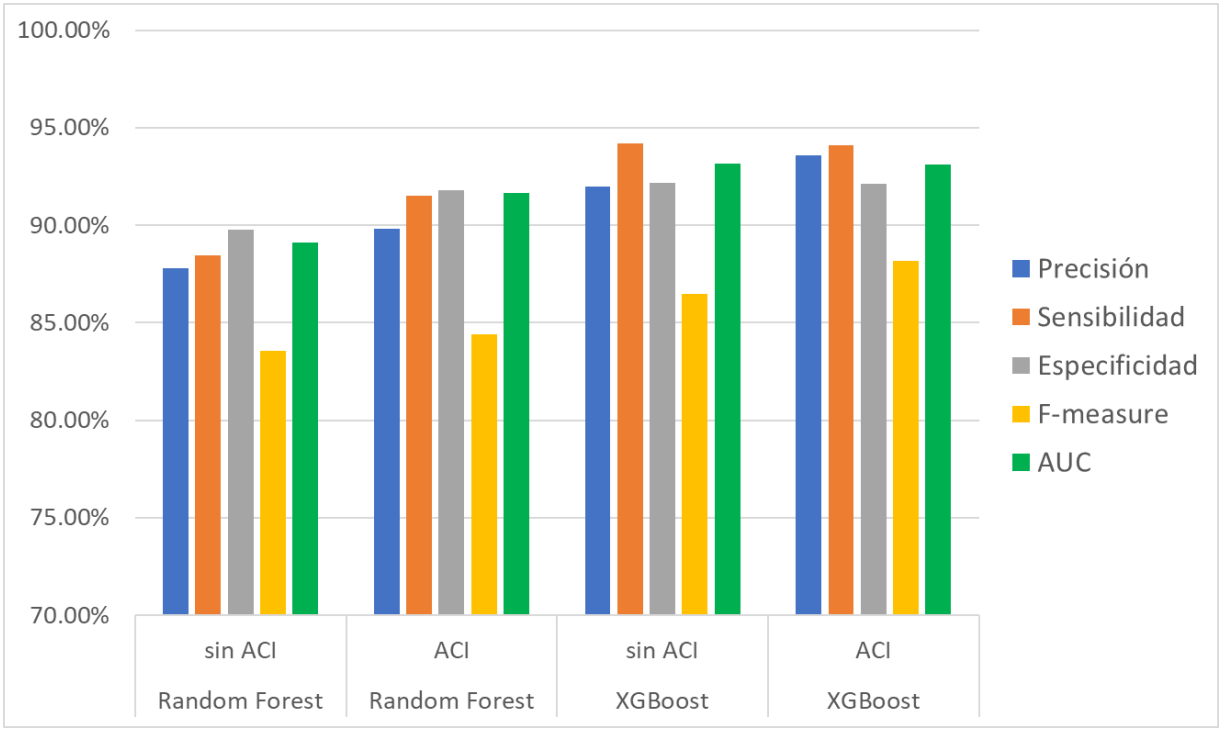



Figura 5.19. Comparación de los clasificadores

Una vez realizada la evaluación, esta permite identificar los genes con mayor peso en la clasificación. Es decir, estos genes podrían ser los marcadores seleccionados con un grado de relevancia como biomarcadores de cáncer de hueso. Es importante destacar que se requiere de la validación biológica y asignación de valores de riesgo para su utilización en la clínica.

