Análisis y predicción de la deserción de empleados. Un caso de estudio en la industria de software colombiana.

Analysis and prediction of employee Churn rate. A case of study in the Colombian software Industry.

Eliana Marcela Sierra Buriticá

TESIS

Asesor, docente

Paula María Almonacid Hurtado

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

Análisis y predicción de la deserción de empleados. Un caso de estudio en la industria de software colombiana.

Eliana Marcela Sierra Buriticá
Facultad de Ciencias, Universidad Eafit, Medellin, Colombia

Resumen

El objetivo de este estudio es realizar el análisis y la predicción de la deserción de empleados de una compañía de software en Medellín, con base en una base de datos privada que contiene 19 características de 1497 trabajadores, donde 900 están activos en la empresa y el restante han abandonado su puesto de trabajo.

En primer lugar, se realizó un análisis descriptivo y exploratorio, donde se encontró que había algunas variables que no aportaban información al modelo, como por ejemplo: Tipo de identificación, fecha de inicio de contrato, entre otras, también en esta parte se identificó la correlación de algunas variables y se procedió a eliminarlas del conjunto de características descriptivas del problema, ya que al dejarlas sería dejar información redundante en el modelo. En segundo lugar, se entrenarónn 4 modelos de machine learning (Niave Bayes, Random Forest, Decision Tree, Logistic Regression) y se compararon los resultados obtenidos por cada uno, para así encontrar el que mejor se ajusta a la problemática de la deserción laboral, en este paso se encontró que el mejor clasificador de machine learning es un árbol de decisión (Decision Tree) con 14 capas, ya que métricas como su curva de aprendizaje y curva de ROC dieron resultados superiores al de los otros dos modelos entrenados.

Palabras claves: Deserción, Industria de software, Análisis exploratorio, Algoritmos de Machine Learning.

1. Introducción

Hoy en día existen muchas empresas prestadoras de servicios tecnológicos, las cuales a través de soluciones TI crean experiencias sobresalientes en todos los canales digitales. Los servicios más prestados por estas compañías son la omnicanalidad, servicios en la nube, innovación digital, integración API, micro-servicios SOA, entre otros, que incrementan la eficiencia en los diferentes procesos de las empresas.

En su generalidad, estos servicios son prestados principalmente por ingenieros y desarrolladores de software, entre otras profesiones afines. Dicho lo anterior y por estadísticas del mercado, se puede concluir que estas profesiones están siendo muy demandadas por empresas locales y extranjeras que buscan fuerza laboral local, y esto está generando gran rotación de personal en compañías prestadoras de servicios tecnológicos (Oscar Fuentes, 2022).

La alta tasa de rotación de empleados es un tema que presenta diferentes efectos negativos en las organizaciones. En primer lugar, la dificultad de encontrar un reemplazo apropiado y, especialmente para trabajadores con abundante experiencia y conocimientos avanzados, como por ejemplo, profesionales expertos en especificas tecnologias que cada vez están siendo altamente demandadas por el mercado, otra de las razones es el desgaste que genera para el personal de recursos humanos, ya que esto aumenta su flujo de contratación, y en consecuencia de esto se aumenta el costo de adquisición lo que conlleva a que el capital de trabajo de las organizaciones requiera un esfuerzo adicional. También se puede identificar la distribución injusta de la carga laboral a los empleados vigentes; Como consecuencia de lo anteriormente expuesto surge la importancia de tratar de predecir el número de renuncias futuras que se pueden presentar en una organización o empresa.

La minería de datos se reconoce por su modelo de predicción teniendo como objetivo descubrir tendencias o patrones ocultos en un gran conjunto de datos (Alamsyah and B. Nurriz, 2017). Hay muchas técnicas en la minería de datos y el modelo de clasificación es uno de los modelos que se utiliza ampliamente en muchas áreas para encontrar el análisis de tendencias y la planificación futura. Por lo tanto, los modelos de clasificación de minería de datos se utilizan en este estudio (A. M. Florence.T and M. Savithri.R, 2013).

El objetivo principal de este trabajo es crear un modelo de clasificación que logre predecir el comportamiento de deserción de empleados para una empresa de servicios tecnológicos. Este modelo de clasificación se conoce como aprendizaje supervisado, donde el nivel de la clase o el objetivo de la clasificación ya se conoce (Jantan, A. R. Hamdan and Z. A. Othman,2011).

En este trabajo se entrenaron cuatro modelos de clasificación de los cuales tambien se midió su rendimiento, estos modelos son: Naive Bayes, Árbol de decisión, Random forest y Logistic Regression; Los cuales fueron elegidos tratando tomar diferentes enfoques de modelos paramétricos y no paramétricos con diferentes grados de flexibilidad e interpretabilidad con el fin de explorar las relaciones entre las variables.

2. Estado de arte

La inteligencia artificial se ha transformado; no tanto respecto a lo que podemos hacer con ella sino en lo que hacemos. (Joanna J. Bryson,2017) En un proceso de selección de empleados, Coninciden diferentes áreas de estudio, como lo son: talento humano, psicología, pedagogía, economía laboral entre otras (Andonova, S. Nikolova, Dimitrov , 2019). Todas estas áreas de estudio estan siendo un comeplento con la gestión de la información y los datos, surgiendo la necesidad de la implementación de la inteligencia artificial para facilitar el proceso de análisis de la información recabada. (Casella Hanaki, 2006). Reconocidos autores como Nigel Guenole, Jonathan Ferrar y Sheri Feinzig, hace años que son contratados por las principales empresas a nivel mundial para mejorar su desempeño organizacional. Estos han trabajado en exitosos proyectos relacionados con la retención de empleados, el compromiso laboral, la retroalimentación del desempeño y la cultura organizacional. Aunque en los departamentos de recursos humanos, esta tecnología ya ha sido utilizada con diferentes fines, a lo largo de este trabajo nos centraremos específicamente en la inteligencia artificial aplicada a la retención de empleados. Si bien se han realizado muchos estudios al respecto, estos difieren generalmente en cuanto a los resultados. En un estudio específico en el que se intenta predecir la deserción laboral utilizando los métodos de Random Forest, Support Vector Machines y árboles de clasificación, se concluyó que el método que obtiene mejores resultados tanto a nivel de tasa de aciertos, como en sensibilidad y especificidad, es el algoritmo

de Random Forest (Jain, P. K., Jain, M. Pamula, R. 2020). Sin embargo,según,(Saradhi Palshikar, 2011) define que el modelo de maquina de soporte vectorial tiene mejor performance que el de Naive Bayes y Random Forest. Por otra parte, (Bhuva Srivastava, 2018) determina que el análisis discriminante lineal, conocido como LDA, es el método que obtiene mejores resultados, seguido por la regresión logística, si se toma como referencia la tasa de aciertos. Por otra parte, otro estudio llevado a cabo por Mohbey confronta el rendimiento de cinco algoritmos de machine learning diferentes y finaliza concluyendo que el que obtiene mejores resultados es el método de regresión logística (Mohbey, 2020). Por tanto, se puede concluir que los resultados no son consistentes. La literatura nos señala que este comportamiento se debe principalmente a que la precisión con la que se acierta la deserción de empleados depende directamente de los datos y del modelo utilizado para la predicción (Jain, Jain, Pamula, 2020). Esto quiere decir que el método a utilizar depende de cada conjunto de datos seleccionado y, por lo tanto, no se pueden universalizar los resultados. Por lo anterior cabe la importancia a la hora de la selección de las variables, conocer cual es su comportamiento y explicación de la variable obejtivo con el fin de aunmentar el poder predictivo de los modelos a implementar (Najafi-Zangeneh,Shams-Gharneh, Arjomandi-Nezhad, Zolfani, 2021).

En el desarrollo de proyectos de minería de datos se usan diferentes metodologías, sin embargo, para el desarrollo de este proyecto se trabajará con la metodología CRIP-DM, ya que sigue una serie de pasos que nos permite el desarrollo del proyecto de forma ordenada y clara.

En este estudio se pretende contribuir al análisis y predicción de la deserción de empleados en la industria del software colombiana, considerando los grandes costos, ineficiencias e inestabilidad que dicho problema le ha generado a muchas de las industrias en este sector". Para esto se llevó a cabo un proceso de descripción estadísica univariada y multivariada de las características objeto de estudio para luego implementar diferentes tipos de modelos que describieran adecuadamente la relación en las variables y finalmente logar mejores predicciones de la variable de interés.

3. Marco teórico

3.1. Técnicas empleadas para la predicción de deserciones de empleados

Los modelos de machine learning se pueden dividir en dos grandes categorías: supervisados y no supervisados. La principal diferencia entre estas dos categorías recae en los datos de entrenamiento.

