# Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

## Departamento de Matemáticas y Física
## Maestría en Ciencia de Datos


# Detección Temprana de Necesidad de Tratamiento de Salud Mental con el Aprendizaje Automático para Trabajadores del Sector de Tecnologías de Información

**Tesis** para obtener el **GRADO** de
**MAESTRO EN CIENCIA DE DATOS**

Presentada por:
**Alejandra Paola Galindo Hernández**

Director de Tesis:
**Dra. Diana Paola Montoya Escobar**

Tlaquepaque, Jalisco, Diciembre 2022

**Instituto Tecnológico y de Estudios Superiores de Occidente**
Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

**Formulario de aprobación del Maestría en Ciencia de Datos**

*Título de Tesis:* **Detección Temprana de Necesidad de Tratamiento de Salud Mental con el Aprendizaje Automático para Trabajadores del Sector de Tecnologías de Información**

*Autor:* **Alejandra Paola Galindo Hernández**

Tesis Aprobada para completar todos los requisitos de grado para la Maestría en Ciencias de Datos.

_____
Directora de Tesis, Dr. Diana Paola Montoya Escobar

_____
Sinodal, Dr. Rocio Carrasco Navarro

_____
Sinodal, Dr. Fernando Becerra Lopez

Tlaquepaque, Jalisco, Diciembre 2022

# Agradecimientos

Primeramente, me gustaría agradecer a la Dra. Diana Paola Montoya Escobar, directora de Tesis, por su gran apoyo, grandes ideas, asesorías y guía durante todo este proceso de mi Tesis, a todos mis maestros por darme las herramientas y conocimientos necesarios y a la Universidad Jesuita de Guadalajara, ITESO.

Me gustaría agradecer a mi familia a mis papás Pedro Galindo y Verónica Hernández y a mis hermanas Daniela Galindo y Denisse Galindo porque son una gran fuente de inspiración para lograr mis metas, por su gran apoyo en los momentos que más necesitaba, su paciencia, grandes y sabios consejos, ¡¡Por sus desvelos y preocupación Gracias!!

Si bien no ha sido fácil, ni mucho menos, durante este proceso hubo momentos difíciles, estuve a punto de rendirme, pero gracias a toda su ayuda y apoyo me dieron la confianza de salir adelante, luchar contra mis miedos y lo más importante a confiar en mí.

¡SE LOGRO! Un sueño y un objetivo más cumplido en mi vida.

# Resumen

El problema de salud mental ha existido desde que existe el ser humano, pero con el paso del tiempo y con el avance de la tecnología se han identificado ciertos patrones que tienen las personas más propensas a sufrir de este padecimiento. Actualmente, debido a las presiones que se tienen en el lugar de trabajo, este ha sido un factor que ha incrementado este problema y ha tomado interés para los profesionales de salud mental y las personas de recursos humanos de las empresas en su detección temprana.

Dado toda la problemática del COVID-19, las recientes crisis económicas y problemas a nivel mundial, como la guerra de Ucrania, las personas han tendido a ser más propensas a tener problemas de salud mental por temas relacionados en su lugar de trabajo. En los años recientes, se han llevado a cabo diferentes investigaciones en el tema, buscando predecir a tiempo las personas que pudieran llegar a requerir un tratamiento de salud mental.

Este trabajo, se ha enfocado realizar un modelo de clasificación con modelos de aprendizaje automático (machine learning), para predecir la probabilidad de que una persona de la industria de tecnologías de la información requiera un tratamiento de salud mental. Este se resuelve con características de las personas tomadas por medio de una encuesta laboral, con preguntas sobre el lugar de trabajo o en su vida diaria. Otros autores han realizado investigaciones con este mismo conjunto de datos, buscando el mismo objetivo. En trabajos previos sólo se determinó si la persona necesitó o no un tratamiento por medio de modelos de clasificación con machine learning, mientras, en este trabajo de tesis, adicionalmente se da a conocer las características más relevantes por percentiles de probabilidad para dar mayor idea al personal de salud de cuando pueda requerir una persona tratamiento sin necesidad de la utilización del modelo.

En este trabajo, se utilizaron cinco diferentes modelos de clasificación de Machine Learning: Random Forest, Gradient Boosting Machine, Generalized Linear Model, Support Vector Classifier y Naïve Bayes para predecir la probabilidad de que un trabajador del área de tecnologías de la información requiera un tratamiento de salud mental. Además, obtener las principales características por score de probabilidad de las personas más propensas a recibir un tratamiento. Para cada uno de los modelos, se realizó optimización de hiperparámetros para encontrar la mejor solución de cada modelo. Finalmente, se realizó una comparación entre los modelos evaluados y se encontró el mejor modelo dio resultados con ROC AUC de 73.9% y se compararon con los resultados de trabajos previos realizados por otros investigadores; adicionalmente, se evaluó en texto cómo estaba ordenando los percentiles de probabilidad para determinar si el modelo podría dar buenos resultados en producción.

# Tabla de Contenido

# Índice de Figuras

# Índice de Tablas

# Capítulo 1

## 1.1. Introducción

Tener un trastorno de Salud Mental es una de las enfermedades más críticas e importantes hoy en día que afectan la calidad de vida de las personas. Algunos de los factores más comunes que afectan el estilo de vida y los cuales pueden desarrollar un trastorno de salud mental son el estrés, depresión y la ansiedad. Querer pertenecer a un grupo social, las largas horas de trabajo, guerras y crisis económicas han hecho que las personas tiendan a desarrollar mucho más fácil un trastorno de salud mental.

En diciembre 2020, se detectó el primer caso de COVID-19 en el mundo, dando un efecto negativo en las personas y un exceso en las defunciones. Con ello vino una crisis de incertidumbre que generó desempleo y pobreza, conduciendo a problemas de salud como enfermedades del corazón, ansiedad, estrés y depresión, y con ello, desarrollar algún trastorno mental o/y neurológico.

Una de las consecuencias provocadas por la pandemia fue un alto porcentaje en el desempleo, dejando a tan solo un 60% de la población mundial con trabajo, el cual como trabajador es importante el apoyo y los diferentes beneficios que una empresa pueda brindar y así poder disminuir el estrés laboral; las largas horas de trabajo y las jornadas nocturnas hacen que los trabajadores aumenten sus expectativas laborales, ya que esto puede ocasionar depresión y estrés, y en algunos casos desarrollar algún trastorno de salud mental.

Anteriormente, se han realizado esfuerzos para la detección temprana de que las personas puedan necesitar un tratamiento de salud mental. En 2014 y 2016, se realizaron algunas encuestas [1] para medir las actitudes hacia la salud mental y la frecuencia de trastornos mentales en el área de trabajo en el sector de tecnologías de la información (TI) [1]. Con esta fuente de datos y complementando con otras variables y características, se identificaron algunas hipótesis que podrían ser relevantes para este caso, en este trabajo de tesis se aplicaron cinco modelos de machine learning de clasificación. Esto con el fin de obtener las probabilidades de que una persona pueda requerir un tratamiento de salud mental y analizar las principales características de los scores más altos para así, poder ayudar al área de salud y a las empresas de TI, a detectar de forma temprana, si un trabajador puede ser candidato requerir un tratamiento de salud mental y ayudar a controlarlo a tiempo.

## 1.2. Antecedentes

Más de 450 millones de personas en todo el mundo podrían haber desarrollado un problema de salud mental que dificulta gravemente su estilo vida [2]. El 60% de esas personas que han tenido este tipo de problemas tienen un empleo, y de ellos, entre el 35% al 50 % no reciben ningún tratamiento de salud mental o no el adecuado. Esto ha ocasionado que varias organizaciones y universidades hayan creado varios apoyos y propuestas para generar un ambiente de confianza y lealtad entre los trabajadores y las empresas de TI, no solo para mejorar el rendimiento laboral, sino como apoyo externo debido a los diferentes acontecimientos mundiales en los que nos hemos enfrentado en los últimos años (pandemia COVID-19, recesión económica, guerra de Ucrania-Rusia, etc.) [3].

En recientes años, algunos autores han desarrollado diferentes modelos para detección temprana de problemas en el área de salud mental e incluso para determinar si requiere o no un tratamiento. Por ejemplo, en 2017, Unilever [3], lanzó un conjunto de apoyo para empleadores, llamado *"Cuestiones de salud mental"*, que brinda información y consejos prácticos para ayudar a las empresas a crear entornos laborales de apoyo donde los empleados se sientan cómodos pidiendo ayuda y discutiendo problemas de salud mental sin temor a la estigmatización o la discriminación. El conjunto de apoyo incluye consejos sobre la creación de una cultura que apoye la salud mental, la gestión del estigma, el trabajo con proveedores de servicios y la implementación de enfoques de mejores prácticas.

Así también, Unilever [4], publicó un informe llamado *"Salud mental en el lugar de trabajo"* que encuestó a 100,000 empleados en más de 70 países y descubrió que más de la mitad de los empleados habían experimentado algún tipo de problema de salud mental en el último año, dando como resultado unas de las razones más comunes por las que las personas pueden ausentarse y tener poco rendimiento en el trabajo. Por lo tanto, si los empleados se toman días libres de salud mental para quedarse en casa, descansar y recargar energías después de un tiempo bastante complicado y desafiante en el trabajo, tendría como beneficio no solo al trabajador sino también a las empresas un mejor rendimiento y un alta en la productividad [5].

Por otro lado, en 2018, Unilever India [6], se asoció con la aplicación de salud mental *"Happiness@Work"* para brindar asesoramiento y capacitación física a los trabajadores que sufren de depresión o ansiedad. La empresa también ofreció a sus empleados una alta variedad de clases de asesoramiento, apoyo, yoga y meditación. Ofrecer este tipo de talleres de tranquilidad y bienestar mental teniendo como objetivo mejorar y mantener una relación estable entre los trabajadores.

En México se ha identificado esta problemática, pero han sido pocos los estudios de investigación dado que no se ha recolectado los datos suficientes para realizar un análisis detallado de esto. El objetivo de este trabajo de tesis es entrenar un modelo con datos generales a nivel mundial y luego poder hacer un backtest con datos de México para poder predecir de forma temprana las personas más probables a requerir un tratamiento.

## 1.3. Trabajos Relacionados

En los últimos años se han realizado diferentes estudios relacionados a este trabajo de tesis utilizando técnicas de aprendizaje automático (machine learning). En los trabajos que se citan a continuación, todos utilizaron la fuente de datos k Mental Health in Tech Survey [1], que fue el conjunto de datos principal usado en este trabajo de tesis. Para todos los casos, se utilizó, como variable de salida, una variable binaria de si la persona necesita o no tratamiento. Cabe resaltar que todos los trabajos resuelven al final un problema de clasificación, donde por medio de un punto de corte (que por lo regular es 0.5) se identifica si la persona tuvo o no tratamiento para salud mental. A continuación, se describe brevemente el desarrollo de cada una de estas investigaciones.

Shrijoy Chowdhury, et al [7], utilizó doce modelos para predecir si un empleado necesita tratamiento médico o no. El mejor modelo fue SVM obteniendo un *F1_Score* de 88%, que es la media armónica entre precisión y recuperación, un *Accuracy* de 86% y un *ROC AUC* (área bajo la curva) para ver qué modelo funciona mejor entre otros modelos y para el modelo SVM fue de 85% que es el más alto entre todos los modelos.

Por otro lado, Pavan Yeluri, et al [8], utilizó cuatro modelos de machine learning para predecir si los antecedentes familiares y los beneficios de una empresa influyen en los empleados que desean recibir un tratamiento, así también se buscó identificar si los problemas de salud mental del empleado interfieren con el trabajo. Comparando los cuatro modelos, se obtuvo un ROC AUC de 86% en el modelo de XGBoost el cual fue seleccionado como el mejor modelo.

En contra parte, Megan Risdal, et al [9], utilizó ocho métodos de Machine Learning diferentes de los cuales obtuvo el 80% de accuracy en casi todos los modelos, pero el mejor modelo fue Boosting con un 81% el cual se seleccionó para después aplicarle los datos de prueba y así predecir si el paciente deber de ser tratado por su enfermedad mental o no.

Por lo contrario, Aditi Mulye, et al [10], siete modelos de clasificación fueron utilizados para entender, predecir y analizar los diferentes factores de que una persona pueda o no recibir un tratamiento de salud mental de manera sistemática. Comparando los demás modelos el Gradient Boosting Classifier tuvo una precisión del accuracy del 81%

En la Tabla 1 se presenta un resumen de los trabajos previos mencionados anteriormente.

Tabla 1. Trabajos relacionados

| **Técnicas** | **Objetivo** | **Autor** |
|---|---|---|
| Support Vector Machine<br>Decision Tree | Predecir si un empleado necesita un tratamiento médico o no. | Shrijoy Chowdhury [2021] |
| SVM<br>Random Forest<br>Gradient Boosting<br>XGBoost | Identificar las características clave que conducen a problemas de salud mental en el espacio tecnológico | Pavan Yeluri [2022] |
| Logistic Regression<br>KNeighbors Classifier<br>Decision Tree Classifier<br>Random Forests<br>Bagging<br>Boosting<br>Stacking | Predecir si un paciente debe ser tratado por su enfermedad mental o no. | Megan Risdal [2018] |
| Logistic Regression<br>KNeighbors Classifier<br>Decision Tree Classifier<br>Random Forests Classifier<br>Gradient Boosting Classifier<br>AdaBoost Classifier<br>XGB Classifier | En este núcleo, intentaremos comprender qué factores contribuyen a la salud mental de una persona de manera sistemática. | Aditi Mulye [2021] |

## 1.4. Justificación

Al no tener conciencia de cuidar la salud mental en la vida diaria, las tomas de decisiones podrían verse afectadas, lo que conllevaría a no tener una buena calidad de vida. La mayoría de la población mundial tiene un trabajo, el cual se necesita para sobrevivir, pero una alta demanda, largas horas laborales y realizar algún trabajo extra pueden generar estrés, ansiedad y depresión. Por ende, uno de los retos es entender e identificar los factores y características que aumenten las posibilidades de poder desarrollar un trastorno de salud mental dentro de los trabajadores de TI el cual requiere un tratamiento.

Tener una jornada laboral de al menos ocho horas al día, puede afectar la salud mental ya que los trabajares a lo largo de los años han tenido una alta demanda y cambios continuos en los objetivos y actividades diarias que enfrentan a diferentes tipos de retos y problemas a resolver en los cuales pudieran afectar la salud [11].

Algunas iniciativas en temas de salud mental en el lugar de trabajo incluyen: alentar a los empleados a que hablen sobre sus problemas de salud mental, brindar capacitación a los gerentes sobre cómo apoyar al personal con problemas de salud mental y proporcionar recursos para ayudar

a los empleados a manejar el estrés [12]. Los empleadores también pueden apoyar la salud mental y el bienestar brindando acceso a licencia por enfermedad y tiempo de vacaciones para los empleados que lo necesitan para atender un problema de salud mental.

Para este trabajo se aplicarán cinco modelos de machine learning de clasificación con los cuales se puede conocer, entender e identificar la probabilidad y las características de conocer con anticipación si un trabajador del área de TI pueda o no desarrollar algún trastorno mental y poder recibir los medicamentos apropiados para ayudarlo.

## 1.5. Planteamiento del Problema

Vivir con un trastorno de salud mental hoy en día es más común de lo que se cree. La alta demanda en el trabajo, guerras, crisis económicas, enfermedades, COVID-19, el querer pertenecer a un grupo social, cumplir sueños, metas u objetivos, puede generar estrés y ansiedad el cual puede causar alguna enfermedad mental.

Una de las consecuencias provocadas por la pandemia y las diferentes razones previamente mencionadas dejó a un 60% de la población mundial con trabajo el cual se necesita para sobrevivir, mantener un buen equilibrio de la salud mental de los trabajadores es muy importante. Por lo tanto, en este trabajo se implementaron modelos de machine learning de clasificación, los cuales permitirán a identificar las características y la probabilidad de que un trabajador pueda desarrollar algún trastorno de salud mental.