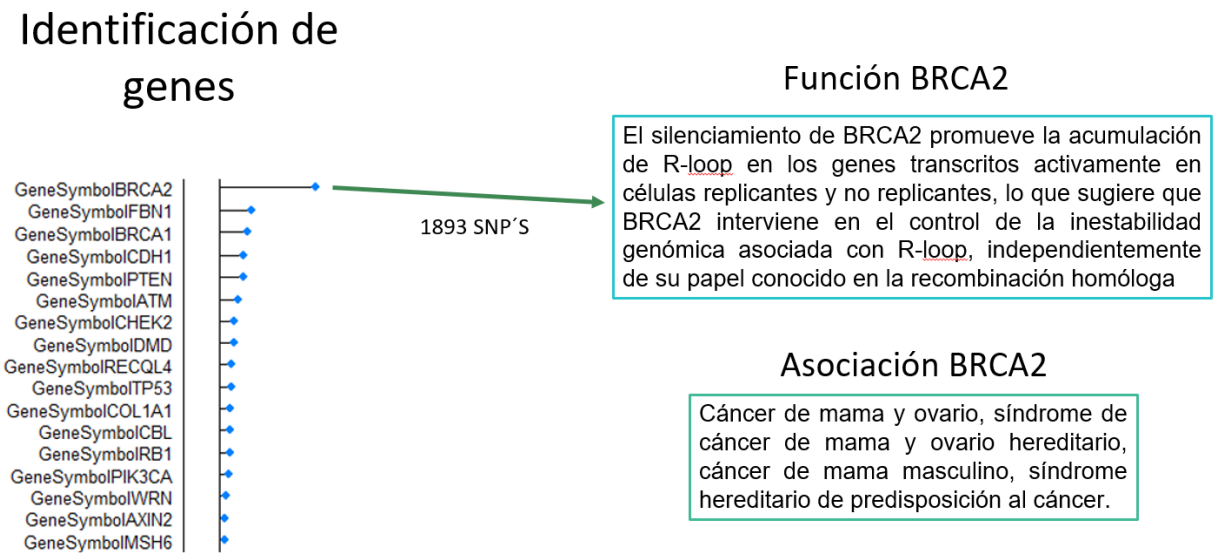



Figura 5.20. Identificación de genes

Adicionalmente al reconocimiento de genes, se puede establecer la asociación biológica, metabólica y clínica, esto se realiza por medio del atributo “Clase”, la cual cuenta con un PubMed, donde se puede realizar la búsqueda de la información de un gen especifico (ver Figura 5.20), información que ya se validó por el experto, el Co-asesor de este trabajo. Esto permite que, la identificación de estos *SNP´*s, se puede sugerir un tratamiento farmacológico.

# CAPÍTULO 6. CONCLUSIONES

En este capítulo se muestra las conclusiones, cumplimiento de los objetivos, aportaciones, perspectivas y productos adicionales.

## 6.1 Conclusiones

- Se cumplió con los objetivos general y específicos.
- Se obtuvo mejores resultados con XGBoost- ACI, superando en un 3.81% a Random Forest- ACI.
- Se eficientó el proceso del entrenamiento a través de la paralelización, logrando una reducción de 4.03 horas (29.3%) en XGBoost con ACI y 5.23 horas (30.9%) en Random Forest con ACI.

## 6.2 Cumplimiento de los objetivos

Se cumplieron los objetivos, general (Tabla 6.1) y específicos (Tabla 6.2).

Tabla 6.1. Cumplimiento del objetivo general.

| Objetivo general | Cumplimiento |
|---|---|
| Investigar, aplicar y evaluar estrategias de aprendizaje automático artificial para el manejo de datos genómicos relacionados con el diagnóstico temprano de cáncer de hueso en humanos. | Se investigaron, aplicaron y evaluaron dos algoritmos, Random Forest y XGBoost, capaces de procesar grandes cantidades de datos, tanto cuantitativamente, cualitativamente y tipo cadena. Donde se evaluaron con cinco métricas (precisión, sensibilidad, especificidad, F-measure y AUC. |

Tabla 6.2. Cumplimiento de los objetivos específicos.

| Objetivos específicos | Cumplimiento |
|---|---|
| a) Análisis y estudio de al menos dos algoritmos de aprendizaje automático artificial aplicado a información relacionada a cáncer de hueso y su diagnóstico. | Se analizaron y estudiaron cinco algoritmos de aprendizaje automático cuyas características permitieran procesar grandes cantidades de datos descritos. Se revisaron cinco algoritmos (Rpart, Perceptron multicapa, Adaboost, Random Forest y XGBoost), de los cuales se seleccionaron dos (Random Forest y XGBoost). |
| b) Ejecución de los algoritmos seleccionados considerando características de muestras genéticas o *SNP*s. | Se aplicaron Random Forest y XGBoost para grandes cantidades de datos, como lo son las muestras genéticas y se realizó la paralelización |
| c) Evaluación de algoritmos de aprendizaje automático seleccionados. | Se evaluaron los algoritmos con las siguientes cinco métricas:<br>Precisión, Sensibilidad, Especificidad, F-measure y AUC. |
| d) Identificación de patrones en muestras genéticas de cáncer de hueso. | A través de XGBoost se pueden identificar aquellos patrones o genes más significativos (Figura 35). |

