"Año del Bicentenario, de la consolidación de nuestra Independencia, y de la conmemoración de las heroicas batallas de Junín y Ayacucho"

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

(Universidad del Perú, DECANA DE AMÉRICA)

FACULTAD DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SOFTWARE

"MODELOS DE MACHINE LEARNING CLASIFICADORES APLICADOS A LA DESERCIÓN DE EMPLEADOS"

Presenta: Tupac Agüero, Kevin

LIMA-PERÚ

2024

TABLA DE CONTENIDOS

MODELOS DE MACHINE LEARNING CLASIFICADORES APLICADOS A LA DESERCIÓN DE EMPLEADOS

RESUMEN

La deserción de empleados, entendida como la renuncia a una organización, representa un desafío significativo tanto para la eficiencia operativa como para la salud financiera de las empresas. Este fenómeno implica altos costos directos e indirectos, como el reclutamiento, la selección, la capacitación, la pérdida de productividad y la carga de trabajo adicional para los empleados restantes. Por tanto, comprender los factores que influyen en la tasa de renuncia es crucial para desarrollar estrategias efectivas de retención de empleados.

Este estudio tiene como objetivo predecir la variable 'Atrttion' la cual es categórica nominal, basado en factores clave que influyen en la decisión de un empleado de renunciar, los cuales se agrupan en factores relacionados con el empleado (satisfacción laboral, compromiso, bienestar, equilibrio entre vida laboral y personal), factores relacionados con el trabajo (carga laboral, oportunidades de desarrollo, ambiente laboral, reconocimiento y recompensa) y factores externos (condiciones económicas y competencia por talento).

Para a predicción de la renuncia de los empleados se utilizará algoritmos de aprendizaje automático, específicamente k-Nearest Neighbors (KNN) y Random Forests (RF). También será clave la selección de características relevantes y el preprocesamiento adecuado de los datos son fundamentales para el éxito de estos modelos. Se han utilizado técnicas como la normalización de datos, PCA y ANOVA.

ABSTRACT

Employee turnover, understood as the resignation from an organization, represents a significant challenge for both operational efficiency and financial health of companies. This phenomenon involves high direct and indirect costs, such as recruitment, selection, training, loss of productivity, and additional workload for remaining employees. Therefore, understanding the factors that influence the resignation rate is crucial for developing effective employee retention strategies.

This study aims to predict the 'Attrition' variable, which is a nominal categorical variable, based on key factors influencing an employee's decision to resign. These factors are grouped into employee-related factors (job satisfaction, commitment, well-being, work-life balance), job-related

factors (workload, development opportunities, work environment, recognition and reward), and external factors (economic conditions and talent competition).

To predict employee resignation, machine learning algorithms, specifically k-Nearest Neighbors (KNN) and Random Forests (RF), will be used. The selection of relevant features and proper data preprocessing are also crucial for the success of these models. Techniques such as data normalization, PCA, and ANOVA have been employed.

I. INTRODUCCIÓN

La renuncia es un acto mediante el cual una persona decide separarse voluntariamente de su labor habitual, generalmente asociado a una situación de incomodidad generada en el ambiente de trabajo [1]. Este fenómeno impacta de manera negativa tanto en la organización como en el trabajador, generando ausentismo, frustración, disminución de la motivación, falta de creatividad y bajo rendimiento [2]. La intención de renunciar es crucial en el ámbito de las organizaciones, ya que implica costos significativos para la empresa. Estos costos pueden ser directos, como los procesos de reclutamiento, selección, contratación y formación del empleado, e indirectos, como la pérdida de productividad y eficiencia del nuevo empleado durante su periodo de aprendizaje y el tiempo invertido en su supervisión [3].

La tasa de renuncia de los trabajadores es un indicador clave del desempeño de una organización. Los empleados que renuncian representan una pérdida de inversión en reclutamiento, capacitación y desarrollo, además de afectar negativamente la moral y la productividad del equipo restante [4]. La salida de un empleado implica una interrupción en el flujo de trabajo, lo que puede llevar a una pérdida de eficiencia y una carga de trabajo adicional para los empleados restantes, generando una caída generalizada de la moral [5]. Por lo tanto, comprender los factores que influyen en la tasa de renuncia es esencial para desarrollar estrategias efectivas de retención de empleados.

Diversos estudios han identificado múltiples factores que influyen en la decisión de un empleado de renunciar. Estos factores pueden agruparse en tres categorías principales:

1. Factores relacionados con el empleado:

- Satisfacción laboral: Los empleados que están satisfechos con su trabajo, su salario, sus beneficios y las oportunidades de crecimiento tienen menos probabilidades de renunciar [8].
- Compromiso: Los empleados comprometidos con su trabajo y con la organización son menos propensos a renunciar.
- Bienestar: El estrés, el burnout y el agotamiento pueden llevar a la renuncia de los empleados.
- Equilibrio entre la vida laboral y personal: Los empleados que no logran un equilibrio adecuado entre su vida laboral y personal son más propensos a renunciar.

2. Factores relacionados con el trabajo:

- Carga laboral: Los empleados con una carga de trabajo excesiva o con tareas poco desafiantes son más propensos a renunciar.
- Oportunidades de desarrollo: Los empleados que no ven oportunidades de desarrollo profesional en su puesto actual son más propensos a renunciar [9].
- Ambiente laboral: Un ambiente laboral negativo, tóxico o discriminatorio puede llevar a la renuncia de los empleados.
- Reconocimiento y recompensa: Los empleados que no se sienten reconocidos y recompensados por su trabajo son más propensos a renunciar.

3. Factores externos:

- Condiciones económicas: Las condiciones económicas favorables pueden aumentar la tasa de renuncia, ya que los empleados tienen más oportunidades de encontrar un nuevo trabajo.
- Competencia por talento: La competencia por empleados calificados puede llevar a las empresas a ofrecer mejores salarios y beneficios, lo que puede aumentar la tasa de renuncia en otras empresas. El talento requiere de una adecuada movilidad dentro de las organizaciones globales hacia posiciones donde se aproveche su potencial [10].

OBJETIVO

Este estudio tiene como objetivo predecir la probabilidad de renuncia de los empleados según los factores identificados. Comprender estos factores es esencial para que las organizaciones desarrollen estrategias efectivas de retención de empleados, minimizando así los costos asociados a la rotación de personal y mejorando la moral y la productividad del equipo.

JUSTIFICACIÓN

La justificación de este estudio radica en la necesidad crítica de las organizaciones de reducir la tasa de renuncia de empleados, dado que esta representa un desafío significativo para la eficiencia operativa y la salud financiera de las empresas. La alta rotación de personal no solo implica costos directos, como los asociados a los procesos de selección y contratación, sino también costos indirectos que afectan la productividad, la moral del equipo y, en última instancia, la capacidad competitiva de la organización en el mercado.

Al identificar y comprender los factores que influyen en la decisión de los empleados de renunciar, las organizaciones pueden diseñar e implementar estrategias más efectivas para mejorar la retención de personal. Esto incluye la mejora de las condiciones laborales, la creación de un ambiente de trabajo más positivo y el desarrollo de políticas de reconocimiento y recompensa que valoren adecuadamente el desempeño y la contribución de los empleados.

Además, este estudio proporciona un marco para el análisis y la intervención en áreas específicas que son críticas para la satisfacción y el compromiso de los empleados, como la gestión del estrés, el equilibrio entre la vida laboral y personal, y las oportunidades de desarrollo profesional. Al abordar estos factores, las organizaciones no solo pueden reducir la tasa de renuncia, sino también fomentar un ambiente de trabajo más saludable y productivo, lo que en última instancia contribuye al éxito a largo plazo de la empresa.

II. TRABAJOS RELACIONADOS

Se ha optado por dividir las investigaciones encontradas en dos categorías, las cuales son:

- Modelos de Aprendizaje Automático Implementados para el Reclutamiento Laboral: En esta primera sección se revisará la literatura existente relacionada con este concepto, con el fin de adquirir conocimiento de los modelos de machine learning aplicados al reclutamiento laboral.

- Técnicas de preprocesamiento y modelos de Aprendizaje Automático usados para predecir la deserción de empleados: Esta sección se dedicará a la revisión exhaustiva de la literatura existente en este ámbito, con el objetivo de conocer cuáles son los modelos de aprendizaje automático y las técnicas de preprocesamiento que se han aplicado a conjuntos de datos análogos.

Modelos de Aprendizaje Automático Implementados para el Reclutamiento Laboral

Para la toma de decisiones post aprendizaje automático, previamente se requiere analizar la información necesaria, por ello se demanda que el conjunto de datos pase por un preprocesamiento correcto. Al hacer un estudio acerca de las técnicas utilizadas en casos relacionados donde se aplicaron métodos de machine learning al reclutamiento laboral o job recruitment, se observó lo siguiente:

Las técnicas de aprendizaje automático más utilizadas para la selección de personal o reclutamiento del mismo son las siguientes:

KNN: es un algoritmo que a partir de una serie de grupo de datos es capaz de clasificar de una manera acertada en una instancia nueva. Está conformado por varios atributos descriptivos y de un solo atributo objetivo que también se le conoce como clase [11].