Los modelos supervisados necesitan una variable objetivo cuyos valores son conocidos, de modo que el modelo pueda disponer de la información correcta para entrenarse y luego reproducir los resultados para las observaciones de las cuales se desconoce este valor. A diferencia de estos, los algoritmos de aprendizaje no supervisado no necesitan disponer de una variable respuesta en los datos de entrenamiento. En este caso, no se busca reproducir un resultado específico sino descubrir nuevos patrones durante el proceso de aprendizaje. A su vez, los modelos de aprendizaje supervisado se dividen en dos grandes grupos: de regresión y de clasificación. En nuestro caso, dado que se intenta predecir si un empleado específico va a abandonar una empresa o no, se va a trabajar con modelos de aprendizaje supervisados para variable de clasificación binaria. Este tipo de problemática es bastante popular ya que las técnicas de clasificación representan aproximadamente el 70 % de los problemas de machine learning, siendo la mayoría de estos problemas de clasificación binaria (Ponnuru, Merugumala, Padigala, Vanga, Kantapalli, 2020). Algunos de los principales algoritmos aplicables a los problemas de clasificación binaria son la regresión logística, redes neuronales, árboles de clasificación, modelos basados en árboles como el Bagging, Random Forest, Gradient Boosting o Extreme Gradient Boosting, como también el algoritmo de Support Vector Machines. Todos estos se describen y se aplicarán a continuación para intentar obtener aquel que mejor prediga la decisión de abandono de los empleados en una empresa. A su vez, además de usar los algoritmos tradicionales, se utilizarán métodos de ensamblados para probar si, al combinar los distintos algoritmos, se logra aumentar la capacidad predictiva del modelo.

3.1.1. Naive Bayes

Naive Bayes es una técnica de clasificación popular que clasifica los ejemplos en función de la probabilidad de que ocurran (K.P.Murphy 2006). A menudo funciona muy bien para conjuntos de datos complejos que son muy difíciles de aprender utilizando los algoritmos de aprendizaje tradicionales.

Según el autor Keynia, 2017, este algoritmo es una técnica de clasificación conocida en la predicción de deserción, principalmente por su simplicidad y eficiencia en los resultados. Para el caso de la deserción de empleados se basa en utilizar la regla de Bayes que calcula las probabilidades de encontrar abandono y no abandono dado un registro de un empleado, teniendo en cuenta un clasificador bayesiano que permite fácilmente tener el supuesto de que cada una de las variables son independientes y distribuyen normal.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$A, B = \text{Eventos}$$

$$P(A|B) = \text{Probabilidad de A dado B}$$

$$P(B|A) = \text{Probabilidad de B dado A}$$

$$P(A), P(B) = \text{La probabilidad de los eventos independientes.}$$

3.2. Árboles de clasificación

Esta es una de las técnicas de aprendizaje más populares enlas máquinas de aprendizaje. (D. Gupta, D. S. Kohli, and R. Jindal,- 2011) Es el algoritmo de aprendizaje de referencia de los árbol de decisión que a menudo se compara con los nuevos algoritmos que se están desarrollando. Aquí el algoritmo de aprendizaje es utilizado que es una versión avanzada del árbol de decisión tradicional algoritmos de aprendizaje. Los nodos junto con los bordes son las serie de condiciones y las hojas son las etiquetas de clase. (Tambde, A. Motwani, D.2019)

Los árboles de decisión son modelos predictivos que, a partir de un conjunto de reglas sencillas que se van aplicando de forma jerárquica y secuencial, logran dividir las observaciones en función de sus atributos, generando una segmentación para así poder predecir el valor de la variable objetivo. Los árboles de clasificación corresponden al tipo de árboles de decisión para los cuales la variable objetivo es categórica, en nuestro caso, específicamente binaria. De esta forma, se obtienen una serie de segmentos (llamados nodos) que contienen subconjuntos de la muestra. El segmento original contiene a la totalidad de los datos

y recibe el nombre de nodo raíz. “Una vez obtenida la segmentación óptima, entendiendo así aquella que da lugar a nodos con comportamiento homogéneo respecto a la variable objetivo y heterogéneos entre ellos, se asigna un valor de predicción a aquellos nodos que no tienen sucesores (y que reciben el nombre de hojas) de forma que todas las observaciones pertenecientes a dicha hoja serán predichas a partir de dicho valor” (Calviño, 2020). Las principales ventajas de este tipo de modelos son que los árboles son fáciles de interpretar, no necesitan cumplir supuestos en cuanto a la distribución de los datos, no requieren mucha limpieza o preprocesado de los datos, no se ven muy influenciados por valores atípicos, y son capaces de seleccionar predictores de forma automática. Sin embargo, la principal desventaja recae en la calidad de los resultados. Los modelos basados en un único árbol suelen tener resultados bastante inferiores a otros modelos ya que tienden a sobreajustar y generar una alta varianza. Si bien no se pudo mejorar la capacidad predictiva de los árboles, para saltar esta limitación se crearon técnicas de modelado más complejas que se generan a partir de la combinación de múltiples árboles. Entre ellas se destacan las técnicas de Bagging, Random Forest y Gradient Boosting, las cuales se explicarán a continuación (Amat Rodrigo, Ciencia de datos, 2020).

3.3. Random forest

Es una de las técnicas de aprendizaje por conjuntos que consta de varios árboles de decisión en lugar de una sola decisión (un árbol) para la clasificación. Al clasificar todos los árboles en el bosque aleatorio da como resultado una clase a un ejemplo desconocidoy, la clase que tenga el máximo de votos se asignará a la desconocida.(Breiman, L. 2001).

El modelo de Random Forest consta de una adaptación del modelo Bagging. El funcionamiento de ambos es el mismo, por lo cual este también constituye una técnica de ensamblado de algoritmos de regresión y clasificación, siendo los árboles de decisión el método más comúnmente utilizado. La diferencia entre los mismos recae en la construcción de los árboles o el algoritmo con el que se decida trabajar. Además de utilizar diferentes muestras del conjunto de entrenamiento para su construcción, el modelo de Random Forest introduce un proceso aleatorio de selección de variables. (Computational Science and Its Applications,2006) En el caso de los árboles, cada vez que una división va a tomar lugar, para la creación de cada nodo, se elige la variable que aporte mayor utilidad al modelo a partir de un subconjunto sorteado aleatoriamente del conjunto original de variables. Esta técnica agrega mayor aleatoriedad al proceso lo que permite reducir la correlación entre los árboles creados y obtener modelos con menor varianza (Boehmke Greenwell, 2020). Los parámetros por controlar son los mismos que para el método de Bagging solo que en este caso se debe definir además el tamaño de la muestra de variables que serán sorteadas para la construcción de cada nodo. Al ser este modelo una adaptación al de Bagging, las ventajas y desventajas son por lo tanto las mismas, solo que este algoritmo, al introducir mayor aleatoriedad en cuanto a la construcción de los árboles, permite reducir generalmente en mayor medida la varianza.

3.4. Regresión logística

Los modelos de regresión permiten analizar relaciones entre variables dependiente y un conjunto de variables explicativas, pudiendo determinar el impacto que genera un cambio en las variables explicativas sobre la variable dependiente. La regresión logística se centra específicamente en problemas de predicción en los cuales la variable objetivo es binaria, de modo que se pueda crear un modelo lineal que prediga la probabilidad de que la variable objetivo tome valor 1. Se le asigna valor 1 al evento de interés siendo, en este caso, la decisión de abandono del puesto de trabajo. La principal limitante de utilizar un modelo de regresión lineal para predecir la probabilidad es que pueden haber resultados que tomen valores fuera del intervalo (0,1), lo cual no sigue la lógica del concepto de probabilidad. Para saltar este obstáculo se utilizará como función de enlace la función de distribución logística, también conocida como función sigmoidea

La función sigmoidea aparenta una curva en forma de S que puede tomar cualquier valor entre 0 y 1. La ecuación viene dada por $f(x) = \frac{1}{1+e^{-}x}$ donde x es un número real. Cuando x tiende a menos infinito, el cociente tiende a 0, mientras que cuando x tiende a infinito, el cociente tiende a 1. (Rodríguez, 2018)

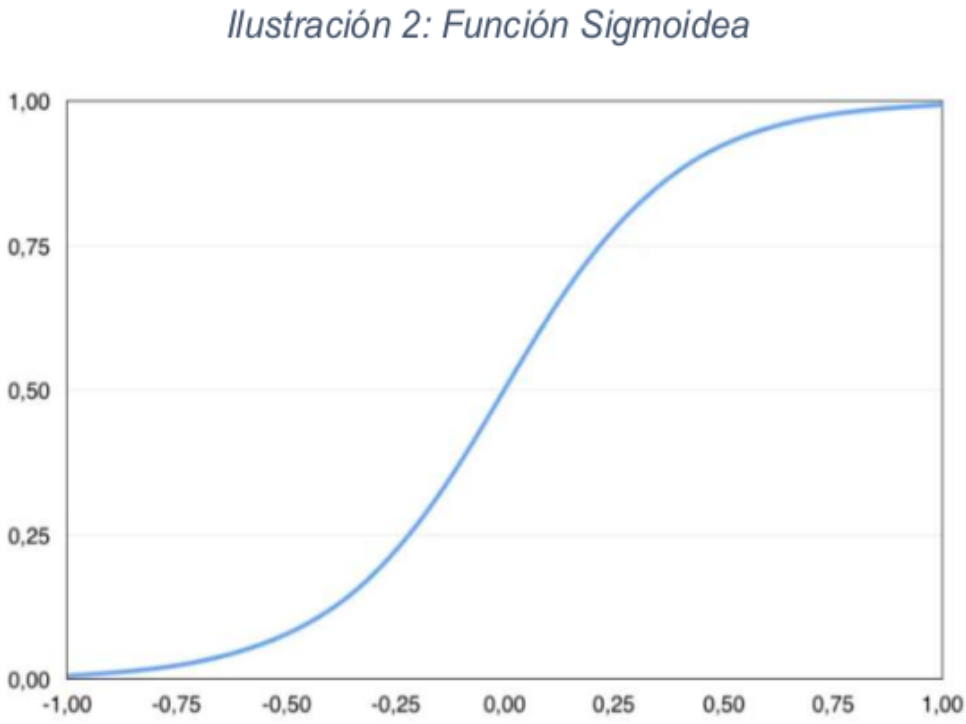

Figura 1: Función Sigmoidea Fuente: Machine Learning para estimación de riesgo

Tras aplicar la función sigmoidea se obtiene que la probabilidad de que ocurra el evento (Y=1), viene dada por: $p1 = P(Y = 1|x_1, x_2, ..., x_m) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+...+\beta_m x_m)}}$, mientras que la probabilidad del no evento viene dada por: $p0 = 1 - p1$.