## 6.3 Aportaciones

Seguidamente se detallan las aportaciones más relevantes realizadas en este trabajo.

a) Manejo de 16,482 datos genómicos relacionados al cáncer de hueso, provenientes de la secuenciación de ADN.

b) Implementación de algoritmo reciente (XGBoost). Donde se realizó una comparación con uno de los algoritmos más utilizados, Random Forest.

c) Modificación de librerías para la implementación de la paralelización. Se paralelizó la etapa de entrenamiento para una rápida convergencia en los clasificadores.

## 6.4 Perspectivas y trabajo futuro

Las perspectivas del presente trabajo se resumen en los siguientes puntos:

- Experimentar con un *hardware* más especializado.
- Comparar XGBoost con otros clasificadores.
- Trabajar con otro banco de datos para verificar la metodología planteada en este proyecto.
- Aprendizaje continuo, actualizar el modelo en cada *set* de datos nuevos.
- Ancestría genética (distribución bio-geográfica (cómo varía la genética a lo largo de regiones geográficas) de la variación genética de las poblaciones humanas).
- Identificar perfiles genéticos /moleculares para el desarrollo de nuevos agentes terapéuticos.
- Identificar biomarcadores.

## 6.5 Productos adicionales

A continuación, se muestran productos académicos que no estuvieron considerados en el plan de trabajo original.

- **Participación como ponencia “aprendizaje automático”**

Participación llevada a cabo en la Universidad Tecnológica del Estado de Morelos, llevado a cabo en julio del 2018. El reconocimiento se muestra en Anexo C.

- **Participación en Escuela de Inteligencia Artificial 2018**

Participación llevada a cabo en la Universidad Tecnológica del Estado de Morelos, llevado a cabo en octubre del 2018. El reconocimiento se muestra en Anexo C.

- **Presentación de póster ICMEAE**

Se realizó la presentación de un póster en el ICMEAE (International Conference on Mechatronics, Electronics and Automotive Engineering), llevado a cabo en noviembre 2018. ISBN: 978-1-5386-9190-8. El reconocimiento se muestra en el Anexo C.

- **Proyecto de innovación y desarrollo tecnológico ICMEAE**

Se participó en el ICMEAE (International Conference on Mechatronics, Electronics and Automotive Engineering) llevado a cabo en noviembre 2018, en la categoría de innovación en el "Advanced Robotics and Drone Competition". El reconocimiento se muestra en el Anexo C.

- **Colaboración en el trabajo realizado en el XXXII Congreso Nacional de Bioquímica**

Se presentó el trabajo "identification and classification of single nucleotide polymorphisms as biomarkers associated with dementia through data mining and Machine Learning", llevada a cabo en Ixtapa, Zihuatanejo, en noviembre del 2018, cuya memoria electrónica para su consulta es: http://smb.org.mx/wp-content/uploads/2018/01/Memoria_XXXII_CNB.pdf, El reconocimiento se muestra en Anexo C.

- **Estancia en el BSMT de la UAEM**

Se realizó una estancia durante los meses de febrero, marzo, abril y mayo del 2019 en el Laboratorio de Biología de Sistemas y Medicina Traslacional (BSMT) de la Universidad Autónoma del Estado de Morelos (UAEM) cuya carta de liberación se muestra en el Anexo D.

# REFERENCIAS

**[Arias, 2016]** Arias, S. O. (2016). Aplicación de la inteligencia artificial en la bioinformática. UGCiencia, pp. 159-171.

**[Arockia, 2015]** Arockia, D. (2015). Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection. Procedia Computer Science. pp. 13-21.