En este trabajo de tesis, se busca encontrar, por medio de las características de una encuesta a empleados de la industria de TI, cuál es la probabilidad de requerir un tratamiento de salud mental y poder así, ayudar a las personas más propensas a tener un problema de salud mental de forma temprana. Asimismo, por medio de las diferentes variables de la encuesta, poder identificar cuáles son las principales características de las personas sean que probables a requerir un tratamiento, para con ello, sin necesidad de tener o utilizar el modelo, poder encontrar patrones o identificar ciertas condiciones que tienen las personas y puedan ser aplicado por personas del sector salud, sin necesidad de requerir un modelo en producción.

## 1.6. Objetivos

### 1.6.1 General

Predecir de forma temprana la necesidad de un tratamiento de salud mental de una persona trabajadora de la industria de TI, encontrando un score de la probabilidad, por medio de modelos de clasificación con machine learning.

### 1.6.2 Específicos

- Obtener el mejor modelo de machine learning para predecir la probabilidad de requerir un tratamiento de salud mental para un trabajador de TI.

- Identificar las características más importantes para determinar qué tan probable es que una persona de esta industria requiera o no un tratamiento.

- Analizar las principales características demográficas que tienen las personas con mayor probabilidad de requerir un tratamiento de salud mental.

- Tener un mapa por score de probabilidad que ayude a las personas de salud a identificar las principales características de las personas más probables a requerir un tratamiento de salud mental.

# Capítulo 2

## 2.1. Introducción

Para la solución del problema que se presenta en este trabajo, se consideraron cinco tipos diferentes de algoritmos de clasificación: Random Forest, Gradient Boosting Machine (GBM), Generalized Linear Models (GLM), Support Vector Classifier (SVC), y Naïve Bayes. Así también, la Optimización Bayesiana y Grid Search fueron utilizadados para optimizar los Hiperparámetros de cada modelo. A continuación, se describen estos modelos para entender su funcionamiento y sus fundamentos matemáticos, que llevaron la resolución del problema planteado en este trabajo de tesis.

## 2.2. Modelo 1 – Random Forest

El Modelo Random Forest está formado por un conjunto (ensamble) de árboles de decisión individuales, cada uno entrenado de manera distinta de los datos de entrenamiento dichas predicciones de cada observación se obtienen las predicciones de todos los árboles individuales que forman un único modelo.

En la década de 1960, se empezaron a utilizar los árboles de decisión para la toma de decisiones en problemas clasificación haciendo las variables de salida categóricas o binarias, gracias a su gran facilidad de entrenar el modelo de forma binaria nos permite comprender e interpretar de forma muy entendible los árboles de decisión [13].

Estos árboles tienen una estructura formada por un nodo inicial, ramas, nodos finales y nodos internos, los cuales se explicarán a continuación [14]:

- **Nodos internos o de decisión:** representan cada una de las características, factores o propiedades y muestra una decisión que el algoritmo tomará.
- **Ramas:** representan las líneas entre los nodos de las cuales indican la decisión en función de una determinada condición o un posible resultado o acción dentro del árbol.
- **Nodos de probabilidad:** representa las probabilidades de los resultados.
- **Nodos finales:** representan el resultado del árbol.

Este algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta [15].

Algunas ventajas de los árboles de decisión son:

- Son fáciles de entrenar
- Analiza todas las posibles soluciones y consecuencias de tomar alguna decisión.
- Ayuda a encontrar las mejores soluciones
- Facilidad de la interpretación
- Identifica las principales variables más importantes
- Se puede graficar

Las principales desventajas de utilizar estos modelos son:

- Estos se pueden volver muy complejos a la hora de predecir.
- El sobreajuste (over-fitting en inglés) es una de las limitaciones que los árboles pueden generar, esto se debe a la alta correlación entre las variables de entrada y el tiempo del diseño [16].
- Alto costo computacional si la muestra es alta.
- En problemas más complejos, no son los mejores modelos para predecir.

Para poder crear un árbol de decisión es necesario comenzar con la decisión inicial, el cual será el nodo principal, en seguida, hacia la derecha se comienzan a crear las ramificaciones que representarían en el modelo posibles decisiones, cada nodo está asociado con un resultado o una clase. En cada ramificación es necesario agregar un nodo de probabilidad y decisión. En este trabajo de tesis se está generando un árbol binario, ya que la variable de salida binaria consiste en si la persona necesita o no tratamiento, para este caso cada nodo debe tener dos decisiones. Tal y como se muestra en la Figura 1 [14].

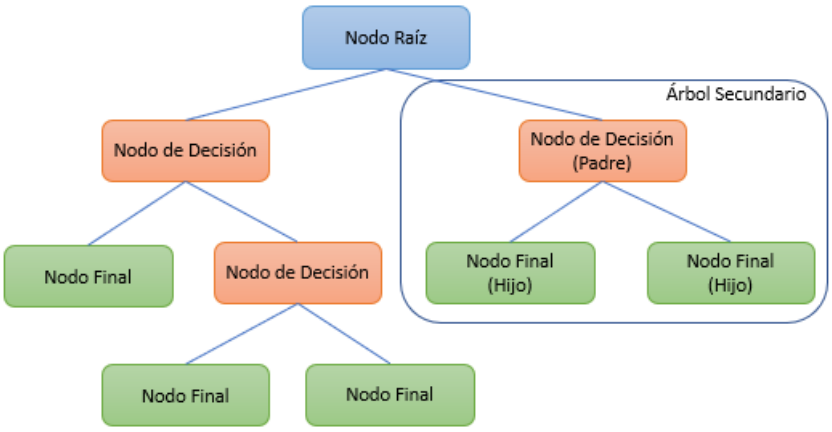


Figura 1. Ejemplo de Árbol Binario

Para el Modelo 1, se decidió utilizar el modelo Random Forest el cual fue desarrollado e implementado en el ambiente de R por Leo Breiman en 2001 Forest. Con el paso de los años este algoritmo de clasificación se ha vuelto el más popular y utilizado, junto con los modelos SVM y las redes neuronales [17].

Este algoritmo en contraste con los otros algoritmos de árboles utiliza los llamados límites de decisión aleatorios en lugar de límites de regresión o de hiperplanos: los límites entre clases no están determinados por ningún modelo anterior, sino que se seleccionan aleatoriamente de todos los límites de decisión posibles dentro del entrenamiento datos. La selección de estos límites de decisión aleatorios se realiza mediante un procedimiento de arranque, que se basa en un gran número de conjuntos de datos más pequeños, cada uno de tamaño igual al conjunto de datos de entrenamiento. Luego, cada uno de estos conjuntos de datos más pequeños se usa para entrenar un árbol de decisión en su respectivo conjunto de datos [17].

Como muchos de los modelos, los árboles de decisión también sufren con problemas de sesgo y varianza. Para poder generar un modelo óptimo es necesario mantener un balance entre estos dos errores. A este balance se le conoce como trade-off, el cual genera un equilibrio entre el error de sesgo y la varianza, utilizando ensambles es una forma de aplicar este trade-off.

Los dos tipos de ensamble más utilizados en este Algoritmo son:

**Bagging.** Se ajusta a múltiples modelos, cada uno con un subconjunto de los datos de entrenamiento y así poder evitar el sobreajuste. Los árboles se forman combinando atributos seleccionados aleatoriamente del conjunto de datos y realizando una selección de características en cada nivel del árbol. Las características seleccionadas se usan para dividir los datos en subconjuntos de igual tamaño, que luego se usan para entrenar nuevos árboles [18].

El algoritmo Random Forest selecciona aleatoriamente características en cada nodo del árbol, en contraste con otros algoritmos de aprendizaje de árboles de decisión. Por lo tanto, un random forest puede encontrar patrones que no son evidentes en otros algoritmos basados en árboles, por ejemplo, patrones ocultos en los datos para clasificar imágenes. Sin embargo, esto también dificulta la interpretación de los resultados obtenidos por el algoritmo. Las ventajas del algoritmo de random forest son que puede manejar datos complejos y es muy resistente a valores atípicos y datos faltantes [19].

**Boostrap.** Se ajustan secuencialmente múltiples modelos sencillos, llamados weak learners, de forma que cada modelo aprende de los errores del anterior. Random Forest utiliza este método para generar límites de decisión aleatorios [17].

Algunas ventajas y desventajas del Random Forest son:

**Ventajas**

Una ventaja que tiene este Modelo de Clasificación es predecir la importancia de los predictores que lo hace un algoritmo muy fuerte, no solo para predecir, sino también para la exploración y análisis de los datos. Dos de estas medidas son: importancia por permutación e impureza de nodos. Para este problema en particular se utilizó importancia por permutación.

- **Importancia por Permutación** [20] identifica la influencia que tiene cada predictor sobre una determinada métrica de evaluación del modelo, la cual es estimada por validación cruzada o out-of-bag error.

  El valor asociado con cada predictor se obtiene de como en la Ecuación (1):

  - Calcular el incremento en la métrica debido a la permutación del predictor $j$.

$$\%Incremento_j = \frac{error - j - error_0}{error_0} * 100 \quad (1)$$

*Fórmula por Joaquín Amat Rodrigo*

- **Incremento de la pureza de nodos** [20] cuantifica el incremento total en la pureza de los nodos debido a divisiones en las que participa el predictor (promedio de todos los árboles).

  La forma de calcularlo es la siguiente [20]: en cada división de los árboles, se registra el descenso conseguido en la medida empleada como criterio de división (índice Gini, MSE entropía, ...). Para cada uno de los predictores, se calcula el descenso medio conseguido en el conjunto de árboles que forman el ensemble. Cuanto mayor sea este valor medio, mayor la contribución del predictor en el modelo.

**Desventajas**

Por otro lado, este tipo de modelo tienen la tendencia de sobre-ajustar (overfit). Esto quiere decir que tienden a aprender muy bien los datos de entrenamiento, pero su generalización no es tan buena. Una forma de mejorar la generalización de los árboles de decisión es usar técnicas de regularización. Para mejorar mucho más la capacidad de generalización de los árboles de decisión, deberemos combinar varios árboles [20].

### 2.2.1 Hiperparámetros

Los hiperparámetros son usados para mejorar la capacidad predictiva o aumentar la velocidad de cálculo. Para este modelo Random Forest se utilizaron los siguientes hiperparámetros [21]:

- ***trees:*** El número total de árboles totales que se encuentran en un Random Forest o Boosted Ensamble. Lo cual vendría siendo igual al número de iteraciones de impulso.
- ***tree_depth*:** Profundidad máxima del árbol.
- ***min_n**:* El número mínimo de puntos de datos en un nodo que se requiere para que el nodo se divida más.

## 2.3. Modelo 2 - Gradient Boosting Machine (GBM)

Es una técnica que propone el uso de modelos "débiles" para resolver tanto problemas de clasificación como regresión, el cuál utiliza la técnica de ensamble usando un modelo bag. Múltiples modelos son credos en paralelo y todos son considerados para obtener una predicción final [23].

El GMB puede utilizar diferentes modelos, para este proyecto se utilizan los árboles de decisión, los cuales se crean de manera secuencial y sucesiva, es decir, cada árbol de decisión va minimizando el error y mejorando al árbol anterior [24].

El algoritmo utilizado para el modelo GBM se conoce como SAMME, el cual se explica en los siguientes pasos:

- **Paso 1:** Se asigna un peso a cada una de las observaciones del conjunto de datos $w = \frac{1}{m}$, donde m es el número total de observaciones del conjunto de datos utilizado.

- **Paso 2:** Se selecciona un subconjunto de los datos $\{X_t, Y_t\}$, donde el número de observaciones es $m_t$. Esta selección se realiza en base al peso asignado, mientras mayor sea el peso, la observación tendrá mayor posibilidad de ser elegida en el subconjunto.

- **Paso 3:** Basado en el subconjunto creado anteriormente, se construye un árbol de clasificación, el cual está limitado a tener profundidad 1.

- **Paso 4:** Se genera un criterio de error $E^t$ para el modelo, basado en el poder predictivo del árbol (Ecuación (2)).

$$E^t = \frac{\sum_{j=1}^{m_t} w_j \; I(\hat{y}_j^{(t)} \neq y_j)}{\sum_{j=1}^{m_t} w_j} \tag{2}$$

$$\text{, donde } I = \begin{cases} 0 & \hat{y}_j^{(t)} = y_j \\ 1 & \hat{y}_j^{(t)} \neq y_j \end{cases} \tag{3}$$

dando mayor peso a las observaciones que se clasificó de manera errónea.

- **Paso 5:** En base al error calculado se calcula $\alpha^t$, la cual se considera como una medida de confianza para el modelo de clasificación (Ecuación (4)).

$$\alpha^t = log\left(\frac{1 - E^t}{E^t}\right) + \log(K - 1) \tag{4}$$

donde $K$ es un hiperparámetro el cual deber ser mayor o igual a 1. Es decir, cuan se obtengan errores pequeños $E^t$ , el valor $\alpha^t$ será grande, y cuando se obtengan errores grandes $E^t$ el valor $\alpha^t$ será menor.

**Paso 6:** Se actualizan los pesos de las observaciones en el subconjunto elegido, en base a los errores generados por el modelo (Ecuación (5)).

$$w_i = w_i \, e^{\alpha^t \, I \, (\hat{y}_j^{(t)} \neq y_j)} \tag{5}$$

Es decir, si la clasificación asignada por el modelo es errónea, el peso de la observación incrementa, de manera contraria, si la clasificación es correcta el peso de la observación se mantiene, por lo tanto, las observaciones con una clasificación errónea tendrán mayor probabilidad de ser elegido en el siguiente modelo.

**Paso 7:** Una vez que los pesos son actualizados, se realiza una normalización de los pesos de todo el conjunto de datos (Ecuación (6)).

$$w = w - promedio(w) \tag{6}$$

**Paso 8:** Se regresa al paso 2. La iteración termina una vez que se han generados el total de modelos establecidos.

Una vez que se generen todos los modelos, la asignación de la clasificación se genera de la siguiente manera (Ecuación (7)).

$$\hat{y}_i = \max_c \left( \sum_{j=1}^{m_t} \alpha^t \, I \, (\hat{y}_j^{(t)} = c) \right) \tag{7}$$

$$\text{, donde } I = \begin{cases} 0 & \hat{y}_j^{(t)} \neq y_j \\ 1 & \hat{y}_j^{(t)} = y_j \end{cases} \text{, c es la clase posible de asignación.} \tag{8}$$

donde se consideran las $\alpha^t$ asignadas a cada observación en su modelo correspondiente, donde las alfas más grandes tienen mayor relevancia en la predicción.

### 2.3.1 Hiperparámetros

Uno de los aspectos más atractivos del Gradient Boosting Machine (GBM) es su alta flexibilidad para ajustar los parámetros [25]. Estos son:

- ***ntrees***: número de árboles que forman el ensemble.
- ***min_rows:*** número mínimo de observaciones que debe tener cada nodo.
- ***stopping_metric***: métrica empleada para cuantificar cuánto mejora el modelo.
- ***score_tree_interval:*** número de árboles tras los que se evalúa el modelo. Por ejemplo, si el valor es 10, entonces, tras cada 10 nuevos árboles que se añaden al modelo, se calcula la *stopping_metric.* Si bien la evaluación del modelo puede hacerse tras cada nuevo árbol que se añade, esto puede ralentizar el entrenamiento.
- ***stopping_tolerance:*** porcentaje mínimo de mejora entre dos mediciones consecutivas (*score_tree_interval*) por debajo del cual se considera que el modelo no ha mejorado.
- ***stopping_rounds:*** número de mediciones consecutivas en las que no se debe superar el *stopping_tolerance* para que el algoritmo se detenga.
- ***ignore_const_cols:*** eliminación de variables con varianza cero
- ***max_depth:*** número máximo de profundidad

## 2.4. Modelo 3 - Generalized Linear Models (GLM)

Los modelos **Generalized Linear Models (GLM)** [26] fueron propuestos por Ronald Fisher en la década de 1920 y fueron desarrollados más adelante en la década de 1950 por Arthur Samuel y James Steele. El GLM es una herramienta eficaz y adaptable para modelar relaciones complejas en sus datos. Por esta razón, se ha convertido en uno de los tipos de modelos más comunes utilizados en el análisis de datos. Así también, permite que el modelo lineal esté relacionado con la variable de respuesta a través de una función de enlace y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho.