Algoritmo de Regresión Logística: esto viene a ser un grupo de técnicas estadísticas las cuales tienen como objetivo explicar las causas de los eventos de la variable dependiente nominal [12]. Estadísticamente, es un enfoque lineal para modelar respuestas escalares (variables dependientes) y variables explicativas (independientes) [13].

Se desarrollaron modelos similares con el fin de reclutamiento laboral, como un sistema de recomendación basado en filtros colaborativos el cual selecciona exclusivamente la información necesaria de cada usuario y lo compara con diversos perfiles buscando similitudes, recomendando los ítems que han resultado de interés para perfiles similares [14].

En [15] se realizó un estudio de utilizando técnicas de machine learning donde se inspeccionan las principales características para un correcto reclutamiento de personal. Los

resultados mostraron que las principales son el nivel y tipo de discapacidad, nivel de estudios, experiencia y capacitación. En el artículo “Apoyo a los subsistemas de talento humano, selección y reclutamiento a partir de un sistema experto” se tomaron 120 muestras de datos con 70 características donde se concluyó que las mejores técnicas de machine learning son la aplicación de SVN con una precisión del 82.1% seguido Regresión logística con parámetros de función cuadrática y cúbica, con precisión de 81.8% y 81.5% respectivamente; y donde el menos preciso es el modelo KNN de 3 vecinos con precisión de 79.1%. Estas características son consideradas después de un previo análisis de identificación de necesidades a partir de cada puesto de trabajo se debe iniciar el proceso de reclutamiento [16]. Esta tarea se debe realizar de la manera correcta, ya que una mala identificación de las características requeridas resultará en un incorrecto modelamiento del sistema; por lo que es necesario contar con una medida objetiva del éxito real de la contratación y el desempeño de los empleados, además de proporcionar información significativa a los propios reclutadores[17]. En [18], por otro lado, se aplican técnicas de inteligencia artificial como Razonamiento basado en casos textual o McCall (este último evaluando hasta 3 ejes principales con sus respectivos de calidad) para desarrollar un modelo de selección de personal apto para la vacante disponible diferenciándolos de los candidatos restantes. Y en [19] se empleó la técnica de Árbol de Decisiones con una alta precisión de 99.8% siendo comparada en el mismo artículo comparado con otras técnicas de selección de características de solicitantes de empleo.

La validación del algoritmo es importante para determinar si el modelo cumple con los requerimientos inicialmente establecidos para su creación, ya que es necesario evaluar su poder predictivo sobre nuevos datos que no han sido vistos [20]. Existen diversas metodologías para lograr dicho fin, como ejemplo en [21], donde se desarrolló un modelo tomando en cuenta el desempeño laboral por medio de las variables de estudio, se recurrió a la validación cruzada, en concreto los métodos de validación que se usaron fueron K-Fold Cross Validation y LOOCV para determinar que tan acertado la predicción del modelo.

Para concluir esta sección, se evidencia que las empresas logran tomar decisiones más acertadas en la contratación de personal al utilizar técnicas de aprendizaje automático y también porque la empresa esté interesada en tener una lista de posibles candidatos a los que se pueda recurrir en un futuro [22].

Técnicas de preprocesamiento y modelos de Aprendizaje Automático usados para predecir la deserción de empleados

Es ampliamente reconocido que el preprocesamiento de datos constituye un pilar fundamental en el manejo de conjuntos de datos en bruto. Antes de poder extraer información o conocimiento de valor, es imperativo que el conjunto de datos en cuestión sea sometido a un preprocesamiento adecuado. Sin embargo, este paso crucial es frecuentemente subestimado por una gran cantidad de investigadores, principalmente debido a la complejidad percibida y al tiempo requerido para su implementación.

Algunas de las técnicas utilizadas en el trabajo fueron:

PCA (análisis de componentes principales) es un método para reducir la cantidad de variables en un conjunto de datos. Lo hace combinando variables altamente correlacionadas. PCA nos permite representar datos a lo largo de un eje y ese eje se llama componente principal. Simplifica dimensiones superiores a dimensiones inferiores [23], también fue usado análisis de varianza (ANOVA), este es el método más exacto para calcular la variabilidad de un sistema de medición porque posee la ventaja de cuantificar la variación debida a la interacción entre los operadores y las partes [24]. Se utiliza dicha técnica cuando se tiene una variable de medición y dos o más variables nominales [25].

En lo que respecta a selección de características, según lo observado algunas resultan más importantes que otras para predecir el attrition, como la motivación, esta afecta la inversión en el trabajo y el compromiso laboral [26], y cómo con los empleados desfavorecidos en el benchmarking de salarios resultan con altas probabilidades de deserción [27], así que es clave identificar y quedarnos con las características relevantes. Para los datos y caso de estudio presentes en [28], las cinco características más importantes identificadas son: nivel de satisfacción, tiempo en la empresa, horas promedio mensuales, número de proyectos y última evaluación.

Luego, analizando únicamente al uso técnico de Machine Learning utilizadas para casos relacionados con la deserción laboral o employee turnover, se observó lo siguiente:

En el artículo [29] se discute acerca de cómo la retención de empleados es vital para el éxito organizacional, y la tasa de rotación se ha convertido en una métrica clave, y como la regresión logística y el algoritmo de bosque aleatorio son los modelos más efectivos para predecir la rotación de empleados, ya que estos pueden manejar relaciones no lineales entre las características, lo que es común en los datos de rotación de empleados. Otros artículos

mencionan un buen desempeño de los Árboles de Decisión, Redes Neuronales, KNN y SVM para mejorar la predicción de abandono de clientes en la industria de telecomunicaciones [30]. Luego fue analizado el caso de un callcenter, lugares con alta rotación laboral, se mostró que según los registros de desempeño operativo de 2407 agentes de ventas mostraron que, aunque el Naive Bayes es mucho más simple que Random Forest, ambos clasificadores se desempeñaron de manera similar, logrando tasas de precisión interesantes en la predicción de la rotación [31]. Notamos que al poder predecir la deserción laboral podemos evitar el costo de oportunidad , y las inversiones de reclutamiento, los resultados obtenidos fueron que el bosque aleatorio tuvo el mejor rendimiento con una precisión del 90.20%, mientras que el de Naïve Bayes tuvo la peor precisión con un 80.20%. [32].

En cambio, para el caso de estudio [33] se observa que el modelo basado en el algoritmo k-Nearest Neighbors (KNN) superó al modelo basado en Random Forest (RF) en varias métricas clave. Específicamente, el modelo KNN mostró una mayor precisión promedio (84% frente al 80% de RF) y una mejor Área Bajo la Curva (AUC) (0.79 frente a 0.82). Además, el KNN obtuvo un F-score promedio de 0.196, mientras que el RF alcanzó un F-score promedio de 0.281.

Por último, concluir que el uso de técnicas para el employee turnover se extienden más allá de sus capacidades puras de predicción para modificar las operaciones de la organización, también al ahorro de costos del análisis de rotación de empleados, los profesionales de recursos humanos pueden utilizar las explicaciones del modelo para desarrollar políticas de retención en toda la empresa y también dirigirse a personas de alto riesgo con iniciativas de retención. [34].

III. METODOLOGÍA

Para abordar el problema de Attrition, vamos a realizar la aplicación del modelo de machine learning Support Vector Machine. La deserción de empleados, definida como la renuncia voluntaria a una organización, es un desafío crítico para las empresas debido a los altos costos asociados, tanto directos como indirectos, tales como el reclutamiento, la selección, la capacitación y la pérdida de productividad. Por ello, comprender y predecir los factores que influyen en la decisión de los empleados de renunciar es esencial para desarrollar estrategias efectivas de retención. Nuestro enfoque se centra en predecir la variable 'Attrition', una variable categórica nominal, utilizando técnicas avanzadas de machine learning. Los factores predictivos incluyen variables relacionadas con el empleado (satisfacción laboral, compromiso, bienestar, equilibrio entre vida laboral y personal), factores relacionados con el

trabajo (carga laboral, oportunidades de desarrollo, ambiente laboral, reconocimiento y recompensa) y factores externos (condiciones económicas y competencia por talento). La correcta selección de características y el preprocesamiento adecuado de los datos, a través de técnicas como la normalización, PCA y ANOVA, son cruciales para el éxito de nuestro modelo, permitiendo a las organizaciones anticipar y mitigar la deserción de empleados de manera efectiva.