**[Asri, 2016]** Asri, H. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, vol. 83, pp. 1064 – 1069.

**[Atkinson, 2001]** Atkinson, W. C. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clinical Pharmacology and Therapeutics, pp. 89-95.

**[Barrón, 2008]** Barrón, M. A. (2008). Desarrollo de un prototipo para la aplicación de técnicas de minería de datos sobre una base de datos real de base poblacional de cáncer, Tesis, CENIDET.

**[Behravan, 2018]** Beharavan. (2018). Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. National Library of Medicine National Institutes of Health. pp 1-16.

**[Bhola, 2015]** Bhola, A. (2015). Machine learning based approaches for cancer classification using gene expression data. Machine Learning and Applications: An International Journal (MLAIJ)

**[Biton, 2014]** Biton, A. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes, Cell Reports. Vol. 9. pp. 1235 - 1245.

**[Bouzalmat, 2014]** Bouzalmat, A. (2014). Comparative Study of PCA, ICA, LDA using SVM classifier. Journal Of Emerging Technologies In Web Intelligence, pp. 64-68.

**[Breiman, 2001]** Breiman, L. (2001). Random Forests. Machine Learning, Kluwer Academic Publishers, pp. 5-32.

**[Burrell, 2013]** Burrell, N. M. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. nature, pp. 338–345.

**[Cano, 2014]** Cano, T. (2014). Capítulo 1.- Redes Neuronales y Reconocimiento de Patrones. Airene, (August), pp. 1–11.

**[Caret, 2018]** Caret (2018), librería de R, https://cran.r-project.org/web/packages/caret/caret.pdf, fecha de consulta octubre 2018.

**[Checa, 2017]** Checa A. (2017). Gen: Desde el código genético hasta la ingeniería genética, Conogasi.org Sitio web: http://conogasi.org/articulos/gen-desde-el-codigo-genetico-hasta-la-ingenieria-genetica/, fecha de consulta febrero 2019.

**[Chen, 2016]** Chen, T. (2016). XGBoost: A Scalable Tree Boosting System. International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

**[Díaz, 2015]** Díaz, H. B. (2015). Algoritmos de aprendizaje automático para clasificación de Splice Sites en secuencias genómicas. Revista Cubana de Ciencias Informáticas, pp. 155-170.

**[Dimitrakopoulos, 2018]** Dimitrakopoulos G. N. (2018), Pathway analysis using XGBoost classification in Biomedical Data.SETN

**[doParallel, 2019]** doParallel (2019), librería de R, https://cran.r-project.org/web/packages/doParallel/doParallel.pdf, fecha de consulta marzo 2019

**[Dplyr, 2018]** Dplyr (2019), libreria de R, https://cran.r-project.org/web/packages/dplyr/dplyr.pdf, fecha de consulta octubre 2018.

**[Draper, 2003]** Draper, B. A. (2003). Recognizing faces with PCA and ICA. Computer Vision and Image Understanding, pp. 115-119.

**[Duval-Poo, 2012]** Duval-Poo, M. A. (2012). combinaciones de clasificadores supervisados: estado del arte. La Habana, Cuba: CENATAV.

**[e1071, 2018]** e1071 (2018), librería de R, https://cran.r-project.org/web/packages/e1071/e1071.pdf, fecha de consulta noviembre 2018

**[FastICA, 2018]** FastICA (2018), librería de R, https://cran.r-project.org/web/packages/fastICA/fastICA.pdf, fecha de consulta noviembre 2018

**[Floor, 2012]** L.Floor, S., J. E. (2012). Hallmarks of cancer: of all cancer cells, all the time? Cell Press, pp. 509-515.

**[FSelector, 2018]** Fselector (2018), paquetería de R, https://cran.r-project.org/web/packages/FSelector/FSelector.pdf, fecha de consulta septiembre 2018.

**[García, 2016]** García, S. (2016). Big Data: Preprocesamiento y calidad de datos. Novática(Revista de la Asociación de Técnicos de Informática), pp.17-23.

