UNIVERSIDAD DE ANTIOQUIA

APLICACIÓN DE LA ANALÍTICA
Departamento Ingeniería Industrial

Natalia Guzmán| 1000046149
Mateo Caicedo Aguirre | 1107101137
Juan José Toro | 1017271652
Sebastián Cuartas | 1152465726

MODELO PARA PREDECIR LA DESERCIÓN LABORAL

PROBLEMA DE NEGOCIO

La empresa enfrenta una alta tasa de rotación del 15% anual, lo que genera costos significativos en términos de contratación, capacitación y pérdida de productividad. Este nivel de deserción impacta tanto la eficiencia operativa como la conexión del equipo. Los directivos buscan reducir esta tasa para mejorar la retención del personal y disminuir los costos asociados. Por lo tanto, se necesita una herramienta predictiva que permita identificar a los empleados con mayor riesgo de retiro, facilitando la implementación de acciones preventivas personalizadas para aumentar la retención de estos empleados clave.

PROBLEMA ANALÍTICO

Para abordar el problema de la alta rotación, se proponen dos enfoques analíticos:

Modelo de Clasificación Supervisada: Utilizando técnicas como regresión logística, árboles de decisión o Random Forest, se busca predecir las renuncias totales para el próximo año. Este modelo permitirá identificar empleados con alta probabilidad de retiro.

Análisis de Variables Influyentes: Además de la predicción, se realizará un análisis para identificar qué variables (como satisfacción laboral, equilibrio entre vida personal y trabajo, nivel de satisfacción con el ambiente laboral, entre otras) tienen un impacto significativo en la probabilidad de que un empleado se retire.

El modelo ayuda a resolver el problema realizando predicción de riesgo lo que permitirá a los directivos priorizar recursos en aquellos empleados con mayor probabilidad de retiro, identificando a aquellos con "alta probabilidad de deserción". También se entregarán resultados prácticos en un formato accesible, como un archivo Excel, que contendrá una lista de empleados con mayor riesgo. Esto permitirá que el área de recursos humanos pueda tomar acciones específicas para retenerlos y poder tomar acciones preventivas con la identificación de las variables que más influyen en los retiros.

DISEÑO DE NEGOCIO

El diseño de negocio se enfocará en clasificar e**l riesgo de deserción** de los empleados en tres niveles:

Probabilidad Alta: Se requiere una intervención inmediata, que puede incluir ajustes salariales, incentivos adicionales u otras acciones directas para retener al empleado.

Probabilidad Media: Se recomienda un monitoreo continuo, ofreciendo capacitaciones o oportunidades de desarrollo personal para mejorar la satisfacción y el compromiso del empleado.

Probabilidad Baja: Se aplicarán medidas preventivas como encuestas de clima laboral periódicas y programas de reconocimiento. Para mantener la satisfacción y prevenir posibles problemas futuros.

Con base en esta clasificación, y a partir de las predicciones y análisis del modelo, el equipo de recursos humanos podrá diseñar intervenciones personalizadas que van desde ajustes salariales hasta programas de desarrollo profesional y bienestar. Al identificar y abordar proactivamente los factores que influyen en la rotación, la empresa no solo mejorará la retención, sino que también fortalecerá su cultura organizacional, convirtiéndola en un entorno más atractivo y comprometido para los empleados.