Procedencia del conjunto de datos

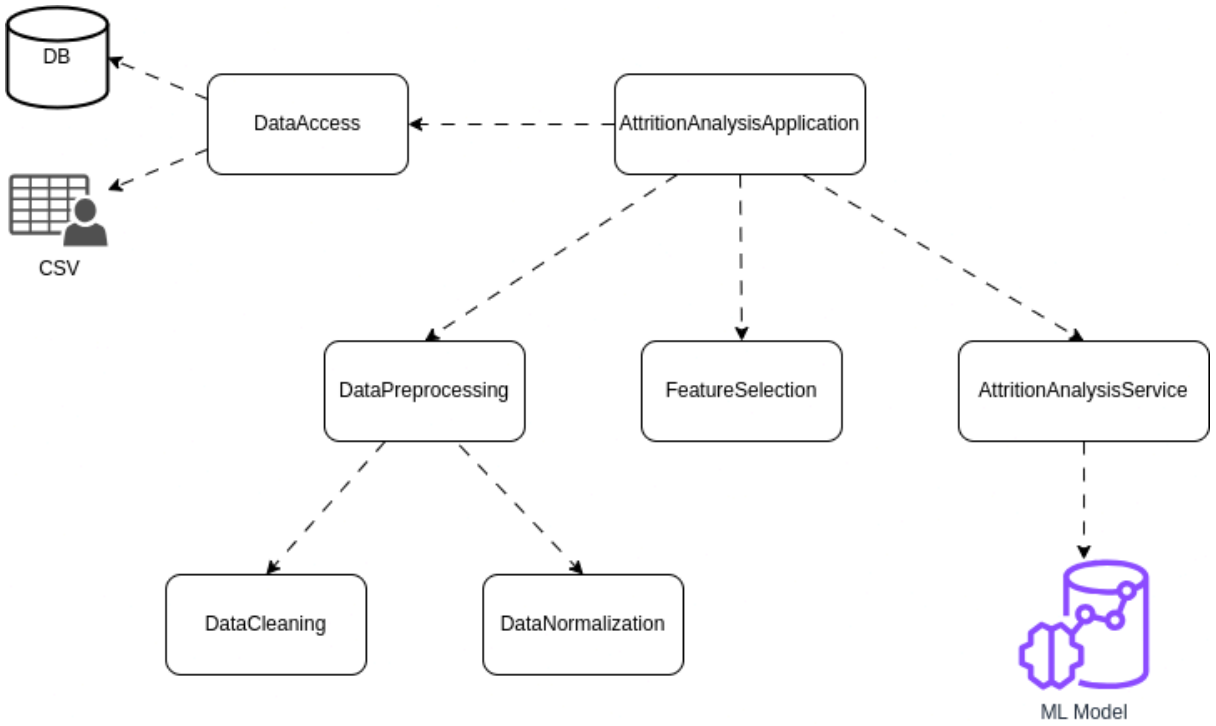
Este es un conjunto de datos ficticio, proporcionado por IBM. Se denomina "IBM HR Analytics Employee Attrition & Performance" contiene casi 30 características de datos categóricos y discretos. Estos datos son valores numéricos y de texto que ayudan a analizar los datos de los empleados desde la contratación hasta el despido o abandono del empleado. Fue extraído del repositorio de datasets presentes en Kaggle.[35].

Arquitectura

La arquitectura del sistema para predecir la deserción de empleados se ha diseñado como un modelo basado en componentes, cada uno con una función específica que contribuye al proceso general de análisis y predicción. A continuación, se describe la arquitectura del sistema, su funcionamiento general y el detalle de cada componente.

1. **Acceso a Datos:**
   Este componente se encarga de recolectar y actualizar datos de diversas fuentes, como bases de datos internas de la empresa y archivos CSV.
2. **Preprocesamiento de Datos:**
   Este componente se encarga de la limpieza y preparación de los datos para asegurar su calidad.
   a. **Limpieza de Datos:** Eliminación de valores nulos y tratamiento de valores atípicos para asegurar la calidad de los datos.
   b. **Normalización de Datos:** Los datos se normalizan para asegurarse de que todas las características contribuyen equitativamente al modelo.
3. **Selección de Features:**
   Este componente Los datos se normalizan para asegurarse de que todas las características contribuyen equitativamente al modelo.
4. **Servicio de Análisis:**
   Este componente será por el que pasará la data antes de entrar al modelo ML.
5. **Model ML:**

Incluye varios algoritmos de machine learning para el análisis predictivo. Se entrenará con la data previamente pre-procesada.



Preprocesamiento

1. **Observación de tipos de datos:** Analizamos los datos para identificar los tipos de datos presentes en el dataset de recursos humanos. Encontramos datos de tipo:

Tipo de datos	Columnas
int64	Age DailyRate EmployeeCount MonthlyIncome MonthlyRate NumCompaniesWorked PercentSalaryHike PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion

	YearsWithCurrManager
float64	DistanceFromHome Education EmployeeNumber EnvironmentSatisfaction HourlyRate JobInvolvement JobLevel JobSatisfaction
object	BusinessTravel EducationField MaritalStatus Over18 OverTime Attrition Department Gender JobRole

Esto nos indica que tenemos tanto datos numéricos como categóricos, lo cual es importante para decidir como pre-procesarlos adecuadamente.

2. **Detección de datos faltantes:** Visualizamos si en nuestras columnas existen datos faltantes, es decir, datos que no fueron llenados. Identificamos que hay 10 columnas con datos faltantes:

	Columnas
Datos faltantes	Department DistanceFromHome Education EmployeeNumber EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel JobSatisfaction

Lo que nos indica la necesidad de tratar estos valores para evitar problemas en el análisis posterior.

3. **Eliminación de columnas con un solo tipo de valor:** Eliminamos las columnas que contienen solo un tipo de valor, ya que estas no aportan información útil al análisis.

Columnas con desviación estándar igual a 0	Columnas
	EmployeeCount Over18 StandardHours

La desviación estándar de estas columnas es 0, lo que significa que todos sus valores son iguales. Por lo tanto, no tienen variabilidad ni contribuyen a diferenciar las observaciones.

4. **Estandarización de datos categóricos:** Convierto todos los datos categóricos(object) a minúsculas. Esto se debe a que encontramos valores en mayúsculas y minúsculas dentro de las mismas columnas, lo cual podría dificultar la interpretación y el análisis de estos datos. Al estandarizarlos a minúsculas, hacemos que los datos sean más consistentes y manejables.
5. **Corrección de valores categóricos:** Corrijo los datos que estaban mal escritos o incompletos en las columnas categóricas. Esto implicó interpretar y reemplazar los valores incorrectos por los verdaderos valores esperados, asegurando así la integridad y precisión de los datos.

Columna	Antes	Después
Department	'sales', 'research & development', nan, 'research & ', 'sa', 'human resources', 'human r'	'sales', 'research & development', nan, 'human resources'
Gender	nan, 'male', 'ma', 'female', 'fem'	nan, 'male', 'female'

6. **Imputación de datos faltantes categóricos:**
    a. Moda: Aplico la moda para rellenar los datos faltantes en las columnas categóricas. Decidimos usar la moda porque los datos faltantes representaban solo el 5.5% del total. Dado el tamaño del dataset, completar estos valores con la moda no tendría un impacto negativo significativo en nuestro análisis.

Columna	Cantidad de datos faltantes	Representan del total
Department	74	5.03%
Gender	87	5.91%

7. **Imputación de datos faltantes cuantitativos:**
    a. Regresión Lineal: Utilizamos la correlación de Pearson para identificar columnas relacionadas. Observamos que "MonthlyIncome" y "JobLevel" tenían una correlación del 95%. Usamos regresión lineal para rellenar los datos faltantes en estas columnas debido a su alta correlación.
    b. Promedio de adyacentes: Como aún no habíamos eliminado los outliers, optamos por rellenar los datos faltantes numéricos con el promedio de los valores adyacentes. Esto nos ayudó a evitar la influencia de los outliers en los cálculos de los valores faltantes.

	Columnas
Promediamos las columnas que tengan datos faltantes por filas adyacentes	DistanceFromHome Education EmployeeNumber EnvironmentSatisfaction HourlyRate JobInvolvement JobSatisfaction

8. **Redondeo de datos:** Redondeo las anteriores columnas numéricas a enteros después de haberlos completado, ya que originalmente eran enteros. Esto nos permitió mantener la consistencia en el tipo de datos.
9. **Conversión de tipos de datos:** Convierto los datos de tipo float a int después del redondeo, ya que Python los había guardado como float. Esta conversión fue necesaria para mantener el formato original de los datos.

	Columnas
Cambiamos de float64 a int64	DistanceFromHome Education EmployeeNumber EnvironmentSatisfaction HourlyRate JobInvolvement JobSatisfaction

10. **Eliminación de outliers:** Observamos una gran cantidad de valores outliers en el dataset. Para ello eliminamos los valores outliers y nos quedamos con los rangos de mínimo y máximo de esas columnas.

	Columnas
Columnas con outliers que las filas donde estaban dichos outliers fueron eliminadas de nuestra data.	YearsAtCompany YearsSinceLastPromotion TotalWorkingYears YearsInCurrentRole

Al aplicar estos criterios, logramos reducir la cantidad de outliers manteniendo suficientes datos para el preprocesamiento y el posterior entrenamiento del modelo. No eliminamos datos en la columna "PerformanceRating" porque solo tenía dos tipos de datos, y eliminarlos habría significado perder casi la mitad del dataset.

Transformación

1. **Revisión de la distribución de datos:** Graficamos cada columna para analizar su distribución, observando distribuciones normales, homogéneas y no normales.
2. **Estandarización con StandardScaler:** Utilizamos StandardScaler para estandarizar los datos debido a que algunas distribuciones no seguían una distribución normal, y la varianza y la desviación estándar estaban muy alejadas de la media.

	Columnas
Estandarizamos los datos mediante StandardScaler	Age DailyRate MonthlyIncome MonthlyRate NumCompaniesWorked PercentSalaryHike PerformanceRating RelationshipSatisfaction StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager DistanceFromHome Education EmployeeNumber EnvironmentSatisfaction HourlyRate JobInvolvement JobLevel JobSatisfaction

3. **Estandarización de datos categóricos:** Utilizamos el método OrdinalEncoder para estandarizar nuestros datos categóricos a datos numéricos.