Utilizando estas ecuaciones, se puede calcular el logaritmo de la razón de probabilidad, de modo que la ecuación quede escrita igual que la de un modelo de regresión lineal, con las probabilidades restrictas al intervalo (0,1): $log\frac{p1}{1-p1} = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$. Para poder interpretar los parámetros del modelo de regresión logística resulta necesario explicar dos conceptos: odds y oddsratio. "Los odds asociados a un determinado suceso A se definen como el cociente entre la probabilidad de que ocurra el suceso y la probabilidad de que no ocurra" (Calviño, 2020). Por lo cual, la función de logaritmos definida anteriormente corresponde al logaritmo de los odds del suceso de interés (Y = 1). El odds-ratio, por otra parte, viene definido "como el cociente entre los odds de un suceso bajo una determinada condición y el odds de ese mismo suceso bajo otra condición" (Calviño, 2020). Esta fórmula nos permitirá evaluar el efecto de dichas condiciones sobre las probabilidades del suceso, pudiendo así interpretar los parámetros. Las principales ventajas de utilizar un modelo de regresión logística son que este es un modelo sencillo de estimar y fácil de comprender. Recordemos que este modelo, además de predecir el valor de la variable objetivo, nos permite realizar una descripción de los datos y explicar las relaciones que puedan existir entre la variable objetivo y las variables input del modelo, independientemente de la clase a la que pertenezcan estas. Si bien la sencillez del modelo representa una ventaja, también puede ser considerada una desventaja. Este modelo no logra identificar relaciones no lineales o complejas entre las variables. Además, suele verse muy influenciado por la existencia de valores atípicos.

3.5. Evaluación de modelos

Para evaluar el desempeño de los diferentes modelos implementados se obtuvieron diferentes métricas pertinentes para el caso de clasificación. En primer lugar, se calculó la matriz de confusión. 2

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 2: matriz de confusión: Fuente RPus-Evaluacíon de modelos de predicción

- VP: Cantidad de predicciones positivas verdaderas.
- FN: Cantidad de predicciones que eran positivas y son predichas como negativas.
- FP: Cantidad de predicciones que eran negativas y son predichas como positivas.
- VN: Cantidad de predicciones negativas verdaderas.

Las métricas más comunes son; Recall, accuracy, precisión y la especificidad. Estas son calculadas a partir de los números de la matriz de confusión.

1. El accuracy es el porcentaje de elementos clasificados correctamente en comparación con el número total de artículos.
$$\text{Accuracy} = \frac{VP + VN}{VP + VN + VN + FN}$$
2. El Recall mide la proporción de verdaderos positivos que el modelo ha clasificado en función del número total de valores positivos.
$$\text{Recall} = \frac{VP}{VP + FN}$$
3. La precisión representa la proporción de verdaderos positivos que son realmente positivos en comparación con el número total de valores positivos predichos.
$$\text{Precisión} = \frac{VP}{VP + FP}$$
4. Es el número de ítems correctamente identificados como negativos fuera del total de negativos.
$$\text{specificity} = \frac{VN}{VN + FP}$$

4. Metodología CRISP-DM

Para el desarrollo de este proyecto de grado se siguió la metodología CRISP-DM, (Cross Industry Standard Process for Data Mining), la cual se compone de una estructura basada en 6 etapas, donde unas de estas etapas resultan siendo bidireccionales. Es decir, en cualquier fase de la metodología se puede reversar a la fase anterior para llevar a cabo revisiones extemporáneas, por lo cual tampoco es necesario seguir estrictamente un orden en las diferentes fases. Por último, la metodología CRISP-DM ha tenido mayor fuerza en los proyectos de minería de datos según encuestas realizadas y publicadas por kdnuggets.com.

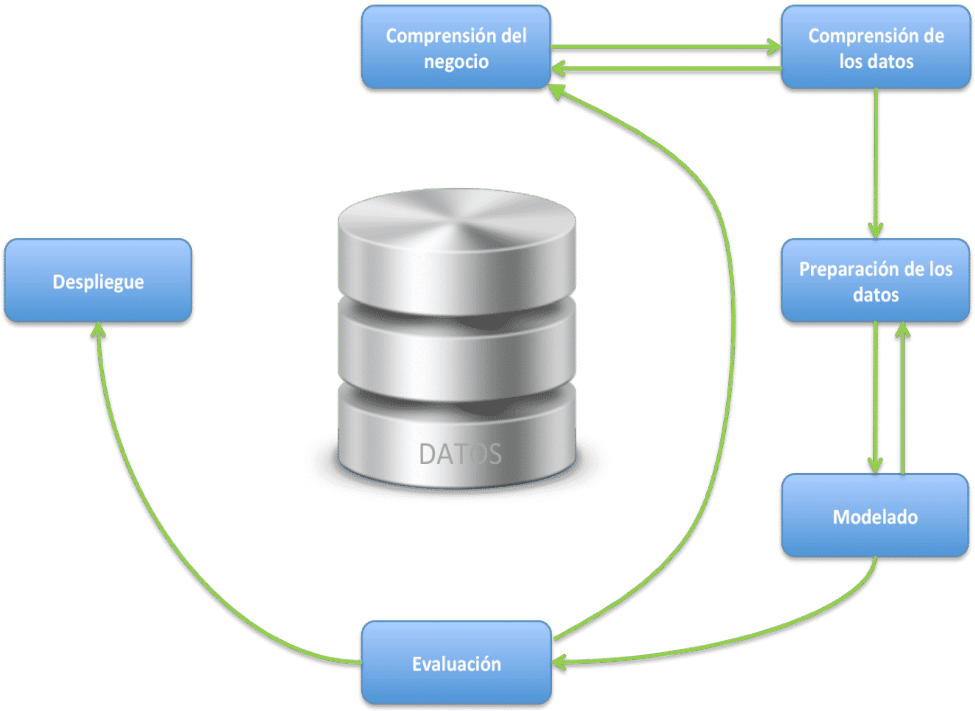

Figura 3: Secuencia de la metodología

4.1. Comprensión del negocio

Esta empresa del sector tecnológico se distingue por: considerar a los empleados su principal activo, apoyándolos con capacitaciones constantes según las nuevas tecnologías que estén siendo demandadas por el mercado; también tiene un enfoque humano muy fuerte, esto quiere decir que los empleados son libres de expresarse y tienen como principio fundamental la escucha de nuevas ideas y/o opiniones. Aun así y pese a lo anteriormente mencionado, los esfuerzos de las empresas del sector de tecnología, como la analizada en este estudio, sufren constantemente con la deserción de sus empleados pues las ofertas que están en el mercado cada vez son más atractivas para los profesionales de este sector e indudablemente desertan para aceptar nuevas ofertas con mejores condiciones.

A partir de los objetivos planteados en el proyecto se busca unificar toda la información posible para el desarrollo del proyecto, esta información puede llegar a ser suministrada por el área de talento de la compañía a trabajar, quien inicialmente es la parte que nos entrega la base de datos, también se continuará realizando una investigación profunda de como se han abordado este tipo de investigaciones.

4.2. Comprensión de los datos

4.2.1. Descripción de los datos

La base de datos con la que se trabajará contiene 19 características de cada uno de los 1497 empleados (registros) y cada registro se encuentra con la información completa (no tiene valores nulos). El nombre, descripción y tipo de variable está en el cuadro 1.

S.NO	Nombre	Tipo	Descripción
1	tipo_identificación	categórica	Nombre del documento del empleado.
2	identificación	categórica	Primeros dos dígitos del documento
3	país	categórica	País de residencia del empleado
4	sexo	categórica	Sexo al que pertenece
5	estado_civil	categórica	Estado civil en el que se encuentra el empleado
6	referido	categórica	Si fue recomendado o no a la hora de ingresar como empleado
7	tipo_contrato	categórica	Tipo de contrato con el que fue ingresado.
8	fecha_inicio_contrato_actual	Date	Fecha en la que comienza a ser parte de la empresa.
9	fecha_nacimiento	date	Fecha de nacimiento del empleado.
10	salario_básico_ordinario	numérica	Salario que recibe mensualmente el empleado de forma común.
11	salario_básico_integral_	numérica	Salario que incluye, además del salario ordinario, incluye prestaciones sociales y extralegales, los recargos por trabajos extras, nocturno, dominical y festivo.
12	no_constitutivos_salario	numérica	Pagos que NO hacen parte de los cálculos de las prestaciones sociales, aportes parafiscales, ni de la seguridad social.
13	salario_total	numérica	Cifra total que recibe el empleado.
14	salario_diario	numérica	Valor al que equivale un día de trabajo del empleado.
15	moneda	categórica	Moneda en la que es pagado el empleado.
16	edad	numérica	Edad con la que ingresó el trabajador
17	permanencia	numérica	Tiempo de permanencia del empleado(años)
18	análisis_permanencia_#	categórica	Rango de permanencia que tiene el empleado
19	Deserción	categórica	Si se ha retirado de trabajar(Variable objetivo)

Cuadro 1: Descripción de las variables de la base datos

4.2.2. Exploración de los Datos

Luego de tener una descripción de las variables, se procede a realizar una exploración de ellas, lo cual implica realizar gráficos de frecuencias, distribución, entre otros. La función principal de este análisis es determinar la consistencia y completitud de los datos. Mediante consultas en python, se han obtenido los datos necesarios para realizar las gráficas que se explican a continuación.