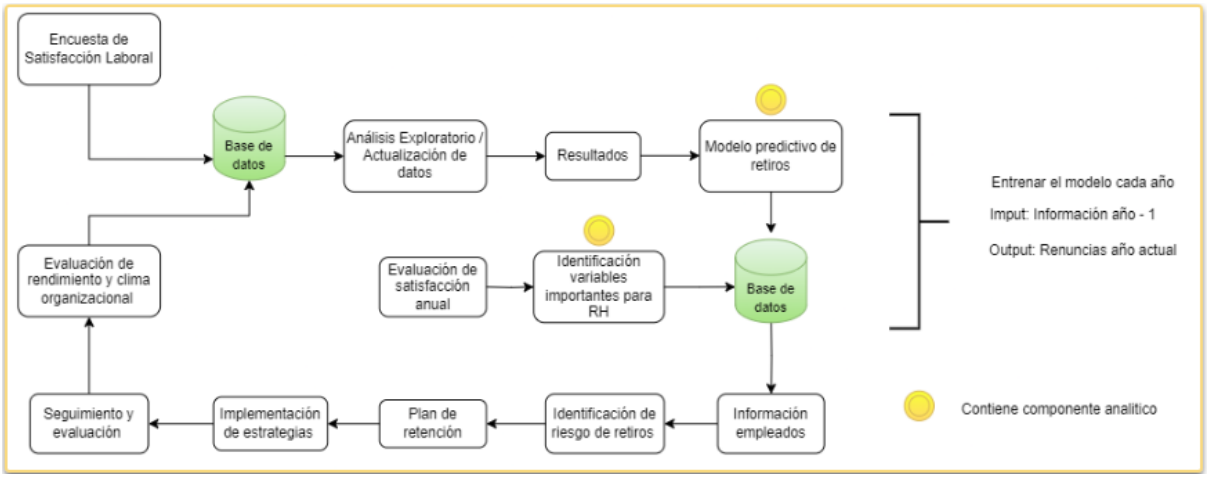


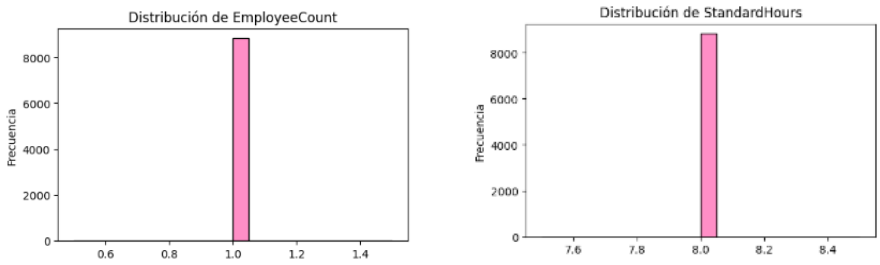
Gráfico 1: Diagrama propuesto

LIMPIEZA Y TRANSFORMACIÓN DE DATOS

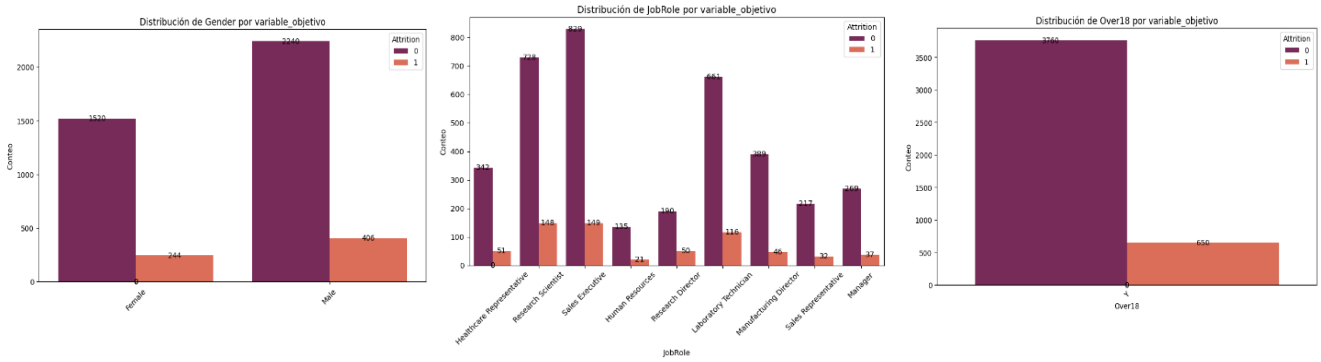
Se inicia con la revisión del conjunto de datos históricos de los empleados tanto de los que continúan trabajando, como los que se retiran. Después se realiza el preprocesamiento de los datos con SQL con el fin de dejar solo una variable que contenga todos los datos para facilitar de gran forma el proceso, esto se logra con la unión usando Left Join. Posteriormente haciendo el análisis se eliminan las primeras variables que son las variables de retiros: [**retirementDate, retirementType y resignationReason]** ya que no son significantes al no arrojar información que consideramos necesaria.