	Columnas
Columnas que fueron transformadas mediante el método OrdinalEncoder	BusinessTravel EducationField MaritalStatus OverTime Department Attrition Department Gender JobRole

Reducción

Para datos Cualitativos

El criterio de aplicación de los siguientes métodos fue influenciado por la efectividad que presentan.

1. **ANOVA (Análisis de Varianza):** El ANOVA es una técnica estadística que se utiliza para analizar las diferencias entre las medias de tres o más grupos. En el contexto del análisis de datos, ANOVA nos permite determinar si existen diferencias significativas en las medias de las variables categóricas. En este caso, ANOVA nos mostró que el score máximo es de 39.89% y el menor es de 0.0071%, sugiriendo que algunas columnas tienen poca influencia en el modelo. Este método descompone la variabilidad total observada en los datos en variabilidad "entre grupos" y "dentro de grupos", permitiéndonos identificar qué variables tienen un mayor impacto en la variable dependiente.
2. **Eliminación por Umbral del 5%:** Decidimos utilizar un umbral del 5% para determinar qué columnas eliminar. Las columnas con un score menor al 5% fueron consideradas para eliminación. En este caso, las columnas a eliminar serían aquellas que tienen un score menor a 5%, ya que su influencia en el modelo es mínima.

   Este criterio nos permite simplificar el modelo eliminando las variables numéricas que no contribuyen significativamente a la predicción de la variable independiente Y.

Columna	Score

RelationshipSatisfaction	4.486703
WorkLifeBalance	3.659993
TrainingTimesLastYear	2.451961
DailyRate	2.079636
NumCompaniesWorked	1.598666
Education	1.354748
PerformanceRating	1.093689
YearsSinceLastPromotion	0.988692
MonthlyRate	0.954776
EmployeeNumber,	0.081800
PercentSalaryHike	0.013303
HourlyRate	0.007095

Para datos Categóricos

Chi-Cuadrado: El test de Chi-Cuadrado se utiliza para evaluar la independencia entre dos variables categóricas. Este método compara las frecuencias observadas con las frecuencias esperadas bajo la hipótesis nula de que no existe relación entre las variables. Un valor de Chi-Cuadrado alto indica una mayor probabilidad de que exista una relación significativa entre las variables. Al aplicar el test de Chi-Cuadrado a nuestro dataset, obtuvimos scores que van desde 50.23% hasta valores menores como 0.057%. Esto nos indica el grado de relevancia de cada columna categórica con respecto a la variable dependiente Y. Notamos que solo la columna "OverTime" tiene suficiente relevancia para ser retenida en nuestro dataset, ya que muestra un score significativo.

Eliminación por Umbral del 5%: Utilizamos un umbral del 5% para determinar qué columnas eliminar. Las columnas con un score menor al 5% fueron consideradas para

eliminación. En este caso, las columnas a eliminar serían aquellas que tienen un score menor a 5%, ya que su influencia en el modelo es mínima. Este criterio nos permite simplificar el modelo eliminando las variables categóricas que no contribuyen significativamente a la predicción de la variable independiente Y. Esto no solo reduce la dimensionalidad del dataset, sino que también mejora la eficiencia del modelo.

Columna	Score
JobRole	2.743701
Department	0.587494
Gender	0.448289
EducationField	0.388971
BusinessTravel	0.057254

Al finalizar todos estos procesos de limpieza y transformación, hemos logrado reducir y optimizar nuestro dataset de manera significativa:

Estado Inicial del Dataset:

- Filas: 1470
- Columnas: 35

Estado Final del Dataset:

- Filas: 1041
- Columnas: 15

Resultados del Preprocesamiento

- **Reducción de Dimensionalidad:** Aplicamos la técnica de reducción de dimensionalidad ANOVA, eliminando columnas que no aportaban significativamente al modelo. Esto redujo el número de columnas de 35 a 15.
- **Limpieza de Datos Faltantes:** Tratamos y rellenamos los datos faltantes utilizando métodos apropiados como la moda, regresión lineal y el promedio de adyacentes, asegurando la consistencia y completitud del dataset.

- **Eliminación de Outliers:** Por filtrado de outliers eliminamos los valores que se consideran atípicos o extremos en una distribución de datos. Este proceso ayuda a mejorar la calidad del análisis y la precisión de los modelos predictivos. Utilizamos el método del rango intercuartílico (IQR) para identificar y eliminar estos valores atípicos.
- **Transformación y Estandarización:** Estandarizamos los datos numéricos y categóricos para asegurar uniformidad y facilitar el análisis y entrenamiento del modelo.
- **Análisis y Selección de Variables Categóricas:** Utilizamos el test de Chi-Cuadrado y establecimos un umbral del 5% para eliminar variables categóricas irrelevantes, dejando solo aquellas con mayor influencia en el modelo.

Preparación para el Entrenamiento del Modelo

El resultado de este exhaustivo preprocesamiento es un dataset más limpio, reducido y relevante, almacenado en un archivo CSV, listo para su uso en futuros entrenamientos de modelos. Esta optimización no solo facilitará un procesamiento más rápido y eficiente, sino que también mejorará la precisión y el rendimiento del modelo, ya que hemos eliminado datos redundantes y ruidosos que podrían haber afectado negativamente el resultado.

Al contar con un dataset bien preprocesado, podemos esperar que los modelos entrenados con estos datos tengan un mejor desempeño, lo que se traducirá en un valor predictivo más alto y resultados más confiables en aplicaciones de machine learning.

Estructura final del conjunto de datos

Después de aplicar el proceso de preprocesamiento descrito, el conjunto de datos original, que constaba de 1470 filas y 35 columnas, se transformó en un conjunto de datos preprocesado con las siguientes características:

1. **Número de Filas y Columnas:**
    a. 1041 filas y 15 columnas.
2. **Tipos de Datos:**
    a. El conjunto de datos preprocesado contiene una combinación de datos numéricos y categóricos. Se observó que existen variables tanto cuantitativas como cualitativas, lo que sugiere la presencia de datos numéricos y categóricos en las columnas.
3. **Ausencia de Datos Faltantes:**

a. Se realizaron varias estrategias para abordar los datos faltantes en el conjunto de datos:
    i. Para datos categóricos, se utilizó el método de imputación por moda.
    ii. Para datos numéricos, se aplicó la regresión lineal y el promedio de los valores adyacentes para completar los datos faltantes.
    iii. Además, se eliminaron outliers basados en umbrales específicos para ciertas columnas.

4. **Transformaciones Aplicadas:**
    a. Se aplicaron diversas transformaciones para estandarizar y limpiar el conjunto de datos:
        i. Se convirtieron todos los datos categóricos a minúsculas para estandarizar el formato.
        ii. Se realizaron transformaciones numéricas, como redondeo y conversión de tipo de datos de float a int después de completar los datos faltantes.
        iii. Se eliminaron columnas con solo un tipo de valor, lo que se consideró innecesario debido a la falta de variabilidad.
        iv. Se utilizó la técnica de ANOVA para reducir la dimensionalidad del conjunto de datos, manteniendo solo las columnas más relevantes.
5. **Integridad y Consistencia:**

Se verificó la integridad y la coherencia de los datos después del preprocesamiento para garantizar que fueran adecuados para el análisis y modelado subsiguientes.

Columna	Descripción
MaritalStatus	Al tener dependientes o una pareja, estas personas tienden a preferir trabajos estables que ofrezcan beneficios a largo plazo. Un entorno de trabajo incierto o la falta de seguridad laboral puede llevarlos a buscar oportunidades más estables.
Age	Al tener una edad mayor los empleados son más estables, más jóvenes tienden a experimentar más en otros puestos o

	compañías.
MonthlyIncome	Los empleados que ganan bien sienten que su trabajo es valorado y, por tanto, tienen menos incentivos para buscar otras oportunidades. En contraste, un sueldo bajo puede llevar a insatisfacción y búsqueda de mejores opciones salariales.
OverTime	Las horas extras pueden afectar negativamente el equilibrio entre la vida laboral y personal, llevando a la insatisfacción y deserción.
StockOptionLevel	El acceso a opciones sobre acciones puede influir en la retención de empleados, ya que estos beneficios pueden ser una forma significativa de compensación a largo plazo.
TotalWorkingYears	La cantidad total de años trabajados puede influir en la deserción, ya que los empleados más experimentados pueden sentir que han alcanzado su máximo potencial en una empresa.
YearsAtCompany	Los años que un empleado ha pasado en la empresa pueden influir en su decisión de quedarse o irse.
YearsInCurrentRole	Estar en el mismo rol durante muchos años sin progreso puede llevar a la desmotivación y posible deserción.
YearsWithCurrManager	La relación con la gerencia puede afectar la satisfacción laboral y la decisión de permanecer en la empresa.
DistanceFromHome	La distancia entre el hogar y el trabajo

	puede influir en la decisión de un empleado de quedarse en una empresa.
EnvironmentSatisfaction	El entorno de trabajo, incluyendo la cultura de la empresa y las condiciones físicas del lugar de trabajo, puede influir en la decisión de deserción.
JobInvolvement	El nivel de responsabilidad y participación en la empresa puede afectar la decisión de un empleado de quedarse.
JobLevel	La experiencia y las habilidades adquiridas dentro de la empresa pueden influir en la decisión de un empleado de quedarse o irse.
JobSatisfaction	La satisfacción general con el trabajo influye directamente en la decisión de deserción.
Attrition	La columna dependiente, si el empleado renuncia o no, de la empresa en donde trabaja.