En el gráfico 4 se muestran las frecuencias de la variable sexo, permanencia, estado civil, país de origen, tipo de contrato y deserción (variable objetivo). Se logra ver que la mayoría de los empleados son de sexo masculino, representando más del 70 % de los cargos de la compañía. También se evidencia que gran parte de los trabajadores están solteros, ya que en esta categoría hay más de 900 personas, la variable permanencia muestra la alta rotación de empleados que tiene la empresa, debido a que la mayoría de trabajadores tienen el nivel de permanencia más bajo. La variable objetivo (deserción) está mostrando que aproximadamente 900 empleados están activos y al rededor de 500 han dejado su puesto de trabajo, finalmete se obtiene que las mayoría de los trabajadores son de Colombia y tiene un un contrato a termino indefinido.

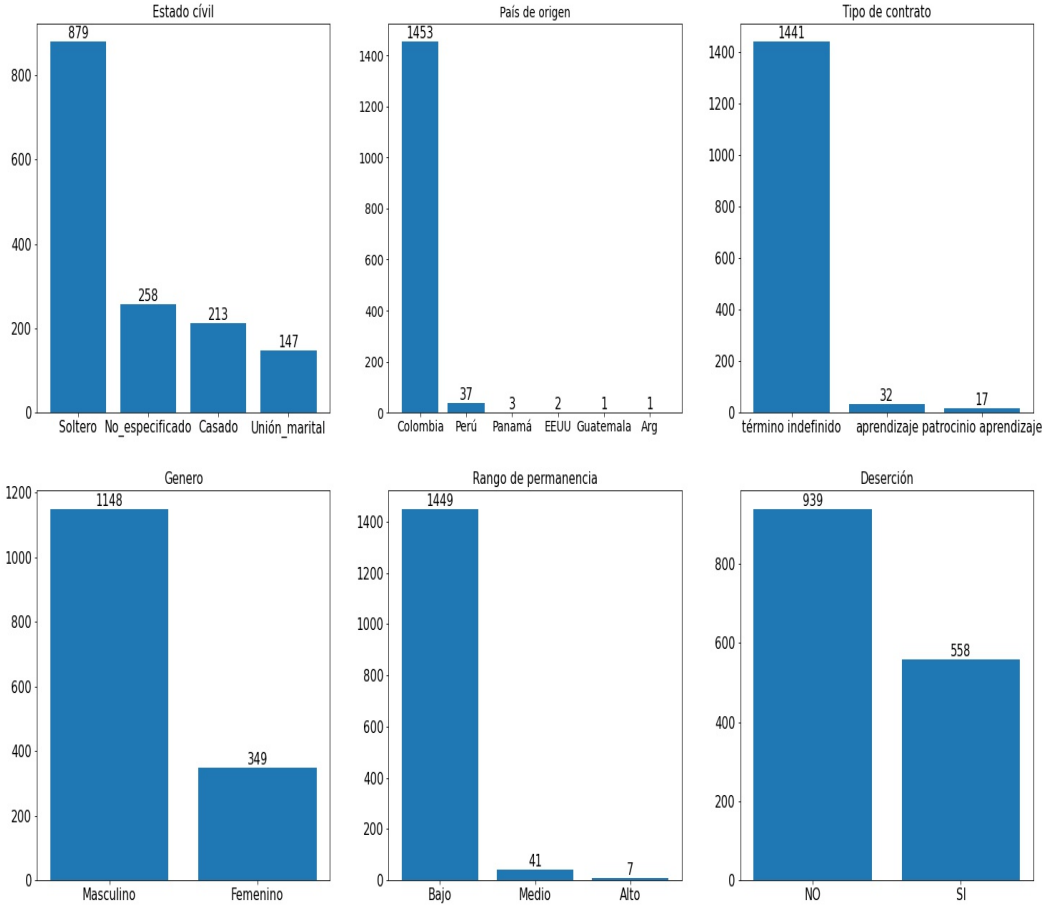

Figura 4: Gráfico de frecuencias: Fuente propia

En el gráfico 5 se logra ver las diferencias significativas visuales que existen en el salario total, respecto al género, el estado civil, el análisis de permanencia y referido. En la parte superior izquierda de la gráfica se logra ver que no hay una diferencia significativa en el salario total de los hombres y las mujeres que están en la compañía respecto a los/las que han dejado su cargo. También se logra ver que en las mujeres existe una mayor variabilidad en su salario total, tanto para las que están en la empresa como para las que han abandonado su puesto de trabajo.

En la esquina superior derecha se ve la diferencia significativa que hay en el salario total de las personas que están casadas y aún están activos en sus cargos, respecto a los que no han abandonado, además esta misma diferencia también se presenta en las personas que tienen una unión marital con una dispersión más baja respecto a los que están casados. Para los trabajadores que no especificaron su estado civil y para los solteros, se nota que no hay una diferencia significativa en los salarios totales.

En la esquina izquierda inferior se ve la diferencia que hay en los salarios totales de los trabajadores que siguen trabajando respecto a los que se han ido y se encuentra clasificados con el estado de permanencia más alto. por lo tanto, esto lleva a presentar una alta variabilidad en sus salarios totales. Finalmente, se observa que no existe diferencia en los salarios totales de los trabajadores que fueron recomendados por algun empleado de la empresa.

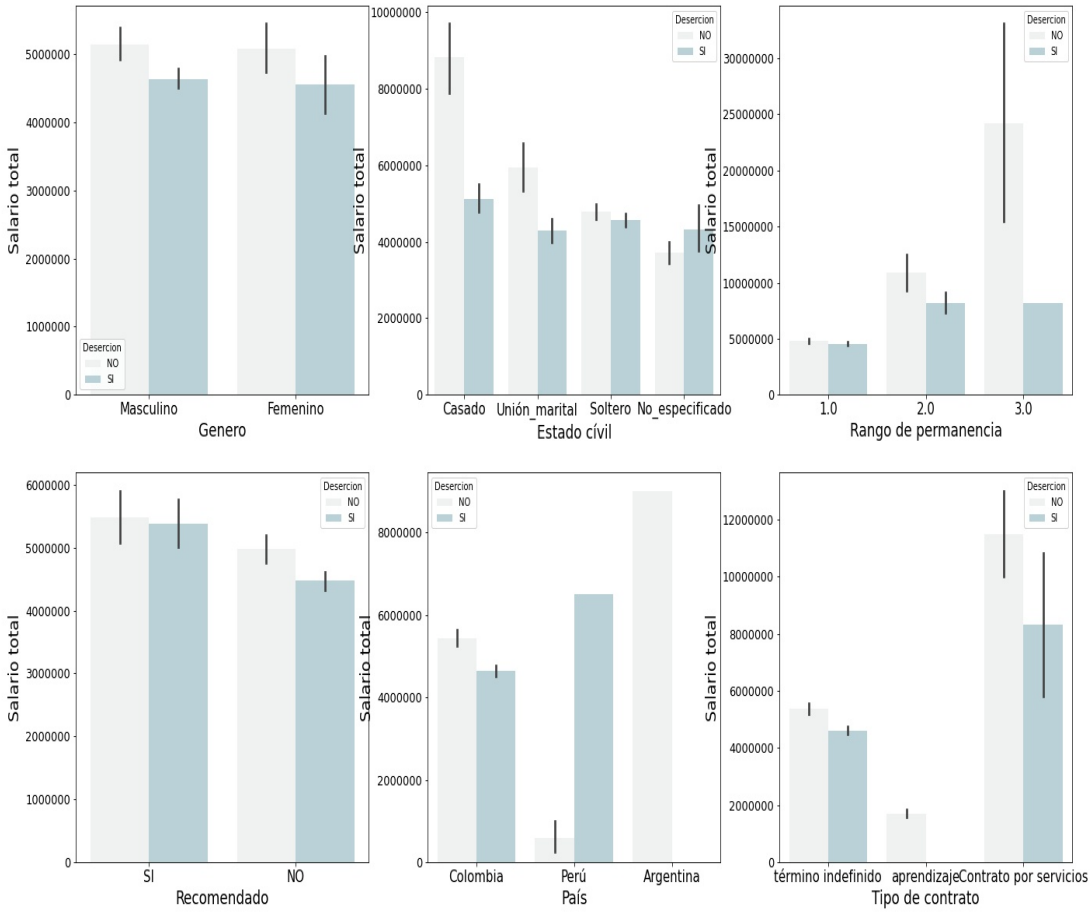

Figura 5: Diferencias significativas visuales: Fuente propia

En el gráfico 6 muestra las distribuciones de densidad del salario total, salario diario, salario básico integral, salario básico ordinario, tiempo de permanencia y edad de cada uno de los 1497 trabajadores, discriminado por la variable de interés (deserción). Del gráfico de la esquina izquierda superior se puede visualizar que el promedio del salario total de los trabajadores que se han retirado es similar al de los que se encuentran activos en la compañía, pero estos contienen una mayor variabilidad. Para el salario diario, salario básico integral y salario básico ordinario se presenta el mismo patrón de la variable salario total, teniendo los promedios similares sin importar si continúa o no activo en la empresa, no obstante, la variable salario básico ordinario es la que presenta una mayor dispersión o variabilidad. Lo anterior nos da indicios de una posible correlación entre dichas variables, lo cual no es positivo para el desempeño del modelo.