Dado un conjunto de observaciones (Ecuación (9)) [27], la media μ de la variable respuesta Y se relaciona de forma lineal con la o las variables regresoras X la podemos representar así,

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (9)$$

O en notación matricial (incorporando $\beta_0$ en el vector $\beta$) [27] (Ecuación (10)).

$$\mu = X^T \beta \qquad (10)$$

Ajustar el modelo consiste en estimar los valores de los coeficientes de regresión $\hat{\beta}$ y la varianza $\hat{\sigma}^2$ que maximizan la verosimilitud, también conocida como likelihood, de los datos, es decir, los que dan lugar al modelo que con mayor probabilidad puede haber generado los datos observados [27].

Un del método empleado es el ajuste por mínimos cuadrados (*OLS*) (Ecuación (11)):

$$\hat{\beta} = argmin_{\beta_0,\beta} (Y - X^T \beta)^2 \quad (11)$$

$$\hat{\mu} = X^T \hat{\beta} \quad (12)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-p} = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu})^2}{n-p} \quad (13)$$

Para que esta aproximación sea válida, se necesita que el error se distribuya de forma normal y que la variable sea constante [27].

GLM está especificado por tres componentes, los cuales se pueden observar en la Figura 2.

1. **Aleatoria.** Distribución de la variable respuesta.
2. **Sistemática.** Función lineal de las variables explicativas.
3. **De Enlace**. Una combinación de los componentes Aleatoria y Sistemática.

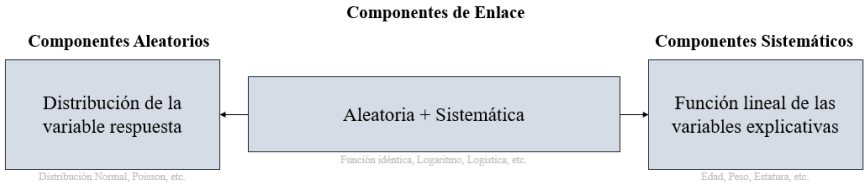


Figura 2. Tipos de componentes en un GLM

### 2.4.1 Hiperparámetros

Los Hiperparámetros utilizados en este Modelo son [25]:

- ***lambda_search:*** Búsqueda del valor óptimo de lambda.
- ***alpha:*** el valor de α que se emplea en el modelo para distribuir la penalización l1 y l2. Por defecto, alpha=0.5
- ***validation_frame:*** método de validación empleado para identificar el mejor modelo, puede ser mediante un único conjunto de validación o mediante validación cruzada (cross-validation).
- ***n_folds:*** número de particiones en la validación cruzada.
- ***lambda_min_ratio y nlambdas:*** determinan la secuencia de valores lambda que se generan automáticamente y sobre los que se realiza la búsqueda. El rango de valores va desde el valor mínimo con el que se alcanza una penalización total *lambda_max* (valor identificado automáticamente por H2O), hasta un valor de lambda igual a *lambda_min_ratio* x *lambda_max.* Dentro de ese rango, se generan un número de lambdas igual al valor especificado con *nlambdas*. Por defecto, se emplean *nlambda*=100 y *lambda_min_ratio* =1e−4.

## 2.5. Modelo 4 - Support Vector Classifier (SVC)

El Modelo de **Support Vector Classifier (SVC)** fue el único modelo de clasificación que necesitó transformación en los datos a comparación con el resto de los Modelos de Clasificación que se utilizaron en este trabajo de tesis. Siendo un modelo de clasificación en el entrenamiento, se va a maximizar el rendimiento y/o margen y minimiza el error.

El SVC es un tipo de aprendizaje automático que utiliza máquinas de vectores de soporte (SVM) como su clasificador [29]. Cada punto de datos en el conjunto de datos se representa como un vector de características, que se representa mediante una cuadrícula bidimensional. Esta cuadrícula contiene los valores de características de los puntos de datos en la cuadrícula y los valores de características se clasifican por sus valores. Los valores pueden ser binarios o de valor real, según la naturaleza de los datos que se utilicen para entrenar el clasificador [28].

Por lo tanto, para el modelo de SVC se decidió utilizar dos Kernels: RBF y Sigmoid para poder optimizar los hiperparámetros cuyo valor optimo se generó por validación cruzada.

- **Radial Basis Function Kernel (RBF)** [30] es uno de los Kernels más utilizamos ya que es muy similar a la distribución Gaussiana, calcula que tan cercano esta un punto con otro, por ejemplo: para dos puntos de X1 y X2 se puede expresar matemáticamente en la Ecuación (14) [30].

$$K(\mathrm{X}, \mathrm{X2}) = exp(-||X1-X2||22\sigma2) \quad (14)$$

  donde

  - $\sigma$ es la varianza y los hiperparámetros.
  - ||X1, X2|| es la Euclidiana (Norma-L2) distancia entre dos puntos

  El máximo valor que el Kernel RBF puede tomar es 1 y esto pasa cuando la distancia entre dos puntos es igual a cero [30].

  - Cuando los puntos son los mismos, no hay distancia entre ellos, y por esto son muy similares.
  - Cuando los puntos están separados por una distancia larga, el valor del Kernel es menor que 1 y cercano a 0, lo cual significa que los puntos son diferentes.

  Para utilizar este Kernel es importante encontrar el valor de $\sigma$ para identificar los puntos que son similares [28].

- **Sigmoid Kernel** proviene de Redes Neuronales, donde la función bipolar Sigmoide se llega utilizar como una función de activación para las neuronas artificiales [28,31] que se basa en la Ecuación (15)

$$K(\mathrm{X1}, \mathrm{X2}) = \tanh(\mathrm{X1TX2}+ \mathrm{coe}f0) \quad (15)$$

Este Kernel satisface y se relaciona con el teorema Mercer, y esto requiere que el Kernel sea positivo. Sin embargo, a pesar de un amplio uso, no es semidefinido positivo para ciertos valores de sus Hiperparámetros [31].

Los Hiperparámetros $\alpha$ y $c$ deben elegirse correctamente, de lo contrario, los resultados y el modelo puede fallar [31].

### 2.5.1 Hiperparámetros

Los Hiperparámetros utilizados en este Modelo son [31]:

- ***Gamma:*** nos facilita encontrar los subespacios que puedan diferenciar los puntos en el espacio y también nos permite añadir mayor complejidad a la hora de separar observaciones.
- ***C:*** controla el número y severidad de las violaciones del margen y del hiperplano que se toleran en el proceso de ajuste.

## 2.6. Modelo 5 - Naïve Bayes

Los modelos de **Naïve Bayes** es un algoritmo de aprendizaje automático de clasificación utilizando el teorema de Bayes, el cuál asume que los eventos que ocurren en un momento anterior tienen mayores probabilidades que los eventos posteriores, siendo capaz de clasificar datos, dado un conjunto de características, de forma probabilística. Por lo tanto, dado un conjunto de características de entrada, genera una distribución de probabilidad de salida que representa la probabilidad de cada una de las clases o categorías presentes. Luego, clasifica el conjunto de datos de entrada según la clase más probable según esta distribución [32].

**Naïve Bayes** se utiliza para problemas de clasificación. A diferencia de otros algoritmos de aprendizaje automático, el clasificador bayesiano no puede producir una sola estimación de probabilidad para cada clase como salida. En su lugar, produce un conjunto de probabilidades que representan la probabilidad de cada una de las clases [29].

**Teorema de Bayes:**

Este teorema fue desarrollado por Thomas Bayes, su principal objetivo es determinar la probabilidad de un evento simultaneo comparado con la probabilidad de otro evento similar [34].

El enfoque Bayesiano considera que los parámetros $\theta$ son los que se consideran aleatorios, y los datos $X$ están fijos como evidencia [35].

En ese sentido, nos interesa modelar la distribución de los parámetros dada la evidencia que se van observando en $P(\theta|X)$, lo cual, por el Teorema de Bayes podemos escribir como en la Ecuación (16) [35]:

$$P(X) = \frac{P(X|\theta)\, P(\theta)}{P(X)} \tag{16}$$

donde:

- $P(\theta|X)$ se conoce como distribución posterior. La posterior indica la probabilidad de los parámetros después de haber observado los datos.
- $P(X|\theta)$ se conoce como función de verosimilitud. La verosimilitud indica que tan bien los parámetros explican los datos.
- $P(\theta)$ se conoce como distribución previa. En la previa, se incluyen todo el conocimiento que se pueda tener acerca de los parámetros.
- $P(X)$ se conoce como distribución de evidencia.

**Tipos de Naïve Bayes Classifier**

- Multinomial Naïve Bayes
- Bernoulli Naïve Bayes
- Gaussian Naïve ve Bayes

**Interpretación de la Probabilidad**

Se analizan dos tipos de probabilidades: Enfoque Bayesiano y Enfoque Frecuentista, ya que ambos enfoques en muchas dimensiones y su aplicabilidad en el análisis de datos.

**Enfoque Bayesiano** [26]: Es la probabilidad relativa a la incertidumbre en un fenómeno. Es subjetiva y expresa el grado de confianza que tenemos en la ocurrencia de un evento. Así también, es la distribución de los parámetros según los datos, capturado por la función posterior, Ecuación (9).

Por ejemplo, desde un enfoque Bayesiano, para un clasificador con datos de entrenamiento ($X_{tr}$, $Y_{tr}$) y datos de predicción $X_p$ , se calcula como se muestra en la ecuación (17):

$$P((X_{tr}, y_{tr})) = \frac{P(y_{tr}|\, X_{tr}) \,, \theta)\, P(\theta)}{P(y_{tr}|\, X_{tr})} \tag{17}$$

y para la predicción se calcula de la siguiente manera: [26] Ecuación (18)

(18)

$$P(X_P, X_{tr}, y_{tr}) = \int P(X_P, X_{tr}, y_{tr})\, d\theta \qquad \text{(Marginalización)}$$

$$P(X_P, X_{tr}, y_{tr}) = \int P(\theta, X_P, X_{tr}, y_{tr}) P(X_P, X_{tr}, y_{tr}) d\theta \qquad \text{(Regla de la cadena)}$$

$$P(X_P, X_{tr}, y_{tr}) = \int P(\theta, X_P) P(X_{tr}, y_{tr}) d\theta \qquad (\theta \perp X_P), (y_P \perp X_{tr}, y_{tr}|\theta)$$

$$P(X_P, X_{tr}, y_{tr}) = E_{P(\theta|X_{tr},y_{tr})}[P(\theta, X_p)]$$

En este caso, la predicción es un promedio ponderado del del modelo por todos los posibles valores de $\theta$. Esto tiene un efecto de regularización, con lo que el clasificador es menos propenso a overfitting. Es decir, la inclusión del conocimiento de los parámetros en la previa $P(\theta)$ nos da un efecto regularizador.

Otra ventaja es que el enfoque Bayesiano nos permite hacer entrenamiento de esquemas online procesando sólo los nuevos datos, como se muestra en la Ecuación (19).

$$P_k(\theta) = P(\theta|x_k) \frac{(P(x_k|\theta)\, P_{k-1}(\theta)}{P(x_k)} \quad (19)$$

El entrenamiento conseguido por los nuevos datos $x_k$ entra como nuevo conocimiento al siguiente paso. Bajo un enfoque frecuentista, tendríamos que procesar todos los datos cada vez que haya nuevos datos.

**Enfoque Frecuentista** [26]: es la probabilidad relativa a la incertidumbre en un fenómeno es objetiva, y se refiere a la frecuencia relativa con la que ocurre un evento, así también maximiza la probabilidad ocurrencia de los datos capturada por la función de verosimilitud.

Principio de máxima verosimilitud (Ecuación (20)):

$$\hat{\theta} = arg\ max_{\theta}\ P(y_{tr}|\theta) \quad (20)$$

Por ejemplo, desde un enfoque frecuentista, para un clasificador con datos de entrenamiento $(X_{tr}, y_{tr})$ y datos de predicción $X_P$, maximizaríamos (Ecuación (21)):

$$\hat{\theta} = arg\ max_{\theta}\ P(y_{tr}|X_{tr},\theta) \quad (21)$$

De esta manera, la predicción sería (Ecuación (22)):

$$\hat{y}_P = \arg max_{yp}\ P(y_p|X_p, \hat{\theta}) \quad (22)$$

Se puede observar que la estimación de los parámetros $\hat{\theta}$ es incorrecta, esto quiere decir que hay overfitting a datos sesgados, por lo tanto, el clasificador tendría un mal comportamiento en la predicción.

## 2.7. Optimización de Hiperparámetros

Los Hiperparámetros son parámetros que afectan el performance/rendimiento del algoritmo ya que se eligen de acuerdo con los requisitos de un problema específico de aprendizaje automático y los recursos disponibles para entrenar un modelo con esos datos. Los Hiperparámetros a menudo se ajustan utilizando métodos de aprendizaje automático, como el descenso de gradiente [27].

Los Hiperparámetros juegan un papel importante en el proceso de aprendizaje automático, porque elegirlos puede generar diferencias dramáticas en la calidad del modelo entrenado. Muchos investigadores han investigado métodos para ajustar Hiperparámetros para diferentes tipos de problemas de aprendizaje automático.

A continuación, se hará una breve explicación de los tipos de Optimización de Hiperparámetros que se aplicaron en este problema.

**Optimización Bayesiana**

El Modelo Random Forest fue ejecutado en el ambiente RStudio fue optimizado con la optimización Bayesiana (tune_bayes), el cual se usa como aprendizaje automático para ajustar los Hiperparámetros del modelo dando un mejor rendimiento en un conjunto de datos de validación [32]. Este tipo de optimización es un enfoque que utiliza el teorema de Bayes para dirigir la búsqueda a fin de encontrar el mínimo o el máximo de una función objetivo.

Una de las ventajas para usar la optimización bayesiana es ajustar los Hiperparámetros aun así cuando la función objetivo sea costosa de evaluar.

**Random Discrete**

Los modelos de GBM fueron optimizados con Random Discrete.

El objetivo del Random Discrete es encontrar la disposición óptima de un conjunto de elementos para maximizar la suma de sus beneficios y minimizar el costo total de hacerlo.

Para este problema se desarrollaron dos modelos de GBM,

1. GBM - H2O VI. En el cual se utilizaron tres de los Hiperparámetros más importantes: *learn_rate*, *max_depth* y *sample_rate*
2. GBM - H2O VF. El cual fue desarrollado con los mejores Hiperparámetros obtenidos de modelo anterior (GBM - H2O VI) y los Hiperparámetros que no fueron definidos fueron optimizados por Random Discrete.

**Cross Validation Method**

En los modelos de SVM Classifier se utilizó Cross Validation Method para poder optimizar los Hiperparámetros, el cual se utiliza como un método de optimización para evaluar modelos o

estimar parámetros basados en una muestra con reemplazo en lugar de en todos los datos disponibles. Así también, es una técnica para evaluar modelos de ML entrenando varios modelos de ML en subconjuntos de los datos de entrada disponibles y evaluándolos en el subconjunto complementario de los datos. Utilicé la validación cruzada para detectar el sobreajuste, es decir, la falta de generalización de un patrón.

## 2.8. Métricas

Para los problemas de clasificación, en este trabajo se evalúan los modelos con diferentes métricas que ayudan a evaluar el performance de los modelos. A continuación, se describen cada una de ellas.

### 2.8.1 Matriz de Confusión

La Matriz de Confusión es una representación matricial de los resultados de las predicciones de cualquier prueba binaria, se utiliza para describir el rendimiento del modelo de sobre un conjunto de datos de prueba cuyos valores reales se conocen.

Cada predicción puede ser uno de cuatro resultados, basado en cómo coincide con el valor real:

1. ***Verdaderos Positivos:*** cuando la clase real del punto de datos era 1 y la predicha es también 1.
2. ***Verdaderos Negativos:*** cuando la clase real del punto de datos fue 0 y el pronosticado también es 0.
3. ***False Positives:*** cuando la clase real del punto de datos era 0 y el pronosticado es 1.
4. ***False Negativos:*** cuando la clase real del punto de datos era 1 y el valor predicho es 0.

| | | Valores Reales | |
|---|---|---|---|
| | | Positivos | Negativos |
| Valores Predicción | Positivos | Verdaderos Positivos | Falsos Positivos |
| | Negativos | Falsos Negativos | Verdaderos Negativos |

Figura 2. Matriz de Confusión

### 2.8.2 Accuracy

El Accuracy es la métrica de evaluación más simple en los problemas de clasificación. Esta métrica mide el porcentaje de predicciones correctas hechas realizadas por el modelo evaluando las muestras totales, como se señala en la ecuación (23). Las ventajas de utilizar la métrica de Accuracy

es que puede ser considerado como uno de los principales factores para determinar qué modelo es el mejor, además, funciona mucho mejor para datos que están correctamente balanceados.

$$Accuracy = \frac{Verdadero\ Positivo + Verdadero\ Negativo}{Total\ de\ Observaciones} \quad (23)$$

### 2.8.3 ROC-AUC

El **ROC**, también conocido como Receiver Operator Characteristic es una de las métricas de evaluación más importante de los algoritmos de clasificación. A comparación con el Accuracy, el ROC AUC determina un umbral bajo la curva el cual permite visualizar el equilibrio entre los datos verdaderos positivos (Sensitivity) y los falsos positivos (Specificity). Los cuales se explicarán más a fondo en las secciones 2.8.4 y 2.8.5.