Algoritmos utilizados por otros investigadores

Existen diferentes investigaciones y estudios ya realizados que también buscaban el mismo objetivo que el presente artículo, es por ello que se realizó una búsqueda exhaustiva y profunda de estudios y documentación similar a esta para obtener la literatura científica necesaria en la que basar el trabajo. Autores distintos utilizaron diferentes técnicas de machine learning para poder definir si un empleado podría renunciar a su empleo, dentro de ese conjunto de técnicas aplicadas se nombran las siguientes.

Hay autores que emplearon la técnica de Bosques Aleatorios como por ejemplo: en [32] el objetivo era estimar la rotación de empleados de recursos humanos bajo un conjunto de datos de empresas de Kaggle, el cual incluye 10 atributos diferentes de 1470 personas y el algoritmo tuvo un accuracy del 90.20%. También es utilizado en [34] donde se trabaja con datos de recursos humanos simulados de kaggle para evaluar un clasificador que logre

predecir qué tipo de empleados tendrán más probabilidades de irse dados algunos atributos. Donde Random Forest tiene una tasa de precisión más alta con 99,27%.

Otra técnica aplicada para la predicción de deserción laboral es Naive Bayes, como se utilizó en [31] donde se aplicaron ambas técnicas mencionadas en el título del artículo tomando datos operativos de los 2407 agentes de ventas del call center, se vio que a pesar de la simplicidad del Naive Bayes, el modelo funciona bastante bien en comparación con Random Forest, ya que este último se aprovecha solo cuando es necesario predecir la rotación con solo un mes de datos, pero con dos meses de información operativa (para predecir una rotación de un tercer mes), el rendimiento de ambos modelos es similar. Incluso el NB tiene un rendimiento ligeramente superior en términos de precisión.

Como tercera técnica utilizada en estudios de esta índole está SVM (Support Vector Machine), en [29] se buscó adaptar el análisis de los modelos sistemáticos de aprendizaje automático con el fin de seleccionar un modelo adecuado que mida el riesgo de deserción. En este se aplicaron diversas técnicas para medir el riesgo de deserción de personal, donde la técnica de SVM obtuvo un accuracy del 86,59% superando a Random Forest y Naive Bayes, ambas técnicas obteniendo un 83.24%.

Mejor modelo visto en la literatura

El modelo escogido es el Support Vector Machine (SVM). Esta elección se fundamenta en una revisión exhaustiva de la literatura científica existente sobre la predicción de la deserción laboral. Las investigaciones mencionadas han aplicado diversas técnicas de machine learning con el mismo objetivo, proporcionándonos una base sólida para la selección del modelo adecuado. Entre las técnicas evaluadas, el SVM ha demostrado un rendimiento superior en términos de precisión y accuracy.

Support Vector Machine, fundamento matemático

Vapnik y colegas introdujeron en la década de los 90’s las Máquinas de Soporte Vectorial. [36]. La formulación básica de las Support Vector Machines (SVM) se centra en la clasificación binaria; sin embargo, actualmente se puede usar para clasificación multiclase usando múltiples clasificadores binarios.

El objetivo de la clasificación mediante máquinas de vectores de soporte (SVM) es buscar de manera eficiente un hiperplano separador en un espacio de características de alta dimensión, el mejor hiperplano separador es aquel que está más separado de los puntos de las clases. Los

puntos más cercanos al hiperplano son denominados vectores de soporte. En otras palabras, el hiperplano óptimo se define de manera que maximiza el margen, y los vectores de soporte son los puntos de datos que se encuentran en los bordes de este margen.

1. **Hiperplano**

   Un hiperplano en un espacio n-dimensional se define como:

   $$w.x + b = 0$$

   dónde:

   w es un vector de pesos (también llamado vector normal del hiperplano) en R^n

   x es el vector de características de una muestra en R^n.

   b es el sesgo o intercepto, un escalar que desplaza el hiperplano desde el origen.

2. **Margen**

   El margen p es la distancia entre el hiperplano y los puntos de datos más cercanos de cualquier clase, que se conocen como vectores de soporte.

   Para un punto x, la distancia perpendicular del punto al hiperplano es:

   $$\frac{\left|w.x_i+b\right|}{||w||}$$

   Dónde $||w||$ Es la norma euclidiana del vector.

3. **Máxima Separación**

   El objetivo de una SVM es encontrar el hiperplano que maximiza el margen. Esto se puede formular como un problema de optimización:

   $$min\frac{1}{2}||w||^2$$

   sujeto a las restricciones:

   $$yi(w\cdot xi + b)\geq 1\forall i$$

   donde:

   $\|w\|^2$ es el cuadrado de la norma euclidiana del vector de pesos, que se usa como regularizador para evitar sobre ajuste.

   yi son las etiquetas de clase (+1 o -1) de los puntos de datos xi.

4. **Multiplicadores de Lagrange**

   Para resolver este problema de optimización, se utiliza la técnica de los multiplicadores de Lagrange. Se introduce una función Lagrangiana:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha i[yi(w \cdot xi + b) - 1]$$

donde:

αi≥0 son los multiplicadores de Lagrange correspondientes a cada restricción de la SVM.

N es el número total de muestras en el conjunto de datos.

5. Solución de la Forma Dual

La solución se encuentra en la forma dual, lo que nos lleva a:

$$max \sum_{i=1}^{N} \alpha i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha i\, \alpha j\, yi\, yj\, (xi \cdot xj)$$

sujeto a:

$$\sum_{i=1}^{N} \alpha i yi = 0 \quad y \quad \alpha i \geq 0$$

6. Determinación del Hiperplano

Una vez encontrados los valores óptimos:

El vector de pesos w se calcula como:

$$w = \sum_{i=1}^{N} \alpha i yi xi$$

El sesgo b se calcula usando los vectores de soporte:

$b = yi - w \cdot xi$ para cualquier vector de soporte xi

7. Función de Decisión

La función de decisión para clasificar un nuevo punto x es:

$$f(x) = sign\,(w.x + b)$$

donde:

$$sign(z) = 1,\ si\ z > 0$$
$$sign(z) = 0,\ si\ z = 0$$
$$sign(z) = -1,\ si\ z < 0$$

Máquina de Soporte Vectorial (SVM) Hiper Parámetros usados:

En este modelo de SVM, se ha seleccionado un valor de C de 1000. Este parámetro controla la regularización del modelo, donde un valor alto implica una menor regularización. Si bien esto permite que el modelo intente clasificar correctamente todos los ejemplos de entrenamiento, también aumenta el riesgo de sobreajuste.

Por otro lado, se ha elegido un valor de gamma de 0.001 para este parámetro, que determina la influencia de cada muestra de entrenamiento. Un valor bajo de gamma significa que la influencia de cada punto es amplia, lo que simplifica el modelo y lo hace más robusto a los outliers.

Finalmente, se ha seleccionado el kernel 'rbf' (Radial Basis Function). Este kernel es útil cuando la relación entre las características y las clases no es lineal, lo que permite al modelo capturar relaciones más complejas en los datos.

Bosques Aleatorios:

Fundamento Matemáticos

Función de Impureza de Gini

Para una partición T en un nodo del árbol, la impureza de Gini se define como:

$$Gini(T) = 1 - \sum_{i=1}^{C} p_i^2$$

Donde p_i es la proporción de elementos en la clase i y C es el número de clases.

Ganancia de Información

La ganancia de información se basa en la entropía, que mide la cantidad de incertidumbre en los datos. La entropía se define como:

$$Entropy(T) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

La ganancia de información para una partición basada en una característica A se calcula como la diferencia entre la entropía antes y después de la partición.

Reducción de Varianza

La combinación de múltiples árboles de decisión ayuda a reducir la varianza del modelo sin aumentar significativamente el sesgo. Este principio se basa en la Ley de los Grandes Números y se puede formalizar como:

$$Var(\widehat{f}RF) = \frac{1}{M} Var(\widehat{f_i})$$

Bosques Aleatorios Hiper Parámetros usados:

En el caso del bosque aleatorio, se ha seleccionado un valor de 300 para el parámetro n_estimators. Este parámetro define el número de árboles en el bosque. Un mayor número de árboles generalmente mejora el rendimiento del modelo al reducir la varianza, a costa de un mayor tiempo de cálculo.

No se ha establecido un límite para la profundidad máxima de los árboles (max_depth). Esto permite que los árboles crezcan hasta que todas las hojas sean puras o contengan menos muestras que min_samples_split, que en este caso se ha establecido en 2.

El valor de min_samples_leaf se ha fijado en 1. Este parámetro ayuda a prevenir el sobreajuste al garantizar que las hojas tengan un tamaño mínimo, evitando así que el modelo se base en detalles irrelevantes de los datos de entrenamiento.

Fundamento Matemáticos

El teorema de Bayes proporciona una forma de calcular la probabilidad de una hipótesis dada alguna evidencia:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Donde:

- $P(H|E)$ es la probabilidad de la hipótesis H dada la evidencia E (probabilidad a posteriori).
- $P(E|H)$ es la probabilidad de la evidencia E dada la hipótesis H (probabilidad verosímil).
- $P(H)$ es la probabilidad de la hipótesis H antes de observar la evidencia (probabilidad a priori).
- $P(E)$ es la probabilidad de la evidencia E.