**[Gartner, 2015]** Gartner, L., G. A.V. (2015). Identification of a Putative Ganoderic Acid Pathway Enzyme in a Ganoderma Australe Transcriptome by Means of a Hidden Markov Model. 9th International Conference on Practical Applications of Computational Biology and Bioinformatics (pp. 107-115). España: Springer International Publishing.

**[Genuer, 2017]** Genuer, R. (2017), Random Forests for Big Data. Big Data Research, Elsevier, 2017, 9, pp.28-46. ff10.1016/j.bdr.2017.07.003ff. ffhal01233923v2.

**[Gil, 2016]** Gil, G. (2016). APRENDIZAJE PROFUNDO El poder del aprendizaje automático unido al poder de cálculo de las computadoras actuales. Puebla, México.

**[Ggplot2, 2018]** Ggplot (2018), librería de R. https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf, fecha de consulta octubre 2018

**[He, 2017]** He, Y., (2017). A support vector machine classifier for the prediction of osteosarcoma metastasis with high accuracy. National Library of Medicine National Institutes of Health.pp. 1357-1364

**[Horne, 2016]** Horne B.D, M. J. (2015). Health effects of intermittent fasting: hormesis or harm? A systematic review. ncbi, pp. 464-470.

**[Iterators, 2019]** Iterators (2019), librería de R, https://cran.r-project.org/web/packages/iterators/iterators.pdf, fecha de consulta marzo 2019

**[Kavakiotis, 2017]** Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, pp. 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

**[Khiabani, 2016]** Khiabani, B. F. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. Journal of Clinical Epidemiology, Volume 71, pp. 76-85.

**[Klein, 2006]** Klein, M. J. (2006). Osteosarcoma Anatomic and Histologic Variants. American Journal of Clinical Pathology, pp. 555-581.

**[Leung, 2016]** Leung, M. K. (2016). Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. Pp. 176 – 197

**[Li, 2017]** Li Z. (2017). Classifying osteosarcoma patients using machine learning approaches. National Library of Medicine National Institutes of Health. pp. 82-85

**[Lodziensis, 2014]** Lodziensis, A. U. (2014). on Selected Methods for Evaluating, Journal Acta Universitatis Lodziensis, vol. 3, pp. 302.

**[Martínez, 2007]** Martínez, J. (2007). La bioinformática como herramienta para la investigación en salud humana. Salud Pública de México, pp. 64-66.

**[MathWorks, 2018]** MathWorks. (2018). Machine Learning. Retrieved from https://la.mathworks.com/discovery/machine-learning.html%0A

**[Medina, 2017]** Medina, R. F., (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. INTERFASES, 165-189.

**[Méndez, 2017]** Méndez, J., (2017). Prevalencia de tuberculosis latente en pacientes con diabetes mellitus en una institución hospitalaria en la ciudad de Prevalence of latent tuberculosis in patients with diabetes mellitus at a hospital in the city of Bogotá, Colombia, pp.165–171.

**[Navin, 2016]** Navin M. (2016). Performance Analysis of Neural Networks and Support Vector Machines using Confusion Matrix. International Journal of Advanced Research in Science, Engineering and Technology, pp. 2106-2109.

**[NIH, 2019]** National Human Genome Research Institute (2019), Fecha de consulta: marzo 2019, sitio web: https://www.genome.gov/glossarys/index.cfm?id=90.

**[NIH, 2019]** Instituto Nacional de Cáncer, polimorfismo de un solo nucleótido, sitio web: https://www.cancer.gov/espanol/publicaciones/diccionario/def/polimorfismo-de-un-solo-nucleotido, fecha de consulta marzo 2019.

**[Oryzon, 2019]** Oryzon G. 2019, Epigenética: Concepto y desarrollo, sitio web: https://www.oryzon.com/es/epigen%C3%A9tica, fecha de consulta marzo 2019.

**[Pan, 2017]** Pan, L. (2017). Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Scientific Reports | 7: 7402 | DOI:10.1038/s41598-017-07408-0.

**[Pérez, 2007]** Pérez, J. (2007). Data Mining System Oriented to Population Databases for Cancer. III Workshop Em Algoritmos e Aplicações de Mineração de Dados (WAAMD) 2007, pp. 101–104.