Medir el área bajo la curva ROC es también un método muy útil para evaluar el rendimiento del modelo y determinar el umbral apropiado para maximizar la relación entre sensitivity y specificity, ya que representa el grado o medida de separabilidad entre clases. Cuanto mayor sea el AUC, mejor será el modelo para predecir las clases negativas como negativas y las clases positivas como positivas.

Para poder evaluar la probabilidad de un modelo que está clasificando perfectamente todas las muestras tendría un valor el área bajo la curva de 1. Por otro lado, un modelo completamente ineficaz tendría una curva ROC cercana a una línea diagonal de 45° grados, el cual el área bajo la curva sería de aproximadamente 0.5.

### 2.8.4 Sensitivity

La métrica de Sensitivity mide y evalúa la habilidad del modelo en predecir los *verdaderos positivos* de cada categoría los cuales son clasificados correctamente. Esta métrica se puede calcular como se presenta en la Ecuación (24):

$$Verdaderos\ Positivos = \frac{Verdaderos\ Positivos}{Falsos\ Negativos + Verdaderos\ Positivos} \quad (24)$$

### 2.8.5 Specificity

La métrica Specificity, al contrario de la métrica de Sensitivity, mide la precisión con que el modelo predice o mide los *falsos negativos*, esto quiere decir, que mide el porcentaje en donde el modelo

se equivocó al clasificar las categorías correctamente con respecto al modelo total, la cual se calcula con la Ecuación (25):

$$Falsos\ Positivos = \frac{Falsos\ Positivos}{Verdaderos\ Negativos + Falsos\ Positivos} \tag{25}$$

### 2.8.6 F1 Score

La puntuación F1, es una métrica que combina entre Precisión y Recall, y se obtiene un valor que puede servir como una forma de comparar el rendimiento de los modelos. Cuanto mayor sea la puntuación, mejor será el modelo para clasificar los datos. Cuando el modelo se prueba en un conjunto de datos diferente al que se entrenó, la puntuación F1 mide su rendimiento en estos nuevos datos [29].

Una puntuación F1 en el rango de 0.8 a 1 se considera aceptable para muchas aplicaciones, pero los modelos más avanzados pueden requerir puntuaciones en este rango o superiores [29].

# Capítulo 3

## 3.1. Introducción

En este capítulo se presenta el análisis de la base de datos, la cual está formado por dos diferentes tipos de conjuntos de datos, el primer conjunto es una encuesta llamada Survey on Mental Health in the Tech Workplace in 2014 [1] y el segundo conjunto de datos indica el Producto Interno Bruto (PIB) para los años 2014 y 2015 [38], ya que durante el análisis y exploración de los datos se pudo identificar una alta relación entre los países con un alto valor de Producto Interno Bruto (PIB) con probabilidad de que una persona pueda o no recibir un tratamiento de salud mental. Hay estudios que afirman que los países con mayor PIB son los países con más índice de suicidios, depresión y ansiedad ya que la exigencia es más grande que en países con un PIB medio o menor. En las siguientes secciones se hablará más sobre este tema.

## 3.2. Descripción de la Metodología en el Modelo

Para realizar este trabajo de tesis, se siguió la metodología de ciencia de datos propuesta por Microsoft [37]. Para esta, se procede a realizar un análisis y exploración de los datos, y con ello, entender las características de nuestras variables, el tipo de dato, valores faltantes, valores atípicos, identificar si tienen sesgo o si se necesita realizar algún escalamiento o transformación en los datos. A continuación, se transforman las variables categóricas a factor para poder facilitar la aplicación del algoritmo al modelo. Antes de correr los modelos, se realizó un preprocesado de los datos y se procedió a realizar el modelo y entrenarlo. En todo este proceso se tuvo que regresar a entender el problema y los datos para encontrar la mejor solución. El Flujo del Proceso de la metodología que se siguió para cada modelo se muestra en la Figura 3.

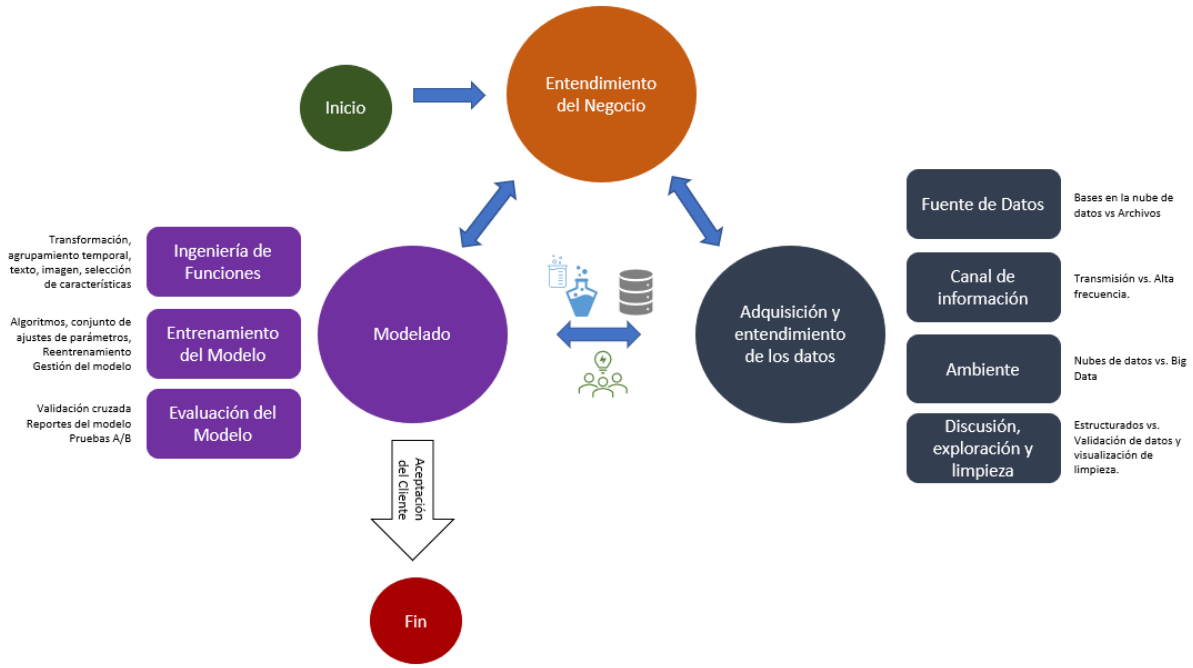


Figura 3. Representación gráfica de la metodología TDSP [37]

En la Figura 4 se presenta un pseudocódigo con todos los pasos realizados, desde la lectura de la fuente de datos hasta finalizar con el entrenamiento del modelo. Antes de crear o definir el modelo, se realiza una partición en los datos, se definen los datos de entrenamiento con un 80% y los datos de prueba con un 20%.

Con los datos de entrenamiento (80%) se procedió a la creación y definición del modelo, dependiendo del tipo de modelo se definieron los Hiperparámetros a optimizar, se evaluó el modelo y se eligió el mejor modelo.

Los datos de prueba (20%) se utilizan para ejecutar y evaluar el modelo. Se inicia con un Random Forest para identificar las variables importantes con nuestra variable de salida. Teniendo identificadas las variables importantes se decide implementar un Random Forest nuevo con las variables seleccionadas, así también se decide correr el mismo procedimiento con los modelos SVM con dos diferentes kernels, GBM, GLM y Naïve Bayes para poder medir el rendimiento y precisión de varios modelos para así comparar y elegir el mejor modelo.

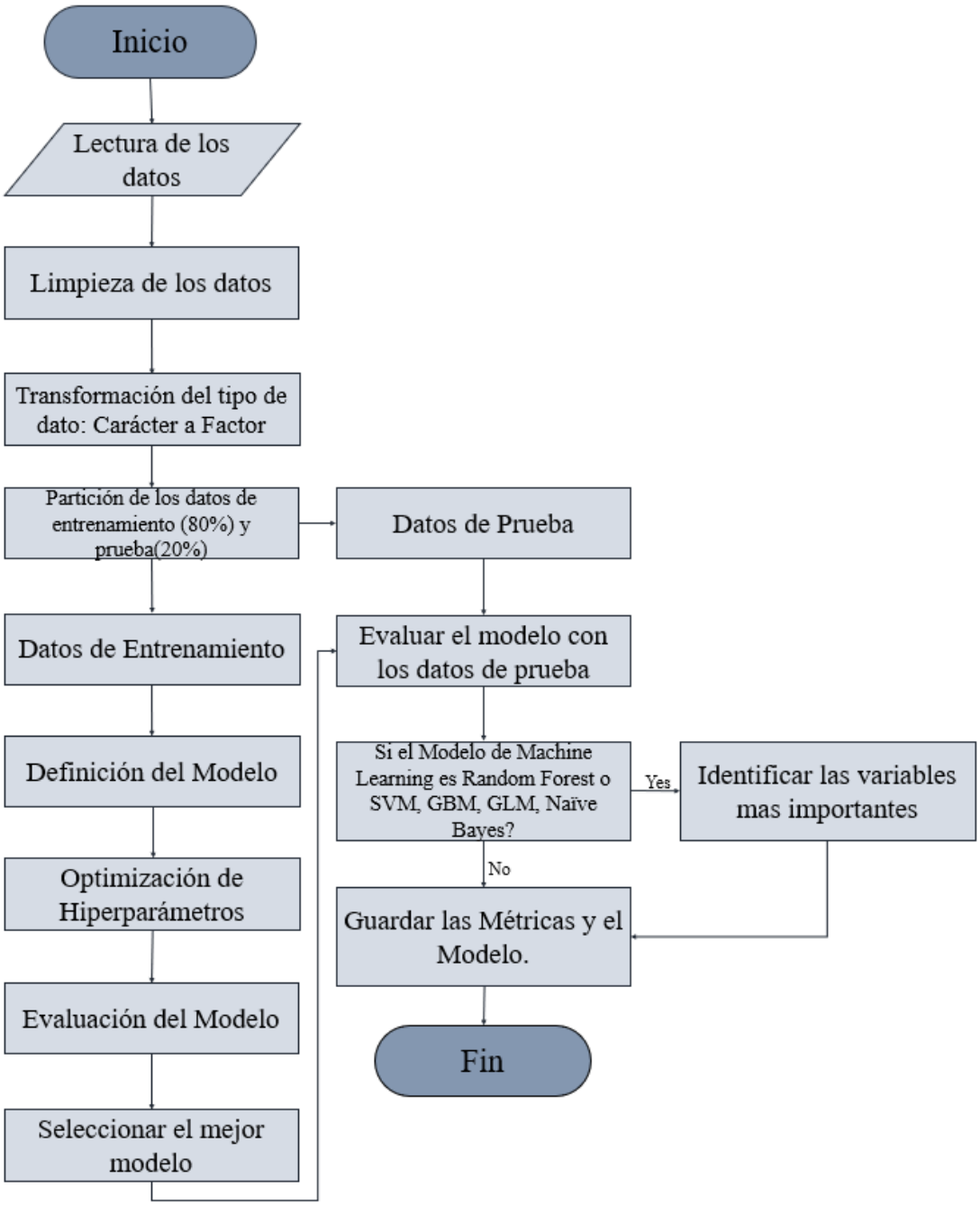


Figura 4. Flujo para la ejecución de los modelos

## 2.5 Conjunto de Datos

Para este problema en específico se decidió utilizar dos diferentes conjuntos de datos. La primera, es una encuesta llamada Survey on Mental Health in the Tech Workplace in 2014[1], la cual fue realizada en Estados Unidos en el 2014 para medir y entender las actitudes hacia la salud mental y la frecuencia de los trastornos de salud mental enfocándose en los empleados que elaboran en el área de TI. Por otro lado, se optó por introducir otras variables como el Producto Interno Bruto (PIB) para los años 2014 y 2015 [38]. La hipótesis para incluir estas variables permitirá a identificar las características de los países en donde las personas son más propensas a recibir un tratamiento de salud mental.

Se decidió incluir dos bases de datos diferentes para poder entender, no sólo las características y las probabilidades de que los empleados de TI puedan o no recibir un tratamiento de salud mental, sino también las compañías en donde dichas empresas de TI están establecidas. Varios estudios afirman que en los países de primer mundo la mayoría de las personas son más propensas a desarrollar un tratamiento de salud mental, esto debido a la alta demanda de tareas, niveles, objetivos, retos, jornadas nocturnas, el pertenecer a un buen nivel socioeconómico, generando un alto nivel de estrés personal para poder cumplir con las expectativas sociales.

Esta base de datos está conformada por 1259 observaciones y 29 variables, de las cuales tenemos un 5.2% de observaciones faltantes, un 89.7% de variables discretas y un 10.3% de variables continuas.

## 2.6 Limpieza de Datos e Imputación de Datos

Al entrenar un modelo de ML con datos erróneos o desbalanceados, se puede enfrentar a tener datos sesgados y dar un rendimiento deficiente. Para este problema en específico, se aplicaron diferentes técnicas para mejorar la calidad de los datos. A continuación, se presenta un resumen de algunas de las técnicas empleadas en el preprocesado de los datos realizadas en este trabajo.

### 2.6.1 Valores Faltantes

Durante la exploración y el análisis de los datos, se identificaron cuatro variables con datos faltantes de veintiocho variables que se tienen en el conjunto de datos, donde la variable *comment* tiene el mayor porcentaje de datos faltantes. En la Tabla 2 se presenta las variables con los valores faltantes.

Tabla 2. Valores faltantes

| **Variables** | **Valores Faltantes (%)** |
|---|---|
| self_employed | 1.43% |
| work_interfere | 20.97% |
| state | 40.91% |
| comment | 86.97% |

Previamente analizada la variable *self_employed* tiene 18 datos faltantes, durante el análisis previo se identificó una gran relación y dependencia entre la variable *self_employed* y *tech_company*, dichos datos faltantes fueron reemplazados de la siguiente manera, si un empleado trabaja en una empresa de tecnología como un empleado dependiente el valor será ¨No¨, de lo contrario si el trabajador no trabaja en una compañía de tecnología el valor sería ¨Yes¨.

De igual manera, las variables *comments, state* y *work_interfere* fueron identificadas con valores faltantes durante el análisis realizado previamente, en el se pudo identificar y descartar la importancia de estas tres variables en el modelo, por lo tanto, se decidió aplicar el método de imputación con el fin de eliminar dichas variables con un gran porcentaje de valores faltantes, para evitar ruido en las nuestras predicciones y no afectar el resultado de los modelos.

Así también se procedió en quitar la variable *Timestamp* ya que es una variable tipo fecha, el cual fue registrado cuando cada empleado contesto la encuesta.

### 2.6.2 Imputación por Moda

Posteriormente se identificaron 50 valores diferentes en la variable *Gender*, esto quiere decir que 50 géneros diferentes contestaron esta encuesta.

Para evitar inconsistencias en el conjunto de datos, se decidió aplicar la imputación por el método la moda, esto quiere decir que se identificaron los valores más frecuentes y se crearon tres grupos de categorías: *Man*, *Woman* y *Others*, los cuales se pueden observar en la Figura 5.