Del gráfico de la esquina superior izquierda refleja que el promedio de edad de las personas que han abandonado sus cargos es superior al de los que aún continúan trabajando, además se puede ver que la mayoría de trabajadores están entre 20 y 40 años de edad. Por otro lado, se tiene que el promedio de tiempo de permanencia de los trabajadores es de aproximadamente 1.5 años y la mayoría oscila entre 0 a 4 años; por lo que está mostrando la alta rotación de personal que contiene la compañía.

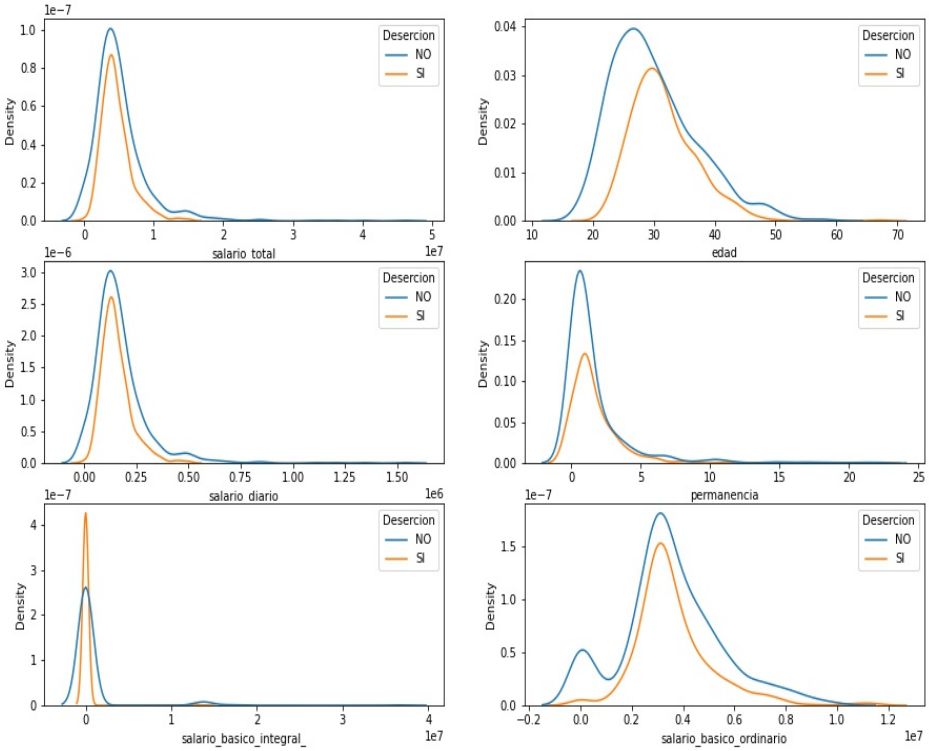

Figura 6: Gráfico de variabilidad: Fuente propia

4.2.3. Resumen númerico de los datos

En el cuadro dos se logra observar las medidas descriptivas de las variables continuas consideradas en los diferentes modelos que se entrenaron en el desarrollo del proyecto. Para la variable edad se tiene que su media es de 30 años con una desviación estándar de 6 aproximadamente, por lo que se puede decir que la edad de los trabajadores y ex trabajadores de la empresa varía de 24 a 36 años de edad, además se tiene que la mediana de los datos es igual a su media, debido a su baja variabilidad, por otro lado, se tiene que el 50 % de los trabajadores tienen una edad inferior a 30 años y finalmente se obtiene que la persona mayor en la empresa tiene 76 años de edad. La variable Salario total tiene una media de de 4.944.243 pesos Colombianos con una desviación estándar de 3.542.267 pesos, entonces el salario de los empleados está entre 1.401.976 y 8.486.510 pesos Colombianos, agregando a lo anterior se tiene que su mediana es de 4.200.000 pesos Colombianos, es decir; el 50 % de los trabajadores tiene un sueldo inferior o igual a 4.200.000 pesos Colombianos, por otra parte, se tiene que el salario total más alto es de 46 millones de pesos Colombianos.

La variable Tiempo de permanencia tiene una media de 1.6 años con una desviación estándar de 2.4 años, es decir; el tiempo de permanencia de un trabajador puede variar entre 7 meses y 4 años, el 50 % de los trabajadores permanecen en la empresa un tiempo de 0 a 12 meses. El máximo tiempo que tiene un trabajador en la compañía es de 22 años.

	Edad	Salario total/Pesos colombianos	Permanencia/años
mean	30	4.944.243	1.6
std	6.445.627	3.542.267	2.4
min	17	0	0
25 %	26	3.000.000	0.390
50 %	30	4.200.000	1
75 %	34	6.000.000	2.0
max	67	46.000.000	22

Cuadro 2: Resumen númerico de las variables continuas

4.3. Preparación de los datos

La limpieza/preprocesamiento de datos es el primer paso o enfoque en la construcción de un modelo predictivo y en el intento de obtener las características que dan cuenta de la clasificación de las etiquetas de las clases a predecir. En total hay 19 variables asociadas a cada empleado, variables como tipo de identificación, identificación, son variables que no aportan ninguna información al modelo. Fecha de inicio de contrato, es una variable que su información está contenida en permanencia, ya que esta mide el tiempo en años de duración en la compañía, fecha de nacimiento es una variable que su información está representada en la variable edad; las variables nombradas anteriormente son variables que se sacarán del conjunto de características que se usarán para montar el modelo.

Segun los resultados arrojados por la matriz de correlación, encontramos que: Las variables salario diario, salario no constitutivos y salario básico integral presentan unas altas correlaciones con salario total, lo cual no estarían entregando información adicional al modelo, por lo que incluirlas o no, no se verían afectos los resultados, es por eso que dichas variables no serán incluidas en el conjunto de características explicativas.

Para el caso de las variables categóricas, estas fueron trasformadas a sistema binario, cada categoría será una nueva variable que contendrá ceros y unos, donde el uno reprenda que si pertenece a esa categoría y el cero que no, para realizar este proceso fue usada la función get.dummis() de la librería pandas que está implementada en el lenguaje de programación python.

Finalmente, para el desarrollo de este modelo se quedó con 29 variables explicativas y 1497 registros.

4.4. Modelado y Optimización

Seguido de la exploración y preparación de los datos que se realizó en las etapas anteriores, se procederá a implementar cada uno de los modelos de machine learning propuestos para este problema (Decision Tree, Random Forest y Bayes Naive y Regresión Logistica).

Evaluar algoritmos con los mismos datos con los que se ha entrenado, conducirá a un sobre ajuste (overfitting). Para evitar esto, se dividirá el conjunto de datos en partes de entrenamiento y de prueba, para el entrenamiento se tomó el 70 % de los datos y para la prueba el valor restante (30 %).

4.4.1. Optimización

Para la implementación de los modelos mencionados anteriormente (Desicion Tree, Random Forest, Bayes Naive y Regresión Logistica) inicialmente se usarán las siguientes parametrizaciones: para Naive Bayes se tomó un var smothing de 5.3x10 a la 6, ya que es el valor que minimiza la función de perdida y a la vez maximiza el accuracy del modelo, dicha optimización fue desarrollada usando la función BayesSearchCv de la librería skopt del python, esta función hace una optimización bayesiana de los hyper parámetros, donde su método se optimiza mediante una búsqueda de validación cruzada sobre la configuración de los parámetros.

El algoritmo Decision Tree Classifier fue entrenado con un 12 capas (maxp depth = 12), para la obtención del valor de este parámetro fue usada la función fmin de la librería hyperopt del python. La función fmin es la función de optimización que itera sobre diferentes conjuntos de algoritmos y sus hiperperámetros y luego minimiza la función objetivo, El algoritmo de búsqueda a utilizar como búsqueda de manera aleatoria se usó una validación cruzada con 10 particiones (n splits = 10) y tamaño de partición 0.3, es decir el 70 % de los datos para entrenar y el 30 % para la validación del modelo. Como criterio de minimización se usó el coeficiente de Gini. El valor mínimo de la función de perdida es a su vez el inverso del máximo del accuracy del modelo, es decir; el valor máximo de desempeño del modelo, donde dicho valor fue de -0.9871, o sea que el accuracy máximo es de 0.9871.