**[Pérez, 2016]** Pérez, Eduardo. (2016). Minería de Datos Orientada al Big Data en el Área de Salud, Tesis de Maestría, CENIDET.

**[Pérez, 2015]** Pérez, J., Iturbide, E., Olivares, V., Hidalgo, M., Martínez, A., & Almanza, N. (2015). A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases. J Med Syst, vol. 39. https://doi.org/10.1007/s10916-015-0312-5

**[Reyes, 2017]** Reyes, R. (2017), R, el lenguaje de la ciencia de datos – SeaCCNA al Día, Sea CCNA, http://www.seaccna.com/r-lenguaje-la-ciencia-datos/, fecha de consulta abril 2019.

**[Rosales, 2014]** Rosales, U. D. D., (2014). Aspectos biológicos y clínicos para comprender mejor al osteosarcoma. Investigacion En Discapacidad, vol. 3, pp. 33–40.

**[Salem, 2017]** Salem, H., Attiya, G., & El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. Applied Soft Computing Journal, vol. 50, pp. 124–134. https://doi.org/10.1016/j.asoc.2016.11.026

**[Serrano, 2017]** Serrano, P. L. (2017). Desarrollo de un recomendador de productos basado en en Extreme Gradient Boosting. España: Universitat Rovira Virrgili.

**[Shaltout, 2014]** Shaltout, N. A. (2014). Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts. World Congress on Engineering. Londres: WCE.

**[Sherry, 2001]** Sherry S, W. M. (2001). dbSNP: the NCBI database of genetic variation. NCBI, pp. 308-311.

**[Spurek, 2018]** Spurek, P. (2018). Fast independent component analysis algorithm with a simple closed-form solution. ELSEVIER, pp. 26-34.

**[Tang, 2013]** Tang J, Y. H. (2013). Histone deacetylases as targets for treatment of multiple diseases. NCBI, pp. 651-662.

**[Tarek, 2017]** Tarek, S. (2017). Gene expression based cancer classification. Egyptian Informatics Journal. pp. 151-159

**[Tidyr, 2018]** Tidyr (2018), librería de R, https://cran.r-project.org/web/packages/tidyr/tidyr.pdf, fecha de consulta octubre 2018

**[Tolosa, 2017]** Tolosa, A. (2017), CROMOSOMAS: Qué son los cromosomas y por qué son importantes, GENÉTICA MÉDICA BLOG, consulta de sitio web: https://revistageneticamedica.com/blog/cromosomas/, fecha de consulta marzo 2019.

**[Turki, 2018]** Turki (2018). An Empirical Study of Machine Learning Algorithms for Cancer Identification. IEEE.

**[Urbanowicz, 2017]** Urbanowicz, R. J. (2017). Relief-Based Feature Selection: Introduction and Review. Journal of Machine Learning Research, pp. 1-18.