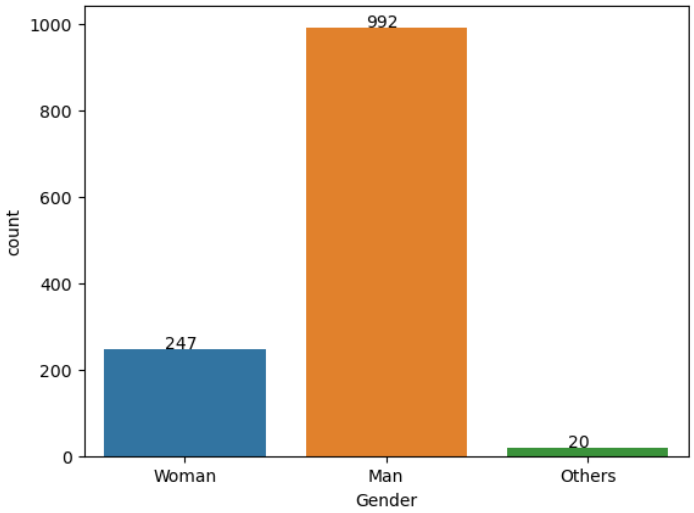

Figura 5. Genero. Nuevas Categorías

### 2.6.3 Valores Atípicos

Siguiendo con la exploración de los datos se identificaron inconsistencias con la variable de edad, la edad mínima capturada en la encuesta es de -1726 y el mayor máximo es de 99999999999, lo cual no tiene sentido. Por lo tanto, se decidió obtener una gráfica tipo box-plot para identificar todos los valores atípicos que están presente en el conjunto de datos y aplicar un método de imputación para reemplazar esas observaciones.

Analizando la edad mínima y máxima para poder trabajar legalmente se decidió reemplazar el valor mínimo -1726 por 18 y el valor máximo 99999999999 por 72 y así poder evitar inconsistencias y valores atípicos en los modelos, lo cual se muestra en la Figura 6.

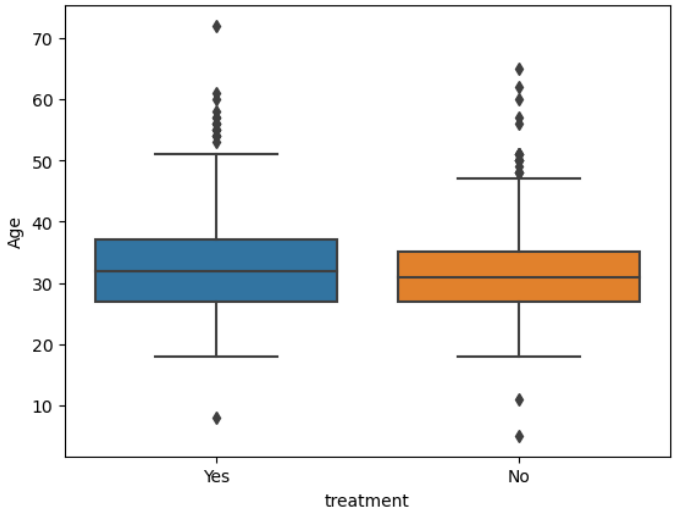


Figura 6. Edad - Valores Atípicos

## 2.7 Exploración de Datos

Después de la limpieza de los datos, se obtuvieron 1259 observaciones y 25 variables, cada registro u observación representa una encuesta por persona con diferentes características. La Tabla 3 muestra una descripción de las variables utilizadas en este trabajo.

Tabla 3. Descripción de las variables

| **Variables** | **Descripción** |
|---|---|
| *Treatment | ¿Ha buscado algún tratamiento para una condición de salud mental? |
| Age | Edad de los encuestados |
| Gender | El género de la persona encuestada |
| Country | País de origen de la persona encuestada |
| Self-employed | ¿Es usted un trabajador independiente? |
| Family history | ¿Tiene antecedentes familiares con alguna enfermedad mental? |
| Work interferes | Si tiene alguna condición de salud mental, ¿Cree usted que interfiere con su trabajo? |
| No employees | ¿Cuántos empleados tiene la empresa u organización donde trabaja? |
| Remote work | ¿Trabaja de forma remota (fuera de una oficina) al menos el 50% del tiempo? |
| Tech Company | ¿En la empresa donde trabaja es una empresa/organización enfocada en IT? |
| Benefits | ¿La empresa donde trabaja proporciona beneficios de salud mental? |
| Care options | ¿Conoce las diferentes opciones de atención de salud mental que ofrece su empresa? |
| Wellness program | ¿Alguna vez en su trabajo ha recibido algún programa sobre la salud mental como parte de un programa de bienestar para empleados? |
| Seek help | ¿En su trabajo proporcionan recursos para obtener más información sobre problemas de salud mental y cómo buscar ayuda? |
| Anonymity | ¿Está protegido su anonimato si elige aprovechar los recursos de tratamiento de salud mental o abuso de sustancias? |
| Leave | ¿Qué tan fácil es para usted tomar vacaciones por una condición de salud mental? |
| Mental-health consequence | ¿Cree que hablar sobre un problema de salud mental con su empresa tendría consecuencias negativas? |
| Phys-health consequence | ¿Cree que hablar sobre un problema de salud física con su empresa tendría consecuencias negativas? |
| Coworkers | ¿Estaría dispuesto a hablar sobre algún problema de salud mental con sus compañeros de trabajo? |
| Supervisor | ¿Estaría dispuesto a hablar sobre algún problema de salud mental con su(s) supervisor(es) directo(s)? |
| Mental-health interview | ¿Hablaría sobre un problema de salud mental durante una entrevista de trabajo? |
| Phys-health interview | ¿Hablaría sobre un problema de salud física durante una entrevista de trabajo? |
| Mental vs Physical | ¿Cree que en su trabajo se toman la salud mental tan en serio como la salud física? |
| pib_2014 | Valor de Producto Interno Bruto del año 2014 |
| pib_2015 | Valor del Producto Interno Bruto del año 2015 |

### 2.7.1 Análisis Exploratorio de la Variable Objetivo

Se realizó un análisis más profundo, como primer acercamiento para determinar si había una posible correlación entre las variables categóricas con la variable objetivo "*treatment*".

En un estudio realizado por Daniel Freeman y Jason Freeman publicado en el 2013. Libro: "*The stressed sex: Uncovering the truth about men, women, and mental health"*, se analizó que las mujeres presentan prevalencias más altas y tienen más probabilidad que los hombres de sufrir depresión, ansiedad, abuso y trastornos de alimento, esto debido a los altos niveles de hormonas que tienen.

Para este problema es especifico se identificó que el 68% del total de las mujeres que participaron en la encuesta son más probables a recibir un tratamiento de salud mental, a comparación con los hombres que fueron el 45% del total en recibir un tratamiento de salud mental, como se muestra en la Tabla 4.

Tabla 4. Relación entre la variable Treatment y Gender

| **Gender** | **Total (n = 1,259)** | **Yes -Treatment (n = 637)** | **No - Treatment (n = 622)** |
|---|---|---|---|
| Man | 78.8% | 35.7% | 43.1% |
| Woman | 19.6% | 13.5% | 6.1% |
| Others | 1.6% | 1.4% | 0.2% |

Así también, se identificó una gran diferencia entre los empleados que tienen un historial familiar de enfermedades de salud mental, el cual al tener un historial médico aumenta la probabilidad del 74% en desarrollar una enfermedad y necesitar algún tratamiento de salud mental. Según el Instituto Nacional de Salud Mental, la probabilidad de que una persona tenga un trastorno mental es mayor si otros miembros de la familia lo tienen, aunque un trastorno mental puede ser hereditario, se puede haber diferencias considerables en la gravedad de los síntomas entre los mientras de la familia, el cual lo podemos observar en la Figura 7.

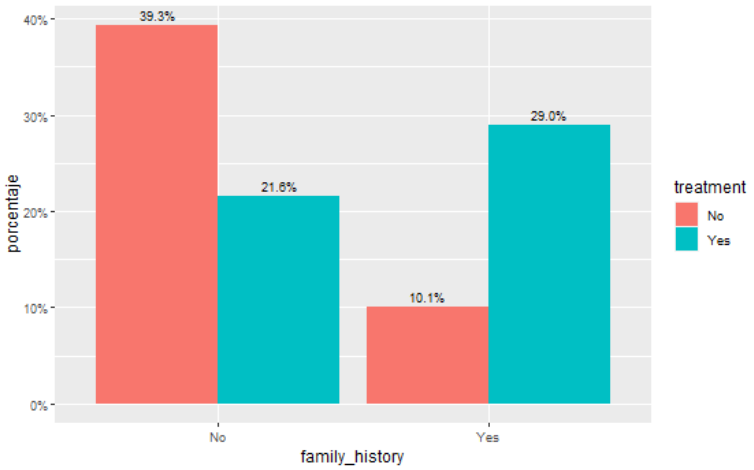


Figura 7. Empleados con Historial Familiar

En estudios realizados por la Organización Mundial de Salud (OMS), aproximadamente un 4.8% de individuos padecen de depresión y ansiedad, la cual algunas de las manifestaciones más notorias de este padecimiento son la pérdida de interés en el ámbito laboral, bajo desempeño, falta de trabajo en equipo y baja autoestima. Esto podría aumentar la probabilidad de pérdidas de empleo, crisis económicas y un bajo desempeño del empleado al no recibir un tratamiento de salud mental o la atención adecuada. Por esto es muy importante que tanto las empresas como los supervisores y compañeros de trabajo generen un ambiente de confianza y lealtad, para poder detectar alguna situación importante y poder tratarla a tiempo.

Infortunadamente, el 70% de los empleados que contestaron esta encuesta han recibido algún tratamiento de salud mental, como se muestra la figura 8.

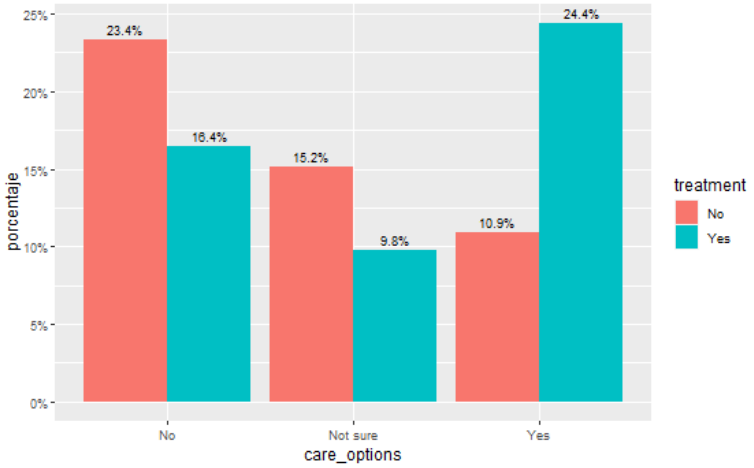

Figura 8. Empleados que han recibido un tratamiento

Por este motivo es muy importante que las empresas empleen apoyos y beneficios para así generar un ambiente con más empatía, fidelidad, con el empleado y la compañía, ya que se recibe el apoyo a una de las enfermedades más especiales, criticadas y silenciosas.

Mantener una buena salud mental beneficia a disminuir la ansiedad y la depresión, estas afectan no solo en el trabajo sino también puede llegar a afectar hasta económicamente de una persona.

Pero solo el 59% de los empleados que reciben un tratamiento de salud mental, tienen conocimiento de dichos apoyos y beneficios. Al contrario del 41% de los empleados que reciben un tratamiento de salud mental pero no piden ayuda, genera depresión, ansiedad y en algunos casos se puede llegar al suicido, sino buscan la ayuda correcta.

Después del COVID-19, que comenzó en diciembre 2019, la mayoría de las empresas tuvieron que modificar el lugar de trabajo y hacer que los empleados se hicieran responsables de su propio tiempo y esfuerzo. Lamentablemente los apoyos psicológicos y cuidados médicos presenciales por la baja demanda se tuvieron que reducir o en algunas ocasiones cancelar. Esto ocasionó que los empleados prefieren pedir ayuda o recibir un medicamento o tratamiento de salud mental anónimamente, por miedo a hacer juzgados o discriminados.

Así también, el 59% del total de los empleados que reciben un tratamiento de salud mental, prefieren no compartir su enfermedad en su trabajo, mantenerlo anónimo como lo confirmamos en los análisis anteriores, ya que piensan que esto les puede generar problemas y consecuencias.
Aunque como se muestra en la Figura 9, el 50% de los empleados que reciben un tratamiento de salud mental prefieren hablarlo con solo algunas personas en específico o simplemente mantenerlo en secreto, como ya lo mencionamos anteriormente, esta es una característica muy importante y con gran impacto en el modelo, ya que muchas variables tienen dependencia o se relacionan, por ejemplo, la variable: *anonymity*.

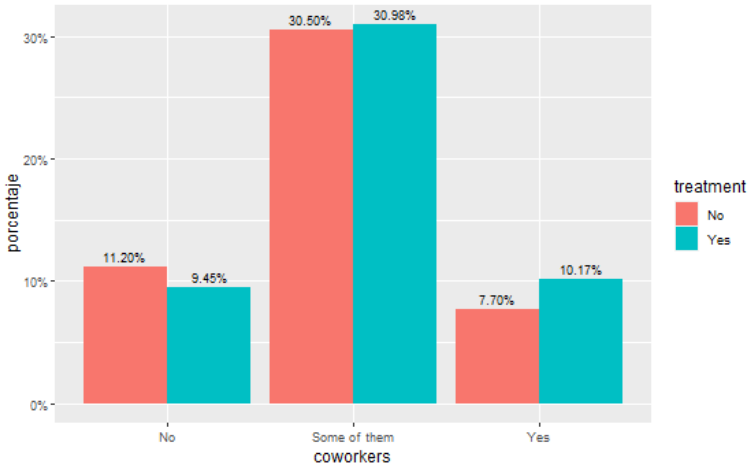


Figura 9. Hablar con un compañero sobre enfermedades mentales

La mayoría de los empleados que trabajan en pequeña empresa con un número de empleados pequeño tienen una correlación alta entre el número de empleados en una compañía y la probabilidad de recibir un tratamiento de salud mental. Como se puede ver en la figura 9.

Esto quiere decir que menor sea el número de empleados mayor es la probabilidad de recibir un tratamiento de salud mental, ya que una de las desventajas de las empresas pequeñas es el bajo presupuesto en los temas de salud, en psicología y actividades de equipo, en la cual los empleados pudieran interactuar entre sí.

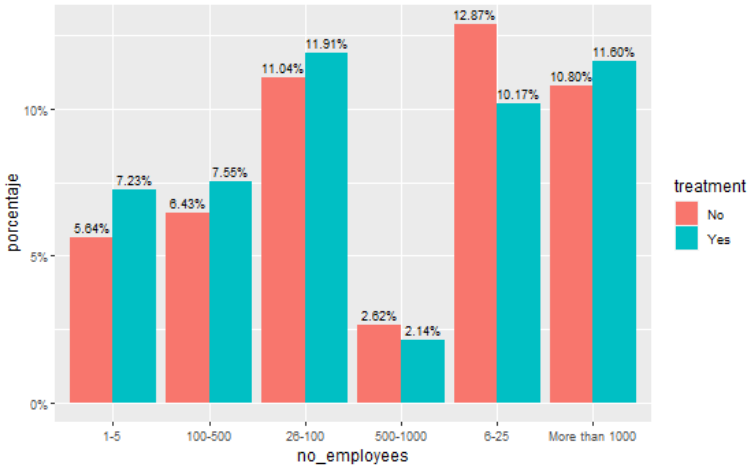


Figura 10. Tipos de empresas

Por otra parte, en la Tabla 5 se identificó que un 70% de los empleados recibieron un tratamiento de salud mental y requirieron ir a un hospital para una revisión médica, y se les hizo muy complicado pedir permiso, ya que los empleados prefirieron mantener el anonimato de su enfermedad o trastorno mental por el miedo a ser juzgados y criticados.