El algoritmo Random forest fue entrenado con 5 árboles (n estimators = 5) y cada uno con 14 capas (max dept = 14), para la optimización de estos hyper parámetros fue usada la misma función que se implementó para el algoritmo anterior, con la misma validación cruzada y sus criterios de minimización, el valor obtenido que minimiza es de -0.9877, por lo que su nivel de desempeño general es de 0.9877 (Accuracy = 0.9877)

Finalmente, el algoritmo Linear Regresión fue entrenado en modo equilibrado (balanced), es decir, utiliza los valores de y para ajustar automáticamente los pesos inversamente proporcionales a las frecuencias de clase en los datos de entrada, además con un $C = 0{,}0001$, que representa el inverso de la regularización, los valores más pequeños especifican una regularización más fuerte finamente la función de optimización utilizada fue newtoncg, que es usada en problemas multiclase. Estos hiperparametros fueron optimizados usando la función GridSearchCV del python que se encuentra la librería sklearn.

4.5. Evaluación de la calidad de los modelos

Para cada uno de estos modelos se le medirán diferentes métricas(Accurracy, Recall, Specificity, Presicion, Curva de Roc y curvas de rendimiento) que se serán comparadas para poder elegir el mejor algoritmo o modelo de clasificación.

4.5.1. Matriz de confusión

En la siguiente gráfica se observan las cantidades de clasificaciones correctas e incorrectas de cada uno de los algoritmos entrenados. Los resultados que contienen, serán usados para hallar las métricas de rendimiento (Accurracy, Recall, Specificity, Presicion, Curva de Roc y curvas de rendimiento) que arrojan cada modelo y así poder compararlos. Por ejemplo, en la matriz de confusión de la esquina superior derecha; se ve que solamente obtuvo 2 falsos positivos, es decir, que las 2 personas no abandonaron el cargo y el modelo lo clasifico en el grupo de los que han dejado sus puestos de trabajo.

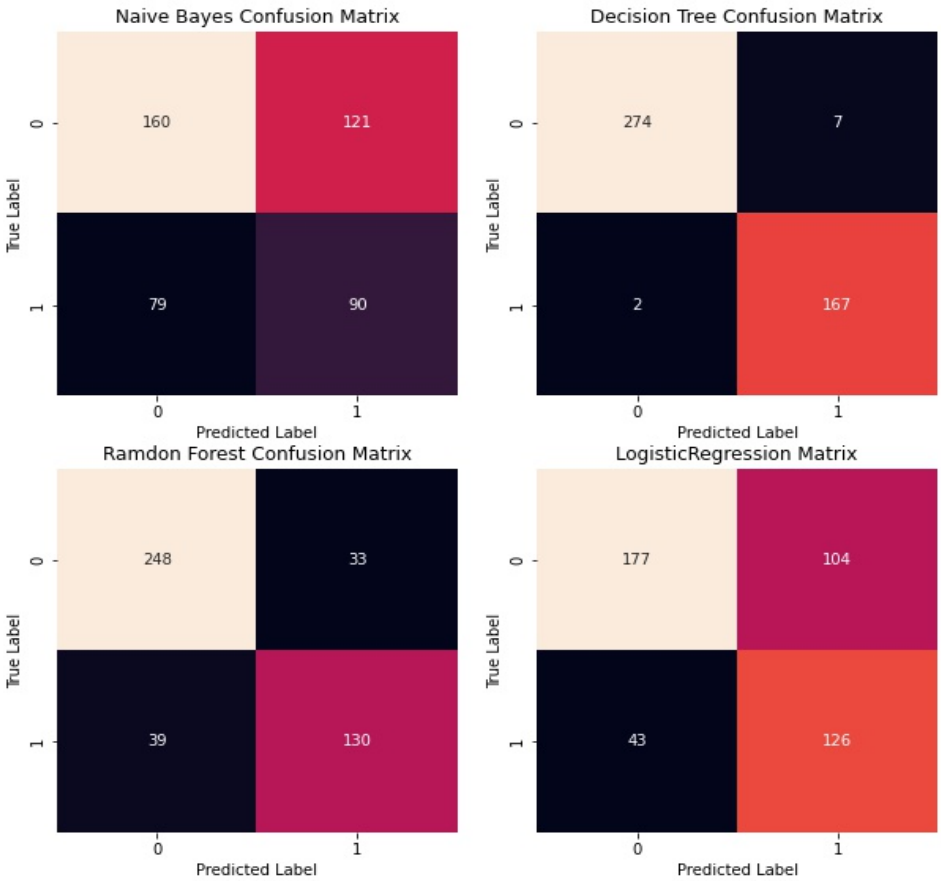

Figura 7: Matriz de confusión: Fuente propia

4.5.2. Comparación de rendimientos

En la siguiente gráfica se observa que los mejores resultados fueron obtenidos por el algoritmo Decision Tree, que fue entrenado usando los hiperparametros optimizados, por ejemplo, se puede ver que el accurracy que tiene es del 98 %, es decir; el modelo acertó el 98 % de las clasificaciones, no obstante, está métrica no es la más eficiente; dado que la información que muestra es muy generalizada, es por eso que se hacen los cálculos del Recall, Precisión y Especificidad. Cada una de estas, se centra en medir la proporción de elementos identificados como positivos o negativos, Por ejemplo, la Precisión obtenida por Decision Tree es del 96 %, o sea que obtuvo un 96 % de clasificaciones correctamente como positivas del total de clasificaciones positivas. También para las dos métricas restantes, Decision Tree generó mejores rendimientos respecto a Naive Bayes y Random Forest y linear Regresion; es decir, el mejor clasificador es Desición Tree; sin embargo, se realizará un análisis más profundo haciendo uso de la curva de ROC y las curvas de aprendizaje para validar esta información se mostrará a continuación.

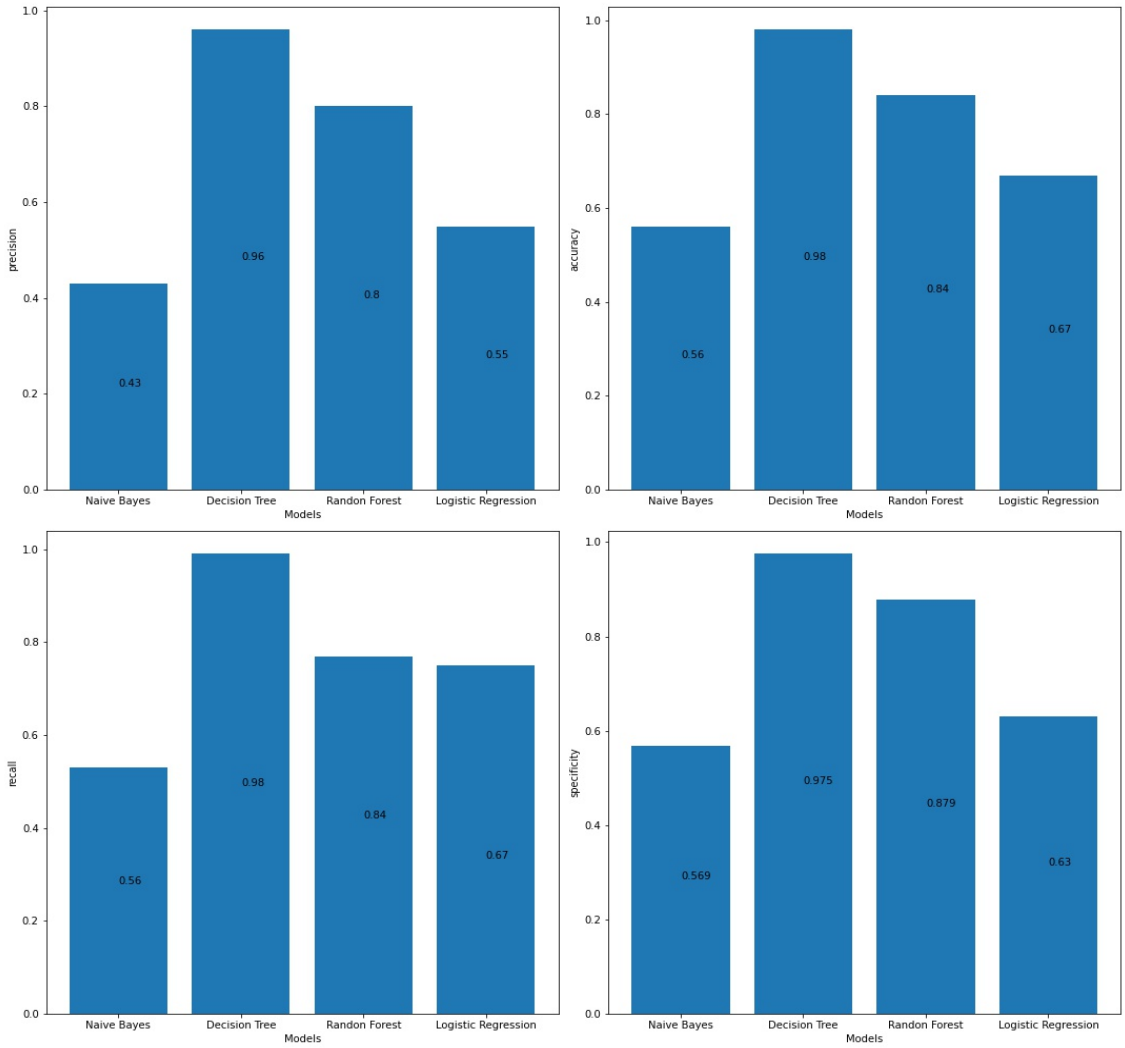

Figura 8: Comparación de rendimientos: Fuente propia

4.5.3. Análisis de las curvas de ROC

La curva ROC dice qué tan bueno puede distinguir el modelo entre dos cosas, por ejemplo, si un trabajador abandonaría el puesto de trabajo o no. Antes de continuar, refresquemos dos términos que son necesarios para entender la curva de ROC, como los son sensibilidad y especificidad.