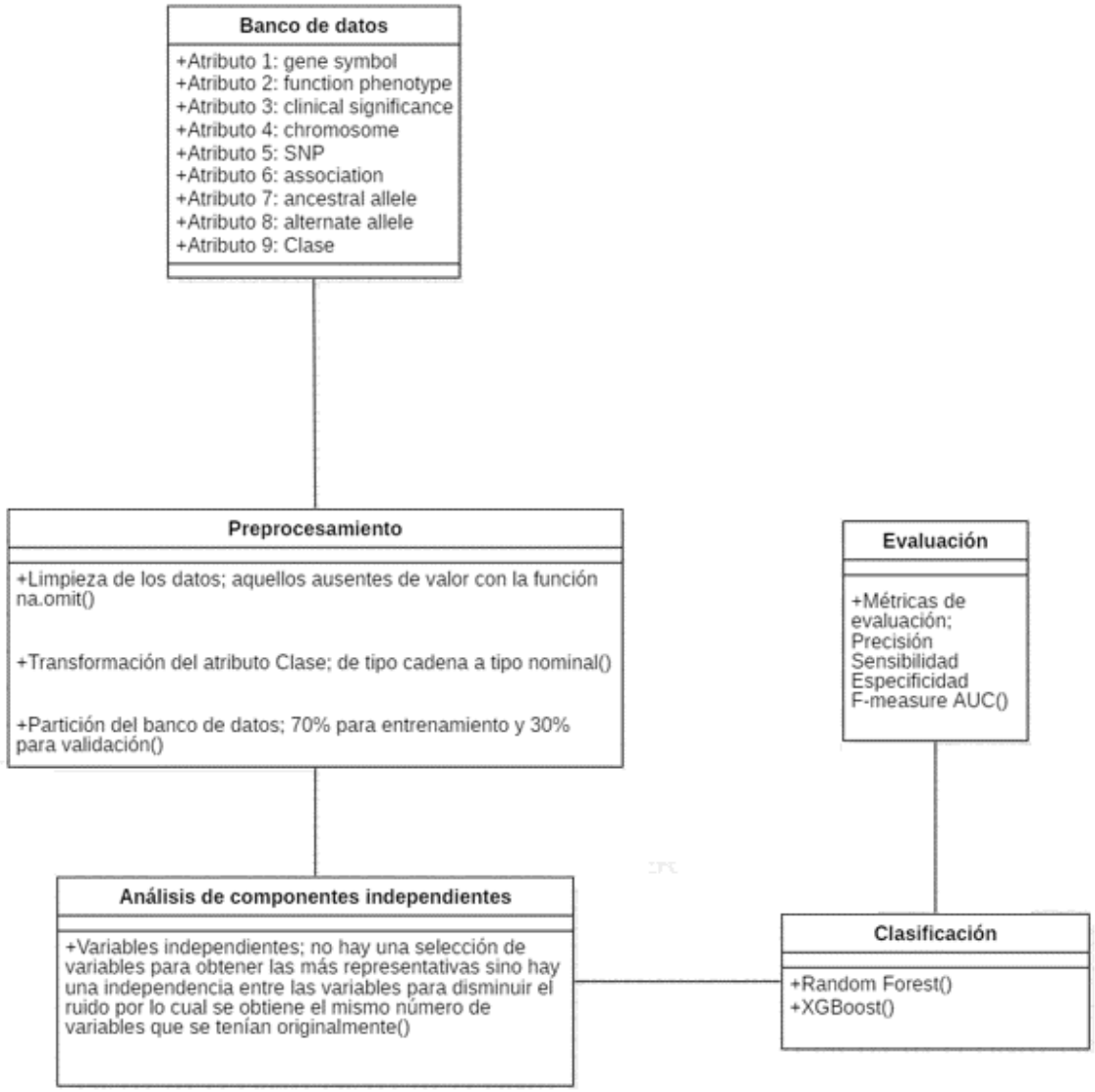
**[Visa, 2011]** Visa S., (2011). Confusion Matrix-Based Feature Selection. Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011 (págs. 120-127). The College of Wooster.

**[Wang, 2009]** Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics.

**[Wyner, 2017]** Wyner, A. J., (2017). Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. Journal of machine learning research, pp. 1-33.

# ANEXOS

## Anexo A. Diagrama de clases de la metodología de solución



## Anexo B. Pseudocódigo de la Paralelización

```
# Integrar la librería paralela
   library(doParallel)
   library(iterators)
# Calcular el número de núcleos
   número de núcleos <- detección de los núcleos
# Crear un cluster
   Cluster (cl) <- hacer cluster(número de núcleos)
# Exportar objetos a paralelizar
   Exportar cluster(cl)
#Entrenamiento
   iniciar temporizador
   entrenamiento del clasificador <- algoritmo (Base de datos, Clase, Parámetros)
   Iterar el entrenamiento en cl
   fin temporizador
   predicción
   matriz de confusión
   métricas de evaluación

# Detener el cluster
   detener Cluster(cluster)
```

## Anexo C. Productos académicos.

## Reconocimiento en ponencia UTEZ

UNIVERSIDAD TECNOLÓGICA
EMILIANO ZAPATA DEL ESTADO DE MORELOS
ORGANISMO PUBLICO DESCENTRALIZADO DEL GOBIERNO DEL ESTADO DE MORELOS

otorga el presente

# RECONOCIMIENTO

Al: ING. CARLOS ALBERTO MONCADA VÁZQUEZ

Por su participación con la ponencia:
**"Aprendizaje automático"**
llevada a cabo en las instalaciones de esta
Universidad Tecnológica.