Tabla 5. Pedir un permiso para acudir a una cita medica

| leave | Total (n = 1,259) | Yes -Treatment (n = 637) | No - Treatment (n = 622) |
|---|---|---|---|
| Don’t know | 44.7% | 20.17% | 24.54% |
| Somewhat easy | 21.1% | 10.41% | 10.72% |
| Very easy | 16.4% | 8.18% | 8.18% |
| Somewhat difficult | 10.0% | 6.51% | 3.49% |
| Very difficult | 7.8% | 5.32% | 2.46% |

Durante el análisis y exploración de los datos se identificó una relación alta entre el PIB y las personas que han recibido un tratamiento de salud mental. Por ende, se procedió a realizar un análisis con la variable *Country*, el cual en la Tabla 6 se pueden identificar a Estados Unidos, Canadá, Reino Unido, Alemania, Francia, Australia como países del primer mundo [39], los cuales se destacan por tener un alto nivel de progreso tecnológico, científico e industrial, la calidad de vida. A continuación, se muestra los principales países con el PIB más alto y el número de trabajadores del sector TI:

Tabla 6. Países con un alto PIB

| País | No. de Tech Workers |
|---|---|
| United States | 751 |
| United Kingdom | 185 |
| Canada | 72 |
| Germany | 45 |
| Ireland | 27 |
| Netherlands | 27 |
| Australia | 21 |
| France | 13 |

### 2.7.2 Análisis del Balance de los Datos

Para poder analizar el balance de clases, se procede a crear un histograma de la variable objetivo. Como se muestra en la Figura 11 los datos/observaciones están bien balanceados (637-Yes, 622-No), por lo tanto, no es necesario aplicar ningún método de balanceo.

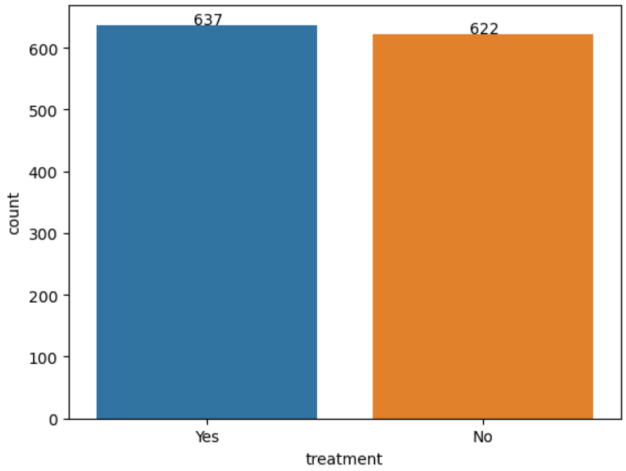

Figura 11. Histograma del Balanceo Datos

## 2.8 Preprocesado de Datos

Durante el análisis y exploración de los datos, se identificó que la mayoría de las variables eran de tipo categórica, las cuales se tuvieron que transformar a factor para poder utilizarlas en el modelo facilitando su algoritmo.

# Capítulo 4

## 4.1 Introducción

Como se mencionó en el Capítulo 2, se utilizaron cinco diferentes modelos de clasificación en este trabajo de tesis:

1. Modelo 1. Se implementaron dos **Random Forest,** uno para determinar las variables con mayor importancia para la variable de salida y otra para medir el rendimiento sin incluir las variables que no tuvieron importancia en el modelo.
2. Modelo 2. **Gradient Boosting Machine GBM**, en donde se incluyeron sólo las variables que tuvieron un impacto importante para la variable de salida.
3. Modelo 3. **Generalized Linear Models GLM**, en donde se incluyeron sólo las variables que tuvieron un impacto importante para la variable de salida.
4. Modelo 4. **Support Vector Classifier SVC**, en donde se aplicaron dos kernels diferentes, RBF Kernel y Sigmoid Kernel, y se incluyeron sólo las variables que tuvieron un impacto importante para la variable de salida.
5. Modelo 5. **Naïve Bayes**, en donde se tomatón únicamente las variables que tuvieron un impacto importante para la variable de salida.

Tres modelos se implementaron en RStudio y cuatro en Python, para poder comparar el rendimiento no solo de los modelos, sino del ambiente en el que se ejecutaron.

## 4.2 Resultados

Como se mencionó en los capítulos 2 y 3, cuatro métricas se utilizaron para poder medir y calcular el rendimiento de los modelos: Accuracy, ROC AUC, F1, Sensitivity y Specificity. Para cada modelo de Clasificación, se siguió en la Figura 3. Process Flow - ML. Todos los resultados serán mostrados a continuación.

### 4.2.1 Resultados Obtenidos en RStudio

Cinco Modelos diferentes fueron desarrollados en RStudio 2022.02.0+443 con el Sistema Operativo Windows 10.0. se utilizaron diferentes paquetes, H2O, scikit-learn, tidyverse, tidymocels, Bayesian-optimization. Las cuales se pueden identificar en la Tabla 7.

Tabla 7. Librerías en RStudio

| **Classifier** | **Librerías** |
|---|---|
| Random Forest | *sklearn.ensemble.RandomForestClassifier* |
| GBM | *h2o.grid.gbm* |
| GLM | *h2o.grif.glm* |

Para nuestro primer modelo implementado en RStudio, el Random Forest la Tabla 8 muestra los resultados de los Hiperparámetros los cuales fueron Optimizados por Bayesiano. Así también este modelo fue utilizado para identificar y predecir las variables que tienen un impacto muy importante en el modelo y la variable de salida.

Tabla 8. Random Forest – Variable Importance – Hiperparámetros en RStudio

| **Hiperparámetros** | **Valor** |
|---|---|
| Random Forest – Variable Importance | |
| tress | 1270 |
| min_n | 31 |
| mean | 0.7667 |
| n | 5 |
| std_err | 0.01161784 |
| .iter | 0 |

El método utilizado para identificar la importancia de las variables fue la importancia por permutación, los cuales se muestran en la Figura 12. Las variables identificadas con un impacto negativo fueron: *tech_company*, *mental_vs_physical*, y *phys_health_consequence*, por lo que no fueron consideradas para los siguientes modelos.

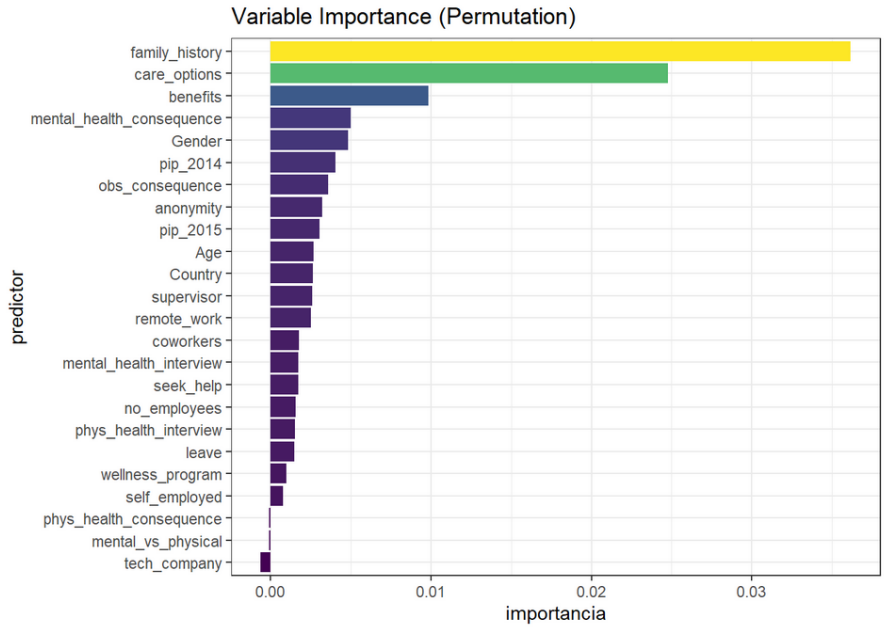


Figura 12. Random Forest – Variable Importance

Para medir la precisión del modelo se obtuvo la matriz de confusión, la cual se puede apreciar en la Tabla 9. La diagonal muestra el número de valores que se predijeron correctamente, en este caso si la variable *treatment == No*, el modelo encontró 96 valores correctos de 132 (96+36), mientras que para las personas que, si han recibido un tratamiento, el modelo encontró 92 correctas de 121 (29+92).

Esto quiere decir que, cuando una persona recibe un tratamiento de salud mental, se tiene una muy mala precisión de predicción, cuando las personas no han recibido un tratamiento de salud mental se tiene una muy buena precisión de predicción.

Tabla 9. Random Forest – Variable Importance - Matriz de Confusión en RStudio

| | **No** | **Yes** |
|---|---|---|
| **No** | 96 | 36 |
| **Yes** | 29 | 92 |

Después de identificar las variables más importantes, se decidió ejecutar un nuevo modelo de Random Forest para comparar el performance. En la Tabla 10 se muestra el mejor modelo con los resultados de los hiperparámetros los cuales fueron optimizados por bayesiano.

Tabla 10. Random Forest v2 – Hiperparámetros en RStudio

| Hiperparámetros | Valor |
|---|---|
| Random Forest v2 | |
| tress | 463 |
| min_n | 10 |
| mean | 0.7717836 |
| n | 5 |
| std_err | 0.0123688 |
| .iter | 0 |

También, se realizó la matriz de confusión para poder comparar el performance y precisión del segundo modelo con las variables más importantes. La Tabla 11 muestra que el porcentaje de recibir o no un tratamiento de salud mental es el mismo.

Tabla 11. Random Forest v2 - Matriz de Confusión en RStudio

| | No | Yes |
|---|---|---|
| **No** | 94 | 34 |
| **Yes** | 31 | 94 |

Para el tercer modelo GBM, fueron desarrollados y optimizandos solo los tres hiperparámetros más importantes: *learn_rate*, *max_depth* y *sample_rate,* así mismo se considerando las variables con mayor importancia para el modelo y la variable de salida. La Tabla 12 muestra los resultados de los hiperparámetros del mejor modelo.

Tabla 12. GBM – Hiperparámetros en RStudio

| Hiperparámetros | Valor |
|---|---|
| GBM | |
| number_of_trees | 1000 |
| number_of_internal_trees | 1000 |
| min_depth | 9 |
| max_depth | 15 |

En la Tabla 13, se muestra los resultados obtenidos de la matriz de confusión, los cuales, a comparación con los dos modelos de Random Forest, se obtuvo una mejor predicción de si una

persona pueda requerir un tratamiento de salud mental, que cuando la persona no requieran un tratamiento de salud mental.

Tabla 13. GBM - Matriz de Confusión en RStudio

| | **No** | **Yes** |
|---|---|---|
| **No** | 68 | 72 |
| **Yes** | 14 | 101 |

El cuarto modelo GBM fue desarrollado con una segunda optimización, tras encontrar los mejores hiperparámetros del modelo tres (GBM), se decidió optimizar el resto de Hiperparámetros utilizando Random Discrete, así también se consideraron las variables más influyentes. La Tabla 14 muestra los resultados de los hiperparámetros del mejor modelo.

Tabla 14. GBM v2 - Segunda Optimización– Hiperparámetros en RStudio

| **Hiperparámetros** | **Valor** |
|---|---|
| GBM – Segunda Optimización | |
| number_of_trees | 800 |
| number_of_internal_trees | 800 |
| min_depth | 4 |
| max_depth | 9 |

Para comparar los dos modelos de GBM, se realizó un análisis de la matriz de confusión para comparar la precisión del performance. La Tabla 15 muestra que el segundo modelo tiene mejor precisión de predicción cuando una persona pueda recibir un tratamiento de salud mental.

Tabla 15. GBM v2 - Segunda Optimización - Matriz de Confusión en RStudio

| | **No** | **Yes** |
|---|---|---|
| **No** | 83 | 57 |
| **Yes** | 24 | 91 |

Por último, se decidió correr un quinto modelo en RStudio el cual fue GLM, considerando las variables más importantes en el modelo. La Tabla 16 muestra los resultados de los hiperparámetros del mejor modelo.

Tabla 16. GLM - Hiperparámetros en RStudio

| **Hiperparámetros** | **Valor** |
|---|---|
| GLM | |
| alpha | 0.95 |
| lambda | 0.005949 |

Para medir la precisión de predicción del modelo, se calculó una matriz de confusión en la cual se puede apreciar que el modelo tiene una mejor predicción cuando una persona no reciba un tratamiento de salud metal, que cuando una persona pueda recibir un tratamiento de salud mental. Los resultados obtenidos se encuentran en la Tabla 17.

Tabla 17. GLM - Matriz de Confusión en RStudio

| | **No** | **Yes** |
|---|---|---|
| **No** | 101 | 39 |
| **Yes** | 23 | 92 |

La Tabla 18, muestra los resultados obtenidos de los modelos de las métricas utilizadas en la cual se puede identificar como el mejor modelo GLM con un ROC AUC de 80% ejecutándolo en RStudio.

Tabla 18. Resultados obtenidos en RStudio

| **Models** | **Accuracy** | **ROC AUC** | **Specificity** | **Sensitivity** |
|---|---|---|---|---|
| Random Forest - Variable Importance | 0.743 | - | 0.727 | 0.76 |
| Random Forest | 0.743 | - | 0.734 | 0.752 |
| GBM H2O | 0.663 | 0.756 | 0.486 | 0.878 |
| GBM H2O – Second Optimization | 0.682 | 0.749 | 0.593 | 0.791 |
| GLM H2O | **0.682** | **0.806** | **0.721** | **0.8** |

Las Figura 13, 14 y 15, muestra los resultados de la métrica Accuracy, Specificity y Sensitivity, en las cuales el Random Forest se muestra como uno de los mejores en la métrica de Accuracy con un porcentaje de 74%, esto quiere decir que nuestro modelo ha acertado en la mayoría de los casos

cuando una persona puede o no requerir un tratamiento de salud mental, al contrario, con el performance de la métrica Sensitivity, la cual se muestra en la Figura 14, donde el modelo GBM acertó un 87% en los casos cuando una persona puede recibir un tratamiento de salud mental.

La Figura 15, muestra el Random Forest con un 73% en la métrica de Specificity, donde se puede observar como el modelo hacerla en la mayoría de los casos cuando una persona no puede recibir un tratamiento de salud mental.

Para este trabajo de Tesis los mejores resultados de los modelos se pueden identificar en la Tabla 18, utilizamos el ROC AUC para medir los casos tanto positivos como negativos bajo la curva, y así obtener una mejor predicción. Por esto, el modelo seleccionado fue el GLM con un performance del 80 % y utilizando el modelo Random Forest para la importancia de las variables, la cual se encuentra en la Figura 12.

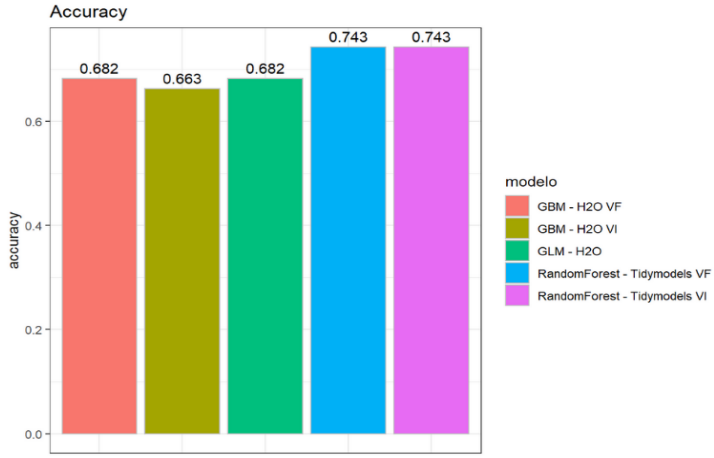


Figura 13. Resultados obtenidos en RStudio - Accuracy

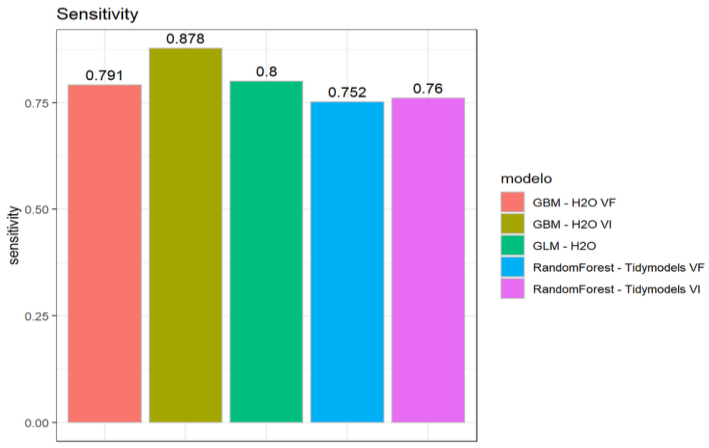


Figura 14. Resultados obtenidos en RStudio - Sensitivity

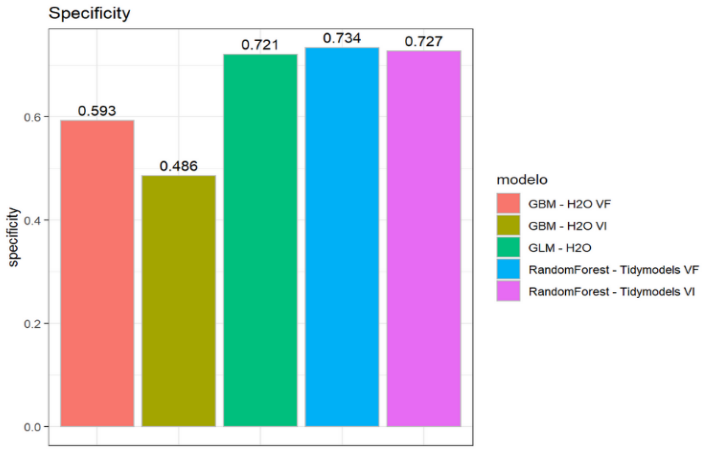


Figura 15. Resultados obtenidos en RStudio - Specificity

### 4.2.2 Resultados Obtenidos en Python

Cuatro modelos diferentes fueron desarrollados en Notebook Jupyter versión 6.4.8 con Python 3. Donde se utilizaron las siguientes librerías, Tabla 19.