La sensibilidad o recall, es la proporción de trabajadores que se identificaron correctamente que dejaron su puesto de trabajo, es decir verdadero positivo, sobre el número total de trabajadores que realmente abandonaron su cargo.

Por su parte, especificidad es la proporción de trabajadores que se identificaron correctamente por no dejar su trabajo, verdadero positivo, sobre el número total de trabajadores que no dejaron su trabajo.

Para trazar esta curva (ROC), en lugar de especificidad usamos "1 – especificidad". De esta forma si aumentamos la sensibilidad, la "1 – especificidad" también aumentará. A esta curva se le conoce como la curva ROC.

Si llevamos este término a la fórmula tendríamos algo así:

$$1 - \text{especificidad} = \frac{FP}{VN + FP}$$

Mientras que la especificidad nos da la tasa negativa verdadera y "1 – especificidad" nos da la tasa positiva falsa. Por lo que la curva ROC está definida por la sensibilidad que es la tasa de verdadero positivo y "1 – especificidad" es la tasa de falso positivo.

Área bajo la curva (AUC) El AUC es el área bajo la curva ROC. Este puntaje nos da una buena idea de qué tan bien funciona el modelo.

En la figura 9 se ve que el AUC asociado a Decision Tree es de 0,98, lo que significa que hay 98 % de probabilidad de que el modelo pueda distinguir entre clase positiva y clase negativa. También se observa que dicho AUC es el más elevado entre los cuatro algoritmos de clasificación, por lo que hay evidencia que Decision Tree es más acto para predecir si un trabajador abandona o no su puesto de trabajo.

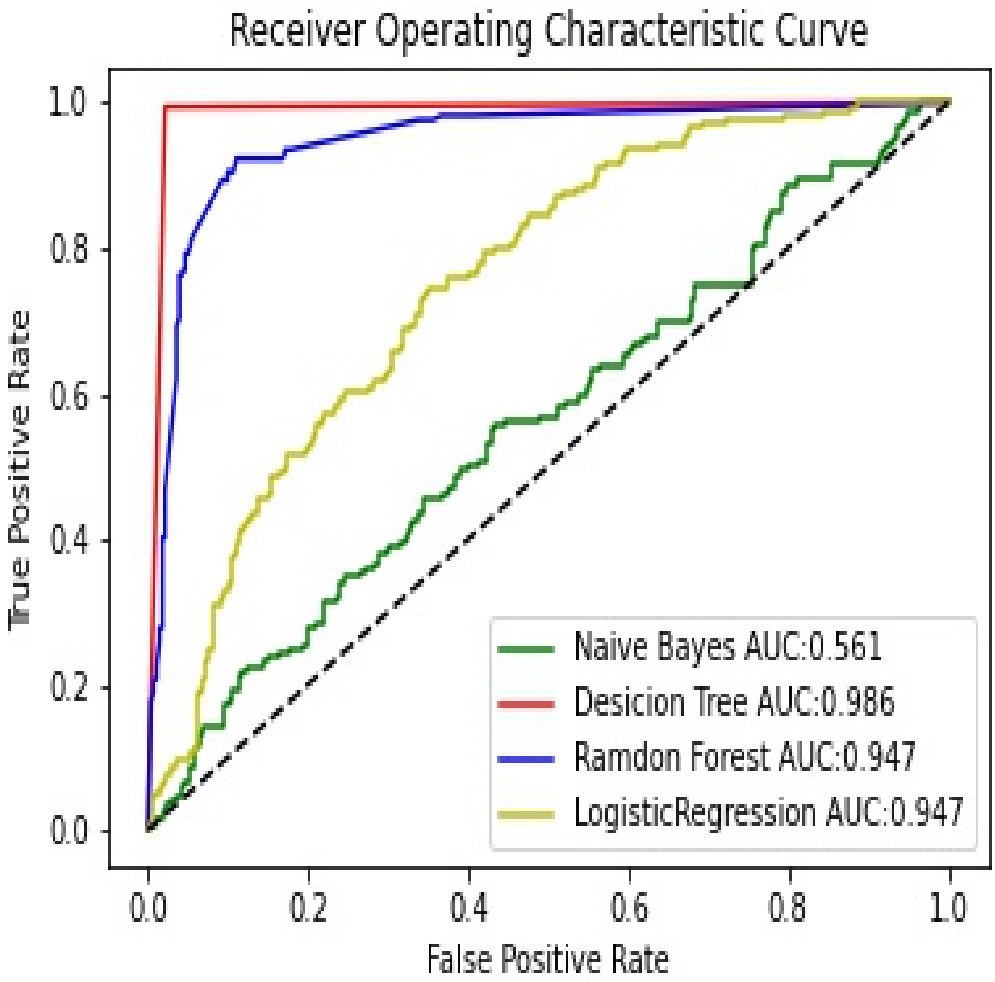

Figura 9: Curvas de ROC: Fuente propia

4.5.4. Análisis de las curvas de aprendizaje

Los modelos de machine learning tienden a presentar un sobre ajuste en el conjunto de datos de entrenamiento, esto sucede cuando el modelo se memoriza los datos de entrenamiento y no permite un buen desempeño de clasificación de los datos de prueba.

Las curvas de aprendizaje permiten analizar si el modelo presenta un sobre ajuste o no, ya que estas grafican el nivel de precisión en diferentes conjuntos de datos de entrenamiento contra el número de observaciones.

Cuando la curva de entrenamiento está muy cerca a 1 para los diferentes tamaños muestrales estaría indicando un overfiting en el modelo. Por otro lado; cuando las dos curvas están muy cerca o sobrepuesta indica que el modelo está presentando un buen ajuste tanto en los datos de entrenamiento y prueba, como se logra ver en la figura 10

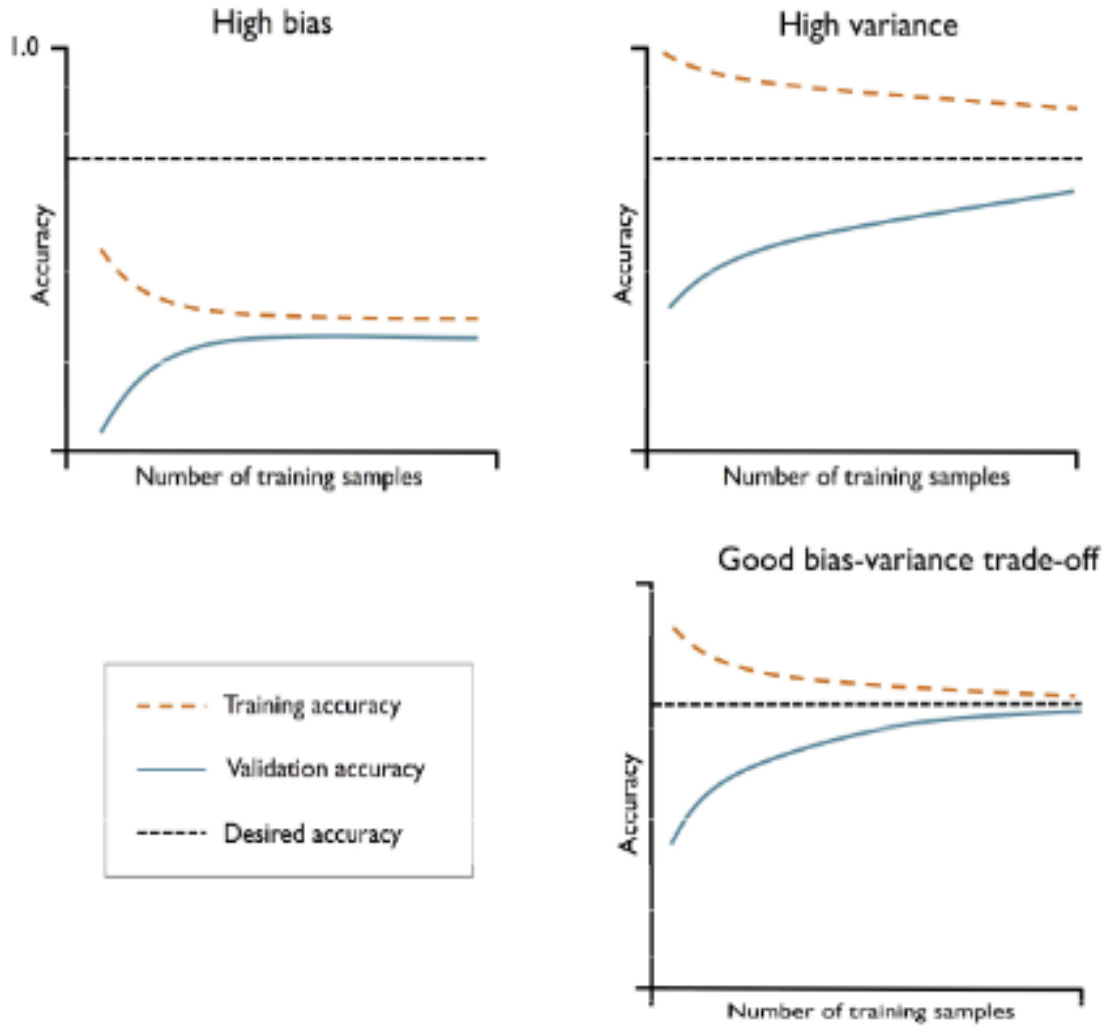

Figura 10: Python Machine Learning (Raschka Mirjalili, 2019)

En la figura 11 se observa que la mejor curva de aprendizaje es generada por el algoritmo Decision Tree, ya que presenta una convergencia en los datos de entrenamiento como en los datos de prueba, para la curva de Random Forest presenta un problema de varianza, esto puede verse a que es un algoritmo complejo, entonces puede presentar desajuste por que no tiene suficientes data, debido que es un clasificador muy fuerte en grandes cantidades de datos. Las curva de Naive Bayes tambien presenta una no convergencia de todos sus conjuntos de datos y finalmente Linear Regresion presenta una convergecia per una auccuracy mas bajo que Dession Tree, así valida los resultados anteriores y se concluye que el mejor modelo para predecir la deserción laborar es Desecion Tree con las parametrizaciones mencionadas anteriormente.