Para clasificar calculamos la probabilidad a posteriori para cada clase y seleccionamos la clase con la mayor probabilidad:

$$\widehat{C} = arg\,max_C P(C) \prod_{i=1}^{N} P(X_I|C)$$

Naive Bayes:

En el modelo Naive Bayes, se ha seleccionado un valor de 1e-09 para el parámetro var_smoothing. Este parámetro añade una pequeña cantidad a la varianza calculada de cada característica. Esto estabiliza el modelo y evita divisiones por cero en los cálculos de probabilidad, lo que puede ocurrir cuando la varianza de una característica es muy baja o nula.

IV. RESULTADOS

Conjunto de datos

1. Conjunto de Datos

El conjunto de datos utilizado consta de N observaciones. Cada observación incluye X características relevantes para el análisis.

- Número total de observaciones: 1041
- Número de características por observación: 14

2. División Train/Test

En esta partición, el 80% de los datos se utiliza para el entrenamiento del modelo y el 20% restante para la prueba del mismo.

- Datos de entrenamiento (80%): 832
- Datos de prueba (20%): 209

Esta división se ha realizado de manera aleatoria para asegurar la representatividad de los datos en ambos subconjuntos.

3. Proceso de División

El proceso de división se ha llevado a cabo utilizando la función train_test_split de la biblioteca Scikit-Learn en Python, asegurando así la reproducibilidad y aleatoriedad en la partición de los datos.

Sometiendo el dataset al modelo

Luego de dividir el conjunto de datos en entrenamiento y prueba, se procedió a entrenar y evaluar los modelos siguiendo los pasos detallados a continuación:

1. **Evaluación mediante Búsqueda por Rejillas y Validación Cruzada:** Se empleó GridSearchCV para llevar a cabo una búsqueda exhaustiva de los mejores parámetros del modelo. Esta técnica utiliza las combinaciones de parámetros definidas previamente y evalúa todas las posibles combinaciones para identificar aquellas que maximizan el rendimiento del modelo. Con los mejores parámetros obtenidos, se procedió a realizar nuevas predicciones sobre el conjunto X_test.
2. **Predicción y Evaluación Final:** Utilizando los parámetros óptimos obtenidos con GridSearchCV, se realizaron predicciones finales sobre el conjunto X_test. Las nuevas predicciones se evaluaron nuevamente con la Matriz de Confusión y las Métricas de Evaluación de clasificación. Esta evaluación final proporciona una medida más precisa del rendimiento del modelo ajustado a los datos.

Acerca del entrenamiento del modelo

En el proceso de entrenamiento de los algoritmos primero se verificó el ajuste para distintos tamaños de datos. Esto resulta fundamental en el desarrollo y evaluación de modelos de aprendizaje automático, ya que proporciona una comprensión profunda de la capacidad de generalización del modelo. Al evaluar el rendimiento del modelo en diferentes tamaños de datos, es posible identificar si el modelo está sobreajustado (overfitting) o subajustado (underfitting).

Cabe mencionar que el conjunto de datos está desbalanceado por lo que a pesar de aplicar técnicas de oversampling se observa un comportamiento atípico (métricas altas para bajo porcentaje de entrenamiento).

Como se puede observar en la siguiente imagen en el Modelo SVC, el valor de la métrica no sube más partir de 65% de datos de entrenamiento.

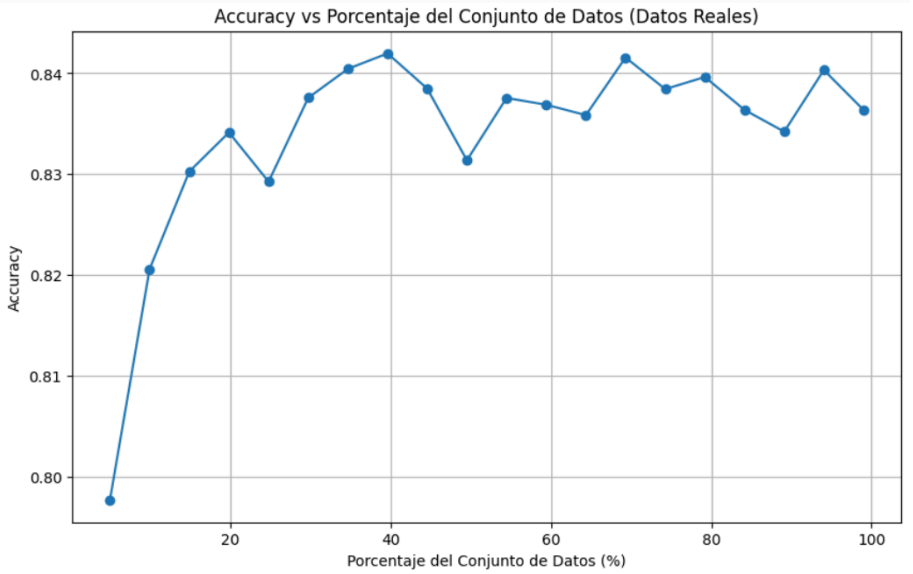


Figura 1: Acuracy para distintos tamaños de entrenamiento en el Support Vector

La siguiente imagen representa al Modelo Random Forest, y el valor de la métrica a partir de 75% no muestra una subida relevante.

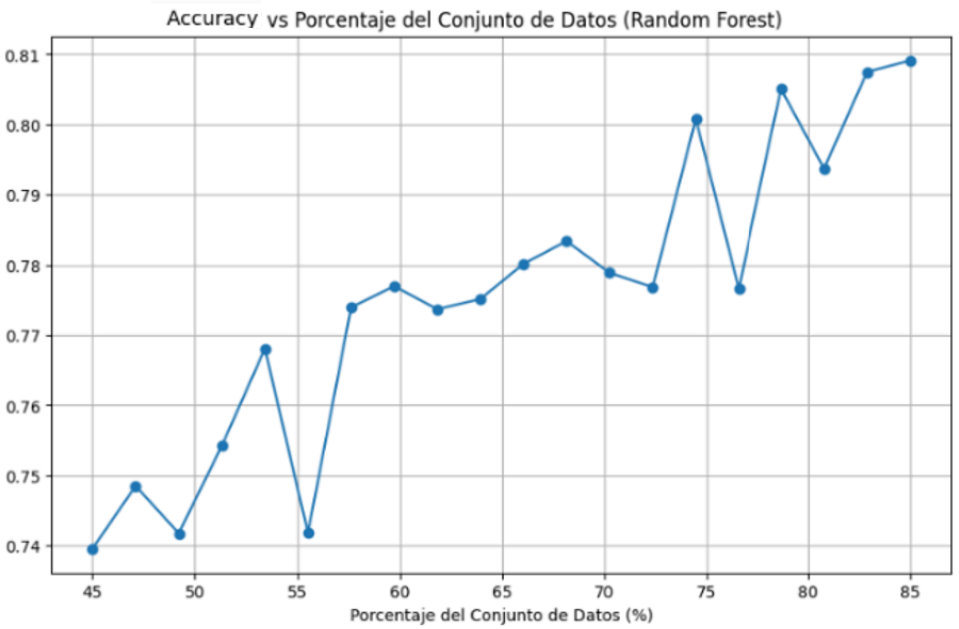


Figura 2: Acuracy para distintos tamaños de entrenamiento en el Random Forest

La siguiente imagen representa al Modelo Naive Bayes, y el valor de la métrica a entre de 75% y 90% es bastante similar.

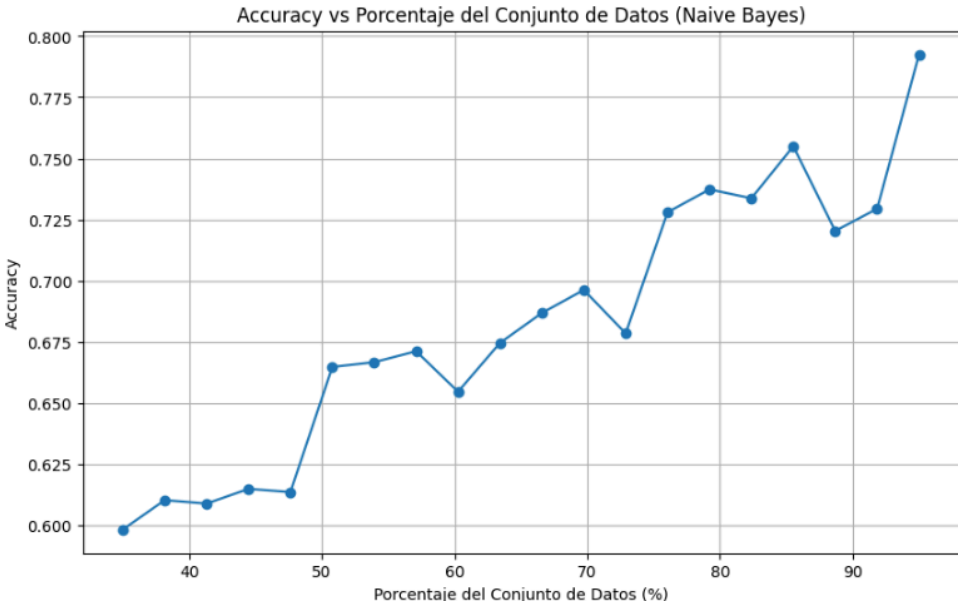


Figura 3: Acuracy para distintos tamaños de entrenamiento en el modelo NB

Luego de verificar el valor del porcentaje de entrenamiento óptimo, se realizó un Grid Search para ajustar los hiper parámetros que maximicen el rendimiento del modelo. Este funciona realizando una combinación de hiper parámetros en la cuadrícula, el modelo se entrena y se evalúa utilizando validación cruzada. Se calcula una métrica de rendimiento (por ejemplo, precisión, AUC, F1-score) para cada combinación de hiper parámetros basada en la validación cruzada. Esta métrica ayuda a identificar qué combinación de hiper parámetros produce el mejor rendimiento.