Tabla 19. Librerías en Python

| **Classifier** | **Librerías** |
|---|---|
| Random Forest | *sklearn.ensemble.RandomForestClassifier* |
| SVC | *Sklearn.svm.SVC* |
| Naïve Bayes | *pgmpy.models.BayesianModel* |

El primer modelo desarrollado en Python fue Random Forest para así comparar los resultados en diferentes lenguajes computacionales y ambientes. La Tabla 20, muestran los resultados de los hiperparámetros del mejor modelo los cuales fueron optimizados por el método Cross Validation utilizando la métrica f1.

Tabla 20. Random Forest – Hiperparámetros en Python

| **Hiperparámetros** | **Valor** |
|---|---|
| Random Forest | |
| max_depth | 5 |
| max_leaf_nodes | 50 |
| n_estimators | 30 |

Para poder medir el rendimiento y la precisión del modelo se realizó la matriz de confusión la cual se muestra en la Tabla 21, en la cual se puede identificar que el modelo encontró 82 datos positivos

de 127 en los casos cuando las personas requieren un tratamiento de salud mental. De lo contrario el modelo acertó 104 veces en los casos negativos de 125 del total de los casos.

Tabla 21. Random Forest - Matriz de Confusión en Python

| | **No** | **Yes** |
|---|---|---|
| **No** | 104 | 21 |
| **Yes** | 45 | 82 |

En la Tabla 22, se pueden identificar los Hiperparámetros del mejor modelo de SVC utilizando el Kernel RBF los cuales fueron optimizados por el método de Cross Validation con 5-folds, utilizando la métrica de f1.

Tabla 22. SVC RBF Kernel - Hiperparámetros en Python

| **Hiperparámetros** | **Valor** |
|---|---|
| SVC – RBF Kernel | |
| C | 100 |
| gamma | 0.001 |

Así también, se calculó la matriz de confusión para identificar que tan preciso fue el modelo. De los cuales hubo un acierto de 85 casos de 127 en donde el modelo acertó cuando una persona puede requerir un tratamiento de salud mental, por lo contrario, cuando una persona no es necesario que reciba un tratamiento de salud mental el modelo acertó 99 veces de 125. Esto se observa en la Tabla 23.

Tabla 23. SVC RBF Kernel - Matriz de Confusión en Python

| | **No** | **Yes** |
|---|---|---|
| **No** | 99 | 26 |
| **Yes** | 42 | 85 |

Por otro lado, la Tabla 24 muestra los hiperparámetros del mejor modelo el cual se utilizó el SCV con el Kernel Sigmoide y fueron optimizados por el método de Cross Validation con 5-folds utilizando la métrica de f1.

Tabla 24. SVC Sigmoid Kernel - Hiperparámetros en Python

| Hiperparámetros | Valor |
|---|---|
| SVC – Sigmoid Kernel | |
| C | 0.5 |
| coef0 | 0.2 |
| gamma | 0.001 |

Igualmente, la Tabla 25 se calculó la matriz de confusión para visualizar los casos tanto positivos como negativos que este modelo acertó. De los cuales 105 casos de 125 el modelo acertó en los casos negativos y 76 de 127 en los que el modelo acertó cuando una persona puede requerir un tratamiento de salud mental.

Tabla 25. SVC Sigmoid Kernel - Matriz de Confusión en Python

| | No | Yes |
|---|---|---|
| **No** | 105 | 20 |
| **Yes** | 51 | 76 |

Adicional se obtuvo una red bayesiana utilizando el modelo Naïve Bayes, para visualizar la red con las variables que tienen más impacto en nuestro modelo y tienen una relación alta con nuestra variable objetivo. La Figura 16 muestra la red propuesta por este modelo, de las cuales 10 variables de 25 fueron identificadas con un mayor impacto, de las cuales como ya habíamos analizado están family_history, Age, Gender, remote_work y wellness_program.

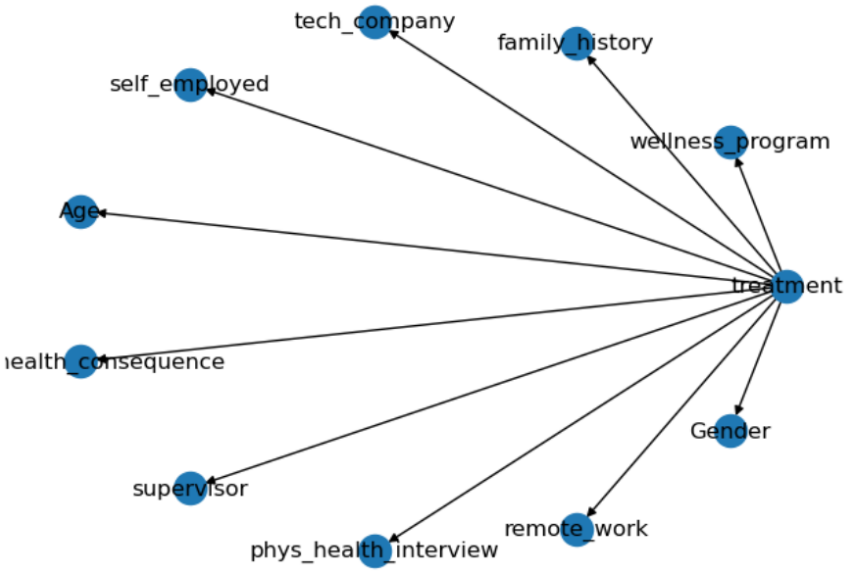


Figura 16. Naïve Bayes - Red Bayesiana en Python

Además, se corrió la matriz de confusión para evaluar el modelo. Obteniendo un 93 de 133 casos en los que el modelo acertó cuando se puede recibir un tratamiento de salud mental, o bien 89 de 119 en los que no se puede recibir un tratamiento. Los resultados obtenidos están en la Tabla 26.

Tabla 26. Naïve Bayes - Matriz de Confusión en Python

| | **No** | **Yes** |
|---|---|---|
| **No** | 89 | 30 |
| **Yes** | 40 | 93 |

La Tabla 27 muestra los resultados obtenidos de los modelos en el cual se puede identificar como el mejor modelo Support Vector Clasifier utilizando el Kernel Sigmoid con un ROC AUC de 73.9% ejecutándolo en Python.

Tabla 27. Resultados obtenidos en Python

| **Models** | **Accuracy** | **ROC AUC** | **F1_Score** | **Specificity** | **Sensitivity** |
|---|---|---|---|---|---|
| Random Forest | 0.738 | 0.714 | 0.598 | 0.699 | 0.832 |
| SVC – RBF Kernel | 0.730 | 0.719 | 0.713 | 0.724 | 0.792 |
| **SVC – Sigmoid Kernel** | **0.718** | **0.739** | **0.669** | **0.727** | **0.840** |
| Naïve Bayes | 0.731 | 0.682 | 0.646 | 0.748 | 0.722 |

La Figura 17 muestra el resumen visualmente de todas las métricas obtenidas de los modelos, siendo el Random Forest un modelo no optimo, este se utilizará para la importancia de las variables, la cual se encuentra en la Figura 12.

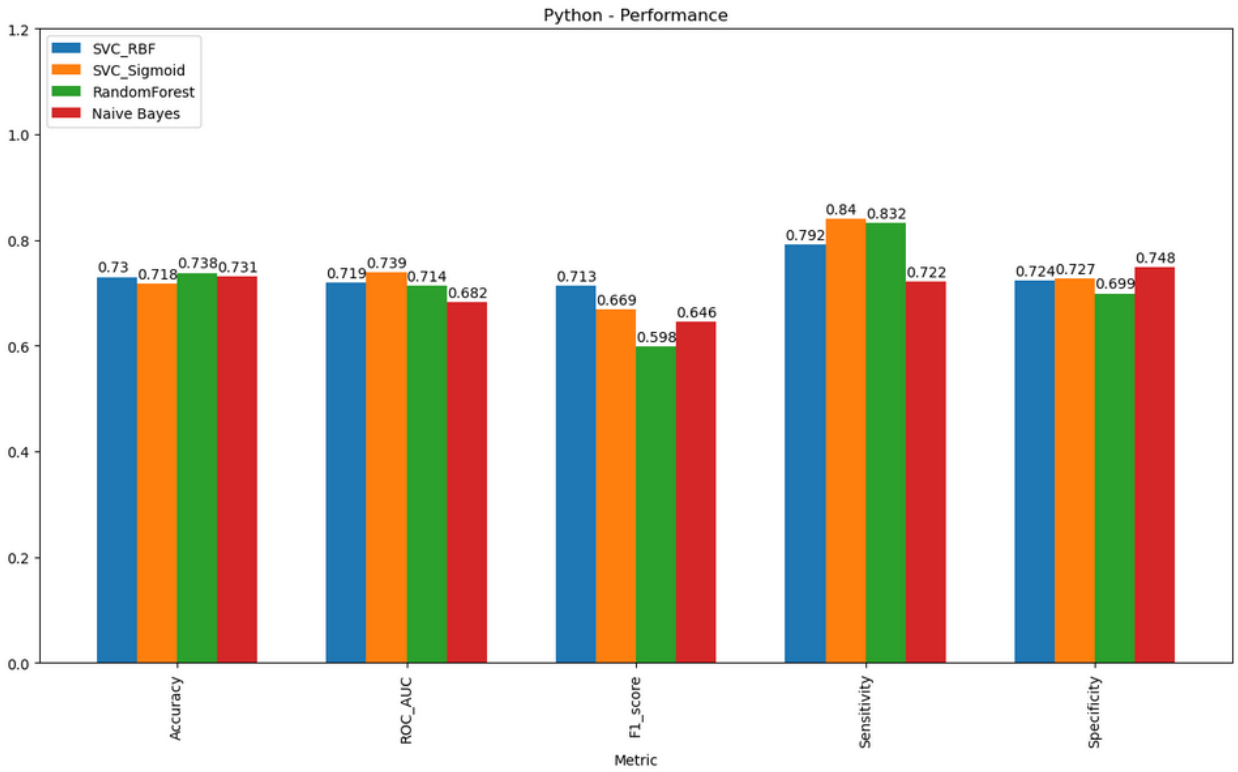


Figura 17. Resultados obtenidos en Python - Métricas

## 4.3 Comparación de resultados

Para poder seleccionar el mejor modelo, primero se hace una depuración entre RStudio y Python seleccionado el mejor modelo, los cuales se muestran en las Tablas 28 y 29.

La Tabla 28 muestra el mejor modelo el cual fue ejecutado en RStudio, el cual fue GLM con un ROC AUC de 80%.

Tabla 28. Mejor Modelo - RStudio

| **Modelo GLM** | **Accuracy** | **ROC AUC** | **Specificity** | **Sensitivity** |
|---|---|---|---|---|
| **GLM H2O** | **0.682** | **0.806** | **0.721** | **0.8** |

La Tabla 29 muestra el mejor modelo ejecutado en Python, el cual fue SVC utilizando el kernel Sigmoide con un ROC AUC de 73%.

Tabla 29. Mejor Modelo - Python

| Modelo SVC | Accuracy | ROC AUC | F1_Score | Specificity | Sensitivity |
|---|---|---|---|---|---|
| SVC – RBF Kernel | 0.730 | 0.719 | 0.713 | 0.724 | 0.792 |
| **SVC – Sigmoid Kernel** | **0.718** | **0.739** | **0.669** | **0.727** | **0.840** |

Para poder seleccionar el mejor modelo, se enfocarán en el rendimiento de la métrica ROC AUC, ya que una de sus ventajas es que puede medir la probabilidad de los casos positivos predichos contra los casos negativos obtenidos, esto quiere decir que ayuda a encontrar un umbral que se adapta al problema en específico y así poder obtener un rendimiento entre los falsos positivos y los verdaderos positivos. Ya que la variable de salida es cualitativa, ¨Si¨ y ¨No¨.

Obteniendo un ROC AUC de 73%, el mejor modelo para predecir si un trabajador de TI puede requerir un tratamiento de salud mental es el Support Vector Machine para clasificación, en el cual utilizamos el kernel Sigmoide, ya que fue el modelo más preciso y con un mejor performance.

Por otro lado, el modelo GLM obtuvo un 80% de ROC AUC, al ser ejecutado en RStudio y H2O, no es posible mandarlo a producción, ya que RStudio es utilizado para realizar análisis y exploración de los datos, adicionalmente para estos dos ambientes es necesario tener un ambiente externo y lo cual es más propenso a tener fallas de conexión.

## 4.4 Análisis del Modelo

Utilizando nuestro mejor modelo SVC con el kernel Sigmoide el cual fue seleccionado por su alto porcentaje en la métrica ROC AUC con un 73% en la sección 4.3, se realizaron los siguientes análisis.

Se decidieron incluir las once variables más importantes para obtener los resultados por percentiles de probabilidad, mostrar un heatmap con las variables más relevantes el cual se encuentra en la Figura 18.

En la Figura 18 se puede observar cómo los percentiles están siendo agrupados correctamente por la probabilidad de cada una de las variables, logrando obtener una alta relación entre los percentiles 1, 2, 3, 4 los cuales se encuentran dentro del promedio teniendo un orden descendente, como lo podemos observar en la Figura 19, con mayor probabilidad y de que las mujeres de 39 años, con historial familiar, trabajando en una empresa donde tienen beneficios, cuidados y apoyos tanto de la empresa como del supervisor para las personas que puedan requerir un tratamiento de salud mental.