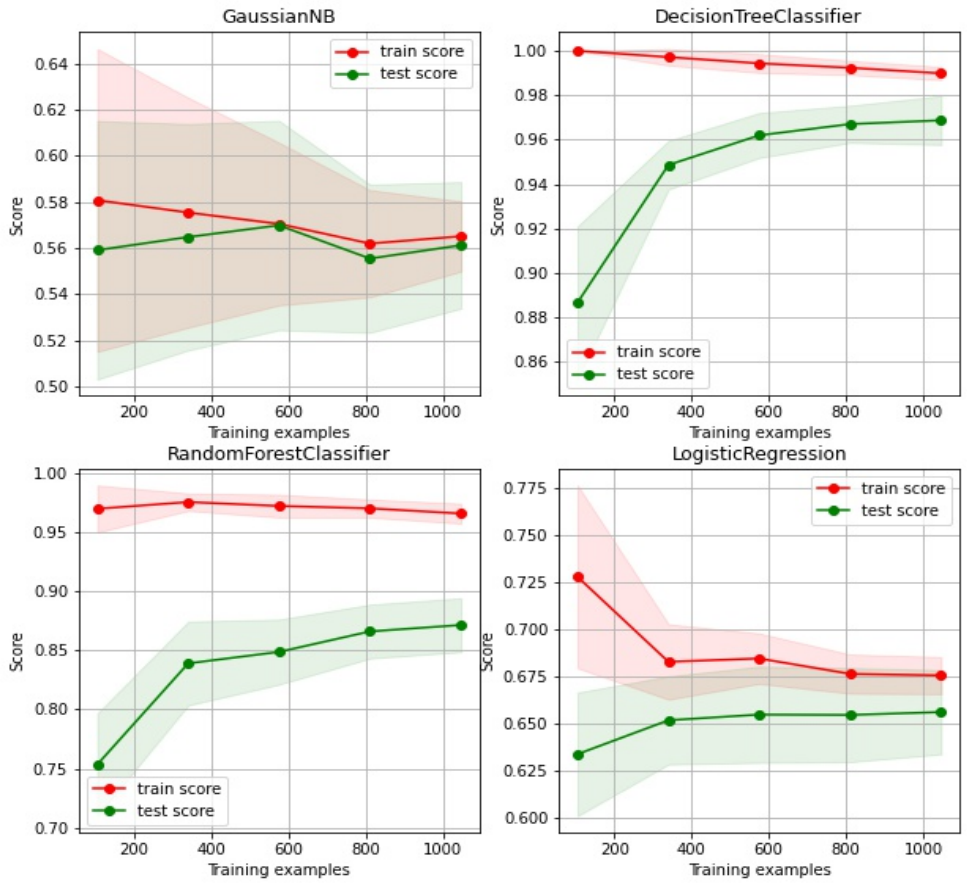

Figura 11: Matriz de correlación: Fuente propia

5. Recomendaciones y Conclusiones

5.1. Conclusiones

El modelo a utilizar es finalmente el que es consistente con los datos de entrenamiento y los datos de prueba. Para este problema particular, las diferentes tecnicas de aprendizaje estadistico buscan un unico objetivo predecir correctamente el derserción por tanto, dado los modelos propuestos se concluye que, el mejor modelo a utilizar para predecir la variable respuesta es arbol de desición con 14 capas, puesto que, ya se ha comprobado es el que mejor desempeño tiene al comparar con los modelos propuestos.

También se logró entender que las variables más importantes del modelo son salario total y tiempo de permanencia, ya que a mayor tiempo en la compañía menor probabilidad de abandonar su puesto de trabajo, sucediendo lo mismo con salario total, por otro lado, se logró identificar que los trabajadores más jovenes son los más que tienen probabilidades de abandonar su cargo laboral.

Particularmente para la empresa analizada, los esfuerzos por mantener unos niveles de deserción que no impacten negativamente las finanzas de la empresa no han dado sus frutos. Actualmente, las empresas de la industria de software ven afectados sus resultados financieros de manera mensual con costos de liquidaciones de empleados que decidieron dejar de apostarle a la empresa para el crecimiento profesional. Así mismo, estas empresas pierden la apuesta de invertir en diferentes capacitaciones del empleado desertor y se ven obligadas a iniciar de nuevo el proceso de adaptación y de empalme.

5.2. Recomendaciones

Finalmente, se recomienda para nuevas investigaciones, en primer lugar, aumentar el tamaño de la base datos y de ser posible realizar un muestreo aleatorio estratificado o simple en diferentes empresas y regiones, para así poder generalizar los resultados obtenidos en los modelos implementados. En segundo lugar, se recomienda entrenar el modelo usando otros tipos de algoritmos, por ejemplo, una máquina de soporte vectorial, redes neuronales, entre otros y también de ser posible incluir otro tipo de variables. Por otro lado, se recomienda realizar el mismo estudio en empresas con diferentes líneas de negocio y de diferentes regiones para así poner lograr una mayor generalización de los resultados, ya que los que se obtuvieron en el desarrollo de este proyecto están centrados a solo una empresa.

Referencias

Alamsyah, A., Nurriz, B. (2017). Monte Carlo simulation and clustering for customer segmentation in business organization. Proceeding - 2017 3rd International Conference on Science and Technology-Computer, ICST 2017. https://doi.org/10.1109/ICSTC.2017.8011861

Oscar Fuentes. (2022). Las profesiones y carreras mejor pagadas y más demandadas del 2022, Blog IeBS Digital school, https://www.iebschool.com/blog/profesiones-solicitadas-bien-pagadas-reclutamiento-seleccion

Andonova, V., S. Nikolova, M., Dimitrov , D. (2019). Entrepreneurial Ecosystems in Unexpected Places Examining the Success Factors of Regional Entrepreneurship (first ed.). Palgrave Macmillan.

Casella, A., Hanaki, N. (Mayo de 2006). american economic association. Recuperado el Febrero de 2021, de https://www.aeaweb.org/articles?id=10.1257/000282806777212341

Galán-Cortina, V. (2015). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario. In Universidad Carlos III de Madrid. Departamento de Informática. https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC Victor GalanCortina.pdf

Florence.T;, A. M., Ms.Savithri.R. (2013). International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) TALENT KNOWLEDGE ACQUISITION USING C4 . 5 CLASSIFICATION ALGORITHM. International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), 4.

Jantan, H., Razak, A., Ali, Z. (2011). Data Mining Classification Techniques for Human Talent Forecasting. In Knowledge-Oriented Applications in Data Mining. https://doi.org/10.5772/14007.

Jain, P. K., Jain, M., Pamula, R. (2020). Explaining and predicting employees' attrition: a machine learning approach. SN Applied Sciences, 2(4). https://doi.org/10.1007/s42452-020-2519-4.

Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., Zolfani, S. H. (2021). An improved machine learning-based employees attrition prediction framework with emphasis on feature selection. Mathematics, 9(11). https://doi.org/10.3390/math9111226.

Murphy, K. P. (2006). Naive Bayes classifiers Generative classifiers. Bernoulli, 4701(October). https://doi.org/10.1007/978-3-540-74958.

Tambde, A.& Motwani, D. (2019). Employee churn rate prediction and performance using machine learning. https://doi.org/10.35940/ijrte.B1134.0982S1119.

Computational Science and Its Applications - ICCSA 2006: International Conference, Proceedings. (2006). In Lecture Notes in Computer Science.

Calviño, A. (2020). Material de la asignatura Técnicas y Metodología de la Minería de Datos (SEMMA).

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.Coussement, K., den Poel, D. V. (2008). Churn prediction in subscription services, 34, 313–327.

Boehmke, B., Greenwell, B. (2020). Hands-On Machine Learning with R. Londres: Chapman and Hall/CRC - Taylor Francis Group/ https://bradleyboehmke.github.io/HOML/bagging.html.

K. P. Murphy,(2006) "Naive Bayes classifiers Generative classifiers," Bernoulli, vol. 4701, no. October, pp. 1–8, 2006.