Una vez obtenemos el modelo con hiper parámetros óptimos, procedemos al entrenamiento de este para su posterior predicción sobre los valores de prueba.

Evaluación al modelo con conjunto de prueba

La evaluación de los modelos se lleva a cabo utilizando un conjunto de prueba independiente, lo que proporciona una medida objetiva de su rendimiento en datos no vistos. Se emplean diversas métricas de rendimiento, tales como precisión, recall, puntuación F1 y accuracy para cada modelo. Estas métricas no solo permiten evaluar la capacidad del modelo para clasificar correctamente las instancias, si no también identificar posibles desequilibrios en la predicción de las diferentes clases.

El proceso de evaluación se realiza luego de la aplicación de GridSearchCV para la optimización de los hiperparámetros de cada modelo. Esta búsqueda sistemática asegura que se seleccionen los parámetros óptimos que maximizan el rendimiento del modelo, permitiendo así obtener una evaluación precisa y robusta del modelo final.

Esta evaluación es para una partición del 20% para el conjunto de test. Un total de 209 filas, de las cuales 175 son de la clase 0 (“Attrition”=0), y 34 de la clase 1 (Attrition=1).

Modelo	Accuracy	Precision	Recall	F1 Score
Máquina de Soporte Vectorial (SVM)	0.86	0.88	0.97	0.92
Bosques Aleatorios	0.83	0.87	0.94	0.90
Naive Bayes	0.76	0.89	0.81	0.85

Tabla 1: Cuadro comparativo de las métricas obtenidas para cada modelo.

Comparación de resultados

En base a los resultados obtenidos por parte de los modelos aplicados para el conjunto de datos, se puede obtener las siguientes comparaciones:
Al utilizar SVM se obtuvo un accuracy del 86% que es un valor ligeramente menor que el modelo de [29] donde este obtuvo un 86,59%, por lo que para predecir si una persona se va o se queda en su labor actual este modelo resulta menos preciso que otros aplicados anteriormente.
En cuanto al modelo de Bosques Aleatorios se obtuvo en Accuracy de 83%, Precision de 87%, Recall de 94% y F1-Score de 90%. En [32] se obtuvo 90% en Accuracy lo cual es mayor que el resultado obtenido en este informe. Aún así, presenta mayor accuracy que el

obtenido en [29] donde se obtuvo 83.24% pero en la comparación de métricas en el mismo estudio se presenta mayor porcentaje en Precision (81%), Recall (75%) y F1 Score (78%); por lo que al utilizar Random Forest se presenta un menor valor de Accuracy pero mayor valor en las métricas restantes en la comparación con otros autores que utilizaron este modelo para el mismo objetivo.

Y para el último modelo utilizado que es Naive Bayes, el modelo presente obtuvo un Accuracy de 76%, Precision de 89% y un Recall de 81%. En [31] estas métricas son menores presentando un Accuracy de 69,7%, Precision de 55,1% y un Recall de 55,8% por lo que el Naive Bayes utilizado en el presente artículo presenta mejores resultados en estas métricas para la predicción de deserción laboral.

Discusión

Metodología

Algoritmos usados

En este estudio se utilizaron varios algoritmos de machine learning, incluyendo Support Vector Machines (SVM), Random Forest y Naive Bayes. Al comparar nuestro trabajo con otros estudios, encontramos que [29] también utilizó SVM, Random Forest y Naive Bayes pero se empleó también Regresión Logística, KNN y Decision Tree. Por otro lado, [31] empleó una gama más amplia de algoritmos, tales como regresión logística, árboles de decisión, KNN, SVM, Random Forest y Naive Bayes. Observamos que el algoritmo Random Forest y el algoritmo Naive Bayes son recurrentes en varios estudios, incluyendo el nuestro y el de [29], lo que sugiere una preferencia común por estos métodos en la comunidad de investigación de machine learning. Y mientras tanto en [33] se utilizaron tan solo los algoritmos de Random Forest y KNN.

Procedencia del dataset

En este estudio, se utilizó el conjunto de datos ficticio "IBM HR Analytics Employee Attrition & Performance" proporcionado por IBM, el cual contiene casi 30 características categóricas y discretas. Este mismo conjunto de datos fue utilizado en [29] y en [33], y es ampliamente empleado en la investigación del aprendizaje automático, estando disponible en plataformas como Kaggle. Este conjunto de datos incluye características relacionadas con la demografía de los empleados, satisfacción laboral, desempeño, entorno laboral, y otros factores que pueden influir en la rotación,

con una variable objetivo binaria que indica si un empleado ha dejado la empresa o no.

Por otro lado, [31] utilizó una muestra de 3543 registros de desempeño y actividad de ventas de agentes de un call center, en campañas de venta de seguros y telefonía celular, recopilados entre mayo de 2010 y noviembre de 2011. Después de eliminar datos inconsistentes, la muestra resultante fue de 2407 registros. Cada registro incluye actividades mensuales de los agentes, con todas las variables estandarizadas y divididas por el total de horas trabajadas en el período de análisis.

Variables usadas

En este estudio se utilizaron 15 variables, los cuales ya fueron especificadas y detalladas previamente entre los originales del dataset y añadidos. En [29], se utilizaron tan solo se utilizaron 6 variables, a diferencia de nuestro estudio este estudio utilizó 4 variables que no fueron consideradas para este estudio las cuales son: Promotion/No promotion, Age, Gender, Tenure. A su vez, en [31] y en [33] también se utilizaron 6 variables del dataset que se utilizó para el pronóstico de deserción: Age, Salary, Marital Status, Gender, Distance y Turnover siendo esta última la variable dependiente.

Resultados

En este estudio, después de evaluar los modelos obtuvimos que: la Máquina de Soporte Vectorial (SVM) obtuvo accuracy del 86%, una precision del 88%, un recall del 97% y F1-Score del 92%. El modelo de Bosques Aleatorios mostró un accuracy del 83%, precision del 87%, un recall del 94% y una F1-Score del 90%. Por otro lado, el modelo Naive Bayes alcanzó un accuracy del 76%, precision del 89%, un recall del 81% y una F1-Score del 85%.

En [29], se reportaron las siguientes precisiones para distintos modelos: la Regresión Logística con un accuracy del 87.71%, el clasificador KNN obtuvo 59.22%, las Máquinas de Soporte Vectorial obtuvieron 86.59%, Naive Bayes obtuvo 83.24%, los Árboles de Decisión alcanzó 80.45% y Random Forest, 83.24%.

En [31], se obtuvieron los siguientes resultados para los modelos: Random Forest alcanzó un accuracy del 66.2%, precision del 81.6% y un recall del 71.1%, mientras

que Naive Bayes mostró un accuracy del 66.4%, precision del 51.1% y un recall del 55.8%.

Finalmente, en el estudio [33], el modelo Random Forest presentó un accuracy del 84%, precision del 47%, un recall del 12% y un F1-Score del 18.7%, mientras que el modelo KNN mostró un accuracy del 80%, un precision del 33.3%, un recall del 24.9% y un F1-Score del 28.1%.

V. CONCLUSIONES

Este estudio tuvo como objetivo predecir la probabilidad de renuncia de los empleados según los factores identificados para ayudar a las organizaciones a desarrollar estrategias efectivas de retención, minimizando los costos de rotación y mejorando la moral y productividad del equipo .La Máquina de Soporte Vectorial (SVM) demostró ser el modelo más eficaz con un accuracy del 86%, seguida de cerca por el modelo de Bosques Aleatorios con un accuracy del 83%. Aunque el modelo Naive Bayes tuvo un accuracy ligeramente inferior del 76%, aún mostró resultados aceptables. A lo largo del proceso, se realizó un exhaustivo preprocesamiento de los datos, incluyendo la identificación y corrección de datos faltantes, la estandarización y corrección de valores categóricos, y la eliminación de outliers. Estas acciones aseguraron la integridad y precisión de los datos, lo cual es crucial para obtener resultados fiables. En general, se logró el objetivo del estudio, identificando los modelos más eficaces y proporcionando a las organizaciones herramientas valiosas para mejorar la retención de empleados.

VI. REFERENCIAS BIBLIOGRÁFICAS

[1] B. D. Narvaéz Montenegro y J. d. C. Castillo Celi, «La terminación de la relación laboral por acuerdo entre las partes, previa renuncia del trabajador» Institucional UNIANDES, Septiembre 2017. [En línea]. Available: https://dspace.uniandes.edu.ec/handle/123456789/6539.