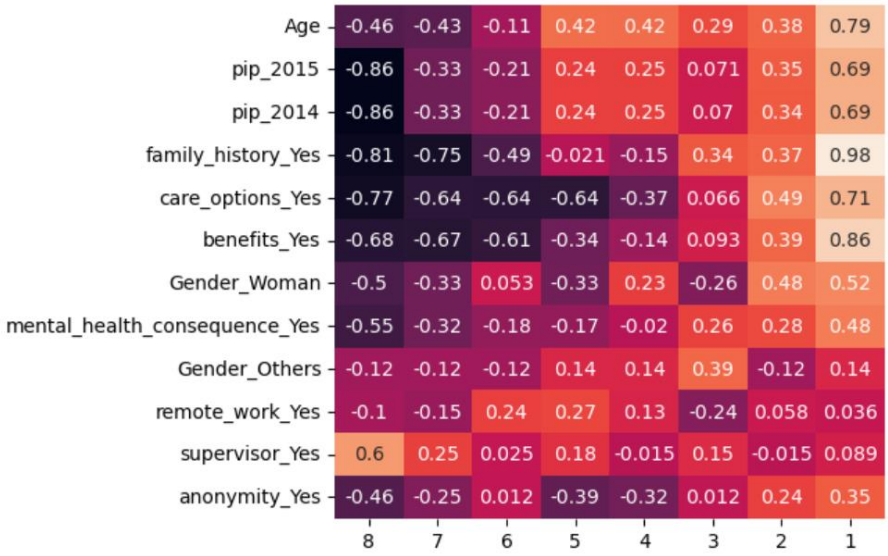


Figura 18. Heatmap por percentiles - SVC Sigmoide

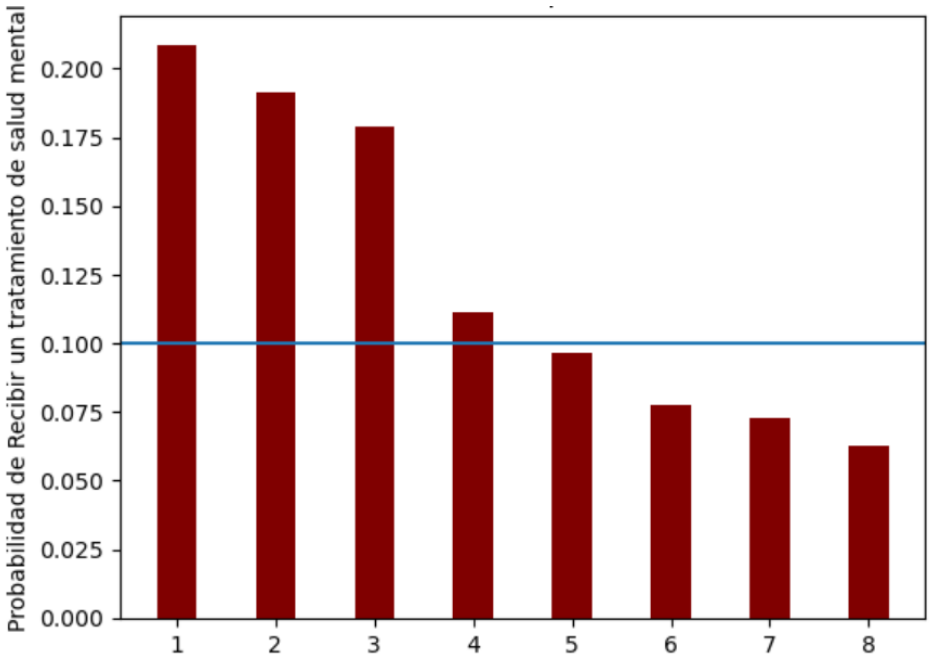


Figura 19. Distribución por percentiles - SVC Sigmoide

Por otro lado, ya que este trabajo es un problema de clasificación uno de los algoritmos más comunes es el Random Forest, por eso se decidió calcular y generar el headmap de dicho algoritmo para comprobar como el modelo están agrupando nuestras clases. Lo cual se puede observar en la Figura 21, que, a pesar de ser un algoritmo muy popular y con muy buenas métricas, para este problema de clasificación se recomiendo utilizar el Random Forest solo para identificar las

variables importantes no para la toma de decisiones, las cuales fueron identificadas en la Figura 12.

Ya que la probabilidad de los percentiles debe tener un orden descendente, en la Figura 21 se puede observar como el percentil 7 y 2, tiene un alto impacto en el modelo dando como resultado que el Random Forest no está ordenando nuestras clases correctamente.

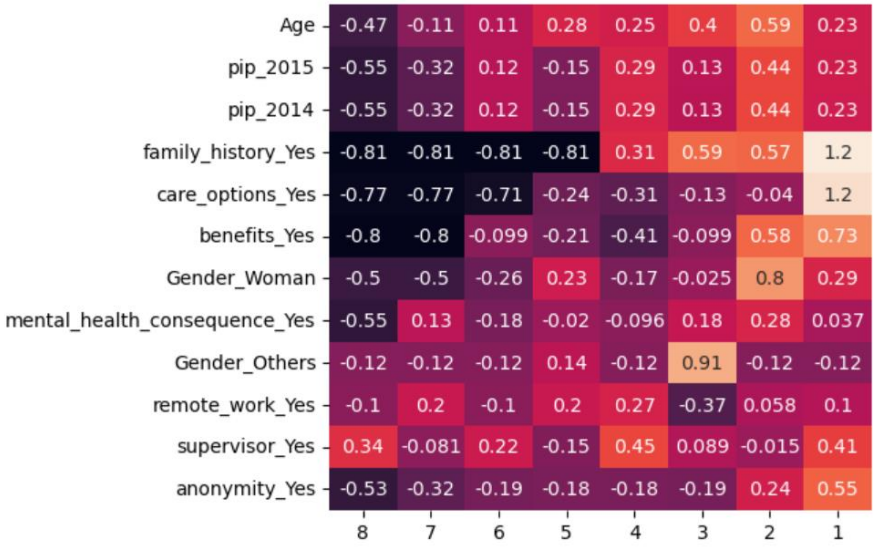

Figura 20. Heatmap por percentiles - Random Forest

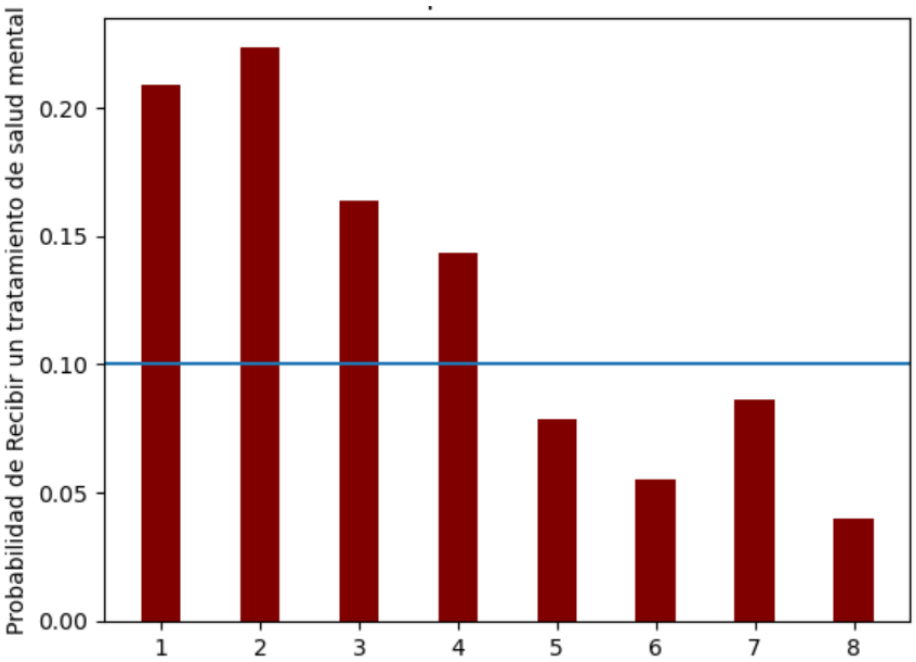

Figura 21. Distribución por percentiles - Random Forest

El modelo SVC con el kernel Sigmoide, tuvo mejor comportamiento ya que los deciles fueron ordenados de manera descendente clasificando nuestras clases de una manera correcta, esto

ayudara tanto al área de salud como al área de recursos humanos a identificar una detección temprana de algún problema de salud mental en los trabajadores.

# Conclusiones

Se implementaron 5 modelos diferentes en el ambiente RStudio y cuatro en el ambiente Python. Obteniendo como el mejor modelo y mejores métricas en Python el Support Vector Clasifier utilizando el kernel Sigmoide para la optimización de los hiperparámetros obteniendo un ROC AUC con un 73%.

Para este trabajo de tesis se identificó una relación entre si una persona requiere o no un tratamiento de salud mental. Para las personas que podrían requerir un tratamiento de salud mental se lograron identificar varios factores y características, en las cuales, tanto el área de salud como las empresas podrían mejorar. Por ejemplo, creando apoyos, actividades físicas al finalizar la jornada laboral, apoyando y generando un ambiente de confianza a los empleados que tuvieron o están en un tratamiento de salud mental y así obtener un buen rendimiento de dichos empelados.

A pesar de obtener buenas métricas en el Random Forest, en esta tesis no se recomiendo utilizarlo ya que el modelo no ordena de manera correcta nuestras clases por deciles. Por lo tanto, utilizó para la identificación y selección de la importancia de las variables de probabilidad.

Hay alta probabilidad y de que las mujeres de aproximadamente 39 años, con historial familiar, trabajando en una empresa donde tienen beneficios, cuidados y apoyos tanto de la empresa como del supervisor para las personas que puedan recibir un tratamiento de salud mental.

Se puede decir que la disminución del absentismo laboral y el aumento de la calidad de vida en el trabajo son dos de los grandes desafíos a los que las empresas y organizaciones deben hacer frente hoy en día y en donde, sin lugar a duda, la Psicología tiene mucho que aportar.

En este trabajo de tesis, se puede ver, que el momento de realizar un modelo, se debe realizar análisis adicionales que permitan tomar decisiones de cuál es el modelo que tiene mejor comportamiento para la predicción. No solo una métrica nos indica que el modelo arroja buenos resultados.

Al tener un análisis adicional por deciles podemos proporcionar al personal de salud alertas que le permitan identificar personas con una alta probabilidad de requerir un tratamiento sin necesidad de correr un modelo (headmap).

Para futuros trabajos se aplicará esta encuesta no solo a los empleados del área de TI, sino también a otro tipo de profesionistas para identificar y comparar los factores y características de que un empleado pueda recibir un tratamiento de salud mental no solo en el área de TI sino, a empleados en general, y así detectar a tiempo si una persona necesita algún tratamiento de salud mental, se aplicara el Random Forest para poder identificar la importancia de las variables y nuestro modelo seleccionado para realizar un análisis más profundo y así detectar el rendimiento y la precisión de nuestro modelo.

*Nelson Mandela:*
*“Siempre parece imposible hasta que se hace”.*

# Referencias Bibliográficas 5


[1] k Mental Health in Tech Survey. (s. f.). Kaggle. Recuperado 29 de octubre de 2022, https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

[2] La salud mental en cifras. (s. f.). Confederacion salud mental españa. Recuperado 13 de

noviembre de 2022, de https://comunicalasaludmental.org/guiadeestilo/la-salud-mental-en-cifras/

[3] Mental Health in the Workplace (https://www.helpguide.org/articles/work/mental-health-in-the-workplace.htm)

[4] To Disclose or Not to Disclose: A Multi-stakeholder Focus Group Study on Mental Health Issues in the Work Environment (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7031172/)

[5] PTO Benefits (https://mitrefinch.com/blog/pto-benefits/)

[6] The Big Shifts driving Workplace Evolution of the Future (https://www.shrm.org/shrm-india/pages/the-big-shifts-driving-workplace-evolution-of-the-future.aspx)

[7] Shrijoy Chowdhury. (Agosto 23, 2021). https://www.kaggle.com/code/shrijoychowdhury/mental-health-survey-svm-decision-tree/notebook. Agosto 23, 2022, de kaggle Sitio web: https://www.kaggle.com/code/shrijoychowdhury/mental-health-survey-svm-decision-tree/notebook

[8] PAVAN YELURI. (2022). A Complete Problem BreakDown(Best Score). 2022, de kaggle Sitio web: https://www.kaggle.com/code/pavanyeluri/a-complete-problem-breakdown-best-score#Identify-the-key-features-that-lead-to-mental-health-problems-in-tech-space

[9] Megan Risdal. (2018). Predictors of mental health illness. 2022, de Kaggle Sitio web: https://www.kaggle.com/code/kairosart/machine-learning-for-mental-health-1

[10] ADITI MULYE. (2021). Mental Health at Workplace . 2022, de kaggle Sitio web: https://www.kaggle.com/code/aditimulye/mental-health-at-workplace

[11] OBSERVATORIO VASCO DE ACOSO MORAL. (s. f.). *La SALUD MENTAL de las y los TRABAJADORES*. Osalan.euskadi.eus. https://www.osalan.euskadi.eus/contenidos/libro/medicina_201320/es_saludmen/adjuntos/salud_mental_trabajadores.pdf

[12] Supporting co-workers (https://www.headtohealth.gov.au/supporting-someone-else/supporting/co-workers#:~:text=The%20ties%20you%20make%20with,a%20lower%20risk%20of%20burnout.)

[13] *Introduction to Artificial Intelligence*. (s. f.). Recuperado 15 de noviembre de 2022, de https://link.springer.com/book/10.1007/978-3-319-58487-4

[14] *¿Qué es un árbol de decisión?* (s. f.). IBM. Recuperado 13 de noviembre de 2022, de https://www.ibm.com/es-es/topics/decision-trees

[15] *Algoritmo de árboles de decisión de Microsoft*. (s. f.). Microsoft. Recuperado 13 de noviembre de 2022, de https://learn.microsoft.com/es-es/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=asallproducts-allversions

[16] *Decision tree methods: applications for classification and prediction*. (s. f.). Recuperado 15 de noviembre de 2022, de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

[17] Random Forests (https://deepai.org/machine-learning-glossary-and-terms/random-forest )

[18] Classification Algorithm (https://www.sciencedirect.com/topics/engineering/classification-algorithm#:~:text=A%20classification%20algorithm%2C%20in%20general,the%20other%20into%20negative%20values )

[19] Malware Classification Using Static Analysis Based Features (https://cs.fit.edu/~pkc/papers/cics17.pdf )

[20] Why do They Engage in Such Hard Programs? The Search for Excellence in Youth Basketball (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3737830/ )

[21] Amat Rodrigo, J. (2017, febrero). *Árboles de decisión, random forest, gradient boosting y C5.0*.

RPubs by RStudio. https://rpubs.com/Joaquin_AR/255596

[22] Prediction of Breast Cancer using Machine Learning Approaches (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/)

[23] The Learning Rate (https://www.andreaperlato.com/theorypost/the-learning-rate/)

[24] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. In *The elements of statistical learning* (pp. 337-387). Springer, New York, NY.

[25] *Grid (Hyperparameter) Search*. (s. f.). H2O.ai. Recuperado 30 de octubre de 2022, de https://h2o-release.s3.amazonaws.com/h2o/rel-yau/2/docs-website/h2o-docs/grid-search.html#gbm-hyperparameters

[26] Generalized linear model (https://en.wikipedia.org/wiki/Generalized_linear_model)

[27] Amat Rodrigo, J. (2018, julio). *Machine Learning con H2O y R*. RPubs by RStudio. https://rpubs.com/Joaquin_AR/406480

[28] scikit-learn developers (BSD License). (2007). *RBF SVM parameters*. Scikit Learn. https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

[29] Effectiveness and Limitations of Statistical Spam Filters (https://arxiv.org/pdf/0910.2540#:~:text=Spam%20is%20not%20only%20clogging,that%20directly%20cause%20financial%20losses.)

[30] scikit-learn developers (BSD License). (2007). *One-class SVM with non-linear kernel (RBF)*. Scikit Learn. https://scikit-learn.org/stable/auto_examples/svm/plot_oneclass.html

[31] scikit-learn developers (BSD License). (2007). *sklearn.metrics.pairwise.sigmoid_kernel*. Scikit Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.sigmoid_kernel.html?highlight=sigmoid+kernel

[32] Department of Computer Science and Engineering JCDM College of Engineering, Sirsa(HR) (https://ijcset.net/docs/Volumes/volume7issue9/ijcset2017070901.pdf)

[33] Chapter 15 (https://www.cse.usf.edu/~r1k/MachineVisionBook/MachineVision.files/MachineVision_Chapter15.pdf)

[34] *Teorema de Bayes*. (s. f.). ConceptoABC. Recuperado 30 de octubre de 2022, de https://conceptoabc.com/teorema-de-bayes/

[35] scikit-learn developers (BSD License). (2007a). *Naive Bayes*. Scikit Learn. https://scikit-learn.org/stable/modules/naive_bayes.html?highlight=naive+bayes

[36] Jiménez Rodrígue, E. (s. f.). *Enfoque Probabilístico Bayesiano*. Jupyter. Recuperado 30 de octubre de 2022, de http://localhost:8888/notebooks/Desktop/Modelos%20probabilisticos/Enfoque%20Bayesiano/2-enfoque-bayesiano.ipynb

[37] ¿Qué es el Proceso de ciencia de datos en equipo (TDSP)? (s. f.). Microsoft Learn website. Recuperado 29 de octubre de 2022, de https://learn.microsoft.com/es-es/azure/architecture/data-science-process/overview

[38] World Bank Open Data. (s. f.). The World Bank. Recuperado 29 de octubre de 2022, de https://data.worldbank.org/

[39] *Lista de 33 países del Primer Mundo – Definición, características y ejemplos*. (2019). https://trabajoypersonal.com/paises-del-primer-mundo/