Emiliano Zapata, Mor., julio de 2018

**Lic. Mireya Espinoza Avilés**
Jefa del Departamento de Tutorías y
Apoyo Psicopedagógico.

## Reconocimiento en escuela de IA UTEZ

UNIVERSIDAD TECNOLÓGICA
EMILIANO ZAPATA DEL ESTADO DE MORELOS
ORGANISMO PÚBLICO DESCENTRALIZADO DEL GOBIERNO DEL ESTADO DE MORELOS

otorga el presente

# RECONOCIMIENTO

A: Carlos Alberto Moncada Vázquez

Por su participación como ponente de la conferencia
**"Análisis de datos para osteosarcoma"**
en el marco del evento: Escuela de Inteligencia Artificial y Robótica 2018,
llevado a cabo en las instalaciones de esta Universidad Tecnológica,
del 25 al 27 de octubre del presente año.

Emiliano Zapata, Mor, octubre de 2018

UTEZ
DIVISIÓN ACADÉMICA DE MECÁNICA INDUSTRIAL

**M. en C. Jaime Vázquez Colín**
Director de la División Académica de
Mecánica Industrial

## Reconocimiento asistente ICMEAE

## Reconocimiento de concursante ICMEAE

## Reconocimiento de colaboración XXXII Congreso Nacional de Bioquímica

SOCIEDAD MEXICANA DE BIOQUIMICA A.C.

Otorga la presente

CONSTANCIA a:

Quien asistió y presentó el trabajo:

**Identification and Classification of Single Nucleotide Polymorphisms as Biomarkers associated with Dementia through Data Mining and Machine Learning**

Por:
Erick Cuevas Fernández, Carlos Moncada Vázquez, Joel Sánchez Mendez, Montserrat Loredo Guillen, Heriberto Manuel Rivera

En la modalidad de presentación oral durante el
XXXII Congreso Nacional de Bioquímica
4 - 9 de noviembre de 2018, Ixtapa Zihuatanejo, Gro.

Atentamente

**Dra. Irene B. Castaño Navarro**
Presidente

# Anexo D. Carta de liberación de la estancia en la UAEM

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

FACULTAD DE NUTRICIÓN

FACULTAD DE NUTRICIÓN

IP Dr. Heriberto Manuel Rivera. Laboratorio de Biología de Sistemas y Medicina Traslacional
Calle Iztaccihuatl numero 100 Col. Volcanes. Cuernavaca Morelos, C.P. 62350
Tel.: (777) 329 7000
Ext Lab 3500; 3493
m2mrivera@uaem.mx
man010ramaster@gmail.com

Cuernavaca, Mor; a 7 de Junio del 2019

**Dr. Gerardo Reyes**
**Profesor Investigador del CENIDET**
**PRESENTE**

Estimado Dr. Reyes, el motivo del presente es para hacer de su conocimiento que el **C. Carlos Moncada**, estudiante del Centro Nacional de Investigación y Desarrollo Tecnológico, realizó su estancia de investigación del **4 de Marzo al 5 de Abril del 2019** en el laboratorio de Biología de Sistemas y Medicina Translacional bajo mi supervisión.

Con base en lo anterior, hago constar que Carlos Moncada concluyó más que satisfactoriamente su plan de trabajo. Mostrando siempre gran interés y entusiasmo. En definida un elemento excelente.

Sin más por el momento, aprovecho la oportunidad para enviarle un cordial saludo.

Atentamente

PITC. Dr. Heriberto Manuel Rivera
Laboratorio de Biología de Sistemas y Medicina Translacional
Facultad de Nutrición-UAEM

*"2019 a 100 años del asesinato del General Emiliano Zapata Salazar"*

UAEM

*Una universidad de excelencia*

RECTORÍA 2017-2023