[2] J. D. Vaamonde, Intenciones de Renunciar al Trabajo: «Propiedades Psicométricas de una Escala y Relaciones con la Percepción de Apoyo y la Identificación Organizacional» Revista Interamericana de Psicología Ocupacional, 3 Mayo 2019. [En línea]. Available: http://revista.cincel.com.co/index.php/RPO/article/view/233.

[3] C. M. Villegas Arriola y M. D. Huaman Bazán, «Satisfacción e inseguridad laboral como variables explicativas de la intención de renunciar al trabajo en jóvenes» Repositorio Académico UPC, 9 Septiembre 2020. [En línea]. Available: http://doi.org/10.19083/tesis/652967.

[4] M. E. Rey Caldeyro, «Publicación: Análisis de predicción aplicado a la deserción de empleados»Docta Complutense Madrid, Septiembre 2021. [En línea]. Available: https://docta.ucm.es/entities/publication/9eefb94d-e1ff-44df-b21b-0374c5a96da9.

[5] M. Nader, S. P. Peña Barnate y E. S. Santa-Bártara, «Predicción de la satisfacción y el bienestar en el trabajo: hacia un modelo de organización saludable en Colombia»ScienceDirect 12 Marzo 2014. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S0123592314000461.

[6] R. Flores Zambada, Factores de la calidad de vida en el trabajo como predictoras de la intención de permanencia «UGTO México,» 1 Marzo 2012. [En línea]. Available: https://doi.org/10.15174/au.2012.363.

[7] M. J. Aguilar Camacho, J. E. Luna Correa, E. Barrera Arias y A. R. Tovar Vega, Deserción laboral de mandos medios. Caso de empresa de autopartes «Dialnet,» 2018. [En línea]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=7785852.

[8] C. Rueda Marín, D. Giraldo Gil, Percepción respecto a los factores organizacionales que generan renuncia silenciosa en un grupo de líderes del área de gestión humana de una empresa privada de la ciudad de Medellín, en tiempo de post - pandemia «Dialnet,» 2023. [En línea]. Available: https://repository.uniminuto.edu/handle/10656/17765.

[9] S. Mori Mego La rotación de personal en una empresa manufacturera, Lima - 2022 «Repo UCV» 2023. [En línea]. Available: https://hdl.handle.net/20.500.12692/94954.

[10] M. Gómez López, Educación ambiental y laboral desde una ética del cuidado en la mejora de una productividad laboral «IDuS» 2023. [En línea]. Available: https://hdl.handle.net/11441/157823.

[11] A. Cabanillas Torres. Desarrollo de un sistema inteligente para evaluación de los perfiles por competencia laboral en un puesto gerencial. «USS» 2022. [En línea]. Available: https://hdl.handle.net/20.500.12802/10467.

[12] G. Pezo Beltrán. Factores laborales que explican la rotación del personal del centro de aplicación Unión. «Universidad Peruana Unión» 2019. [En línea]. Available: http://repositorio.upeu.edu.pe/handle/20.500.12840/2997.

[13] M. Siahaan. An Analysis of Contract Employee Performance Assessment Using Machine Learning «UMA» 2021. [En línea]. Available: https://doi.org/10.31289/jite.v5i1.5357.

[14] S. Rivas Cuevas. Aplicación de métodos NLU en la recomendación de CVs para la selección de personal «UVADOC» 2022. [En línea]. Available: https://uvadoc.uva.es/handle/10324/57977.

[15] K. Matute-Pinos; R. Bojorque-Chasi. Apoyo a los subsistemas de talento humano, selección y reclutamiento a partir de un sistema experto. Caso de estudio. «Scielo» 2021. [En línea]. Available: https://doi.org/10.17163/ings.n26.2021.04.

[16] D. Cortes, C. Agudelo Jimenez, L. Yarpaz, L. Ramirez, L. Rodriguez Pedraza. Mejoramiento del Proceso de Reclutamiento y Selección de Personal de la Empresa Intermegamundo «UNSA Repository» 2020. [En línea]. Available: https://repository.unad.edu.co/handle/10596/41893.

[17] D. Pessach, G.Singer, D. Avrahami, H. Chalutz Ben-Gal, E. Shmueli, I. Ben-Gal. Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming «ScienceDirect» 2020. [En línea]. Available: https://doi.org/10.1016/j.dss.2020.113290.

[18] P. Asto Machaca. Modelo de sistemas M-Learning para el reclutamiento y selección de talento humano / caso: Entrevistas «UNSA Repository» 2020. [En línea]. Available: http://hdl.handle.net/20.500.12773/11528.

[19] S. Krishna Kumar, E. Ramaraj, S. Santhoshkumar, P. Geetha, A decision tree with filter attribution selection model for designing an automated job recommender system for candidate matching through online recruitment «Jatit» 15 Marzo 2024. [En línea]. Available: https://jatit.org/volumes/Vol102No5/30Vol102No5.pdf.

[20] M. Elzaurdi Carrera. Diseño, Implementación y Validación de algoritmos de aprendizaje automático en medicina «Bucle» 2021. [En línea]. Available: http://hdl.handle.net/10366/148561.

[21] S. Diaz Romero, W. Sanyer Mosquera, J. Menéndez Campos. Selección de candidatos para encuestas mediante técnicas de machine learning «DSpace» 2021. [En línea]. Available: http://www.dspace.espol.edu.ec/handle/123456789/54068.

[22] J. Fajardo Vargas, Inteligencia artificial aplicada al proceso de selección de personal «Dialnet» 23 de Setiembre 2022. [En línea]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=9152551.

[23] P. P. J. S. P. M. Shruti Rao, «A comparative study between various preprocessing techniques for Machine Learning» IJEAST, pp. 431, 434, 2020. Available: https://www.ijeast.com/papers/431-438,Tesma503,IJEAST.pdf

[24] M. Botero Arbeláez, O. A. Salazar, y J. A. Mendoza Vargas, «Método anova utilizado para realizar el estudio de repetibilidad y reproducibilidad dentro del control de calidad de un sistema de medición», Sci. tech, vol. 1, n.º 37, dic. 2017. [En línea]. Available: https://revistas.utp.edu.co/index.php/revistaciencia/article/view/4181

[25] Badii, M.H., A. Guillen & J.L. Abreu, Aplicación de ANOVA Anidada en la Investigación Científica , Daena: International Journal of Good Conscience. 9(2) 12-17. Agosto 2019. ISSN 1870-557X [En línea] Available: http://www.spentamexico.org/v9-n2/A2.9(2)12-17.pdf

[26] T. K. Or Shkoler, «How Does Work Motivation Impact Employees' Investment at Work and Their Job Engagement? A Moderated-Moderation Perspective Through an

International Lens,» 20 Febrero 2020. [En línea]. Available: https://doi.org/10.3389/fpsyg.2020.00038 [Último acceso: May 2024].

[27] Q. Meng, H. Zhu, K. Xiao y H. Xiong, «Intelligent Salary Benchmarking for Talent Recruitment: A Holistic Matrix Factorization Approach,» IEEE Explore, 2018. [En línea]. Available: https://doi.org/10.1109/ICDM.2018.00049.

[28] I. Dabbura, «Predicting Employee Turnover,» Medium, 2018. [En línea]. Available Online: https://towardsdatascience.com/predicting-employee-turnover-7ab2b9ecf47e

[29] S. B. Parmod Kumar, «Predicting Employee Turnover: A Systematic Machine Learning Approach for Resource Conservation and Workforce Stability,» 26 october 2023. [En línea]. Available: https://www.mdpi.com/2673-4591/59/1/117. [Último acceso: may 2024].

[30] R. Jafari-Marandi a, M. Aliannejadi b, I. Ahmadian c, M. Mozaffari a, U. Keramati, «Improved churn prediction in telecommunication industry using data mining techniques ", Applied Soft Computing, vol. 24, pp. 994-112, 2016. https://doi.org/10.1016/j.asoc.2014.08.041

[31] Valle, M. A., & Ruz, G. A. . «Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms.» Applied Artificial Intelligence, 29(9), 923–942. 2016. [En línea] Available: https://doi.org/10.1080/08839514.2015.1082282

[32] R. Chakraborty, K. Mridha,a R. N. Shaw y A. Ghosh, «Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches,» 2021. [En línea]. Available: https://doi.org/10.1109/GUCON50781.2021.9573759. [Último acceso: May 2024].

[33] D. E. S. O. Markus Atef, «Early Prediction of Employee Turnover Using Machine Learning Algortihms» hrack.srce, vol. 13, 22. Available: https://doi.org/10.32985/ijeces.13.2.6

[34] T. Wood, «Predicting employee turnover,» Fast Data Science, 2020. [En línea]. Available: https://fastdatascience.com/predicting-employee-turnover.

[35] Ajmal M S, TANMAY DESHPANDE, IBM Data Scientists, February 17, 2023, "IBM HR Analytics Employee Attrition & Performance", IEEE Dataport, doi: https://dx.doi.org/10.21227/2m1g-6v47.

[36] Boser BE, Guyon IM, Vapnik VN.«A training algorithm for optimal margin classifiers,» In Proceedings of fifth COLT. Pittsburgh, PA, 1992, 144–152. https://doi.org/10.1145/130385.130401.