ANÁLISIS EXPLORATORIO

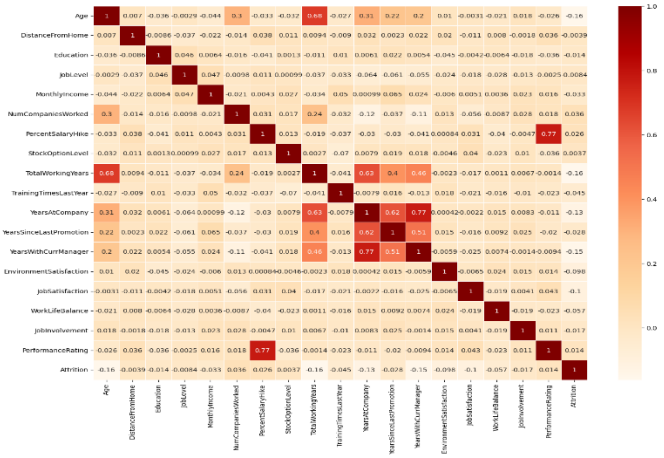
Realizamos una exploración de datos con la información proporcionada, donde se separan por años: **df1** vendrían siendo los empleados encuestados en el año 2015 y **df2** para los empleados encuestados en el año 2016. Se analizan por separado las variables numéricas como categóricas con la relación de la variable objetivo, para las variables numéricas se reemplazan los valores nulos con la media tanto para **df1** y **df2** que son **['NumCompaniesWorked', 'EnvironmentSatisfaction', 'JobSatisfaction', 'WorkLifeBalance'],** colocamos lavarible **['EmployeeID'] en un dataframe aparte** para la seguir con la transformación de datos y después volver a ingresarla en la base de datos, se definen las variables numéricas y las categóricas, se separan en un dataframe para los dos años y hacemos el histograma para los dos años y vemos que hay dos variables constantes por lo cual se eliminaron las variables **['StandardHours', 'EmployeeCount']** en **df_num** y **df_num2.**



Despùes de hacer las pruebas de Chi 2, se crearon gráficas para ver la correlación entre las variables categóricas y la variable objetivo **(Attrition)**, podemos ver que algunas variables no tienen poca relación como por ejemplo [**Gender, JobRole, Over18**] por lo cual se podría eliminar.



Se realiza una Matriz de correlación para las variables numéricas y se identificaron las variables [**PerformanceRating y PercentSalaryHike**] tiene una correlación alta por lo cual podemos decir que brindan información similar y por tanto una de las dos puede ser eliminada, lo mismo con las variables [**YearwithCurrManager, YearsAtCompany y TotalWorkingYears**]



Después pasamos los valores nulos a cero y convertimos a dummies las variables categóricas, por otro lado, escalamos las variables numéricas, después de eso concatenamos las dummies con las variables numéricas escaladas en cada año. después de eso le agregamos a cada dataFrame la variable employeeID

Definimos **X** y **Y** para el entrenamiento usando los datos de 2015, después definimos X_test para la predicción de datos de 2016, se definió el tamaño del conjunto de entrenamiento y el tamaño del conjunto de validación.

Modelos

Se implementaron inicialmente 3 modelos, los cuales son: **Regresión Logistica, Random Forest y XG Boosing Classifier,** cada uno de estos primeramente sin balanceo y luego con el balanceo correspondiente usando "class_weight". Posteriormente comparamos todos los modelos con las métricas **F1 Score, Precision, Recall y AUC RUC**, para seleccionar el modelo ganador. A continuación, una tabla comparativa:

Modelo	F1 Score	Precision	Recall	AUC RUC
Red Log	0.23	0.62	0.14	0.721
Red Log Balanceado	0.35	0.24	0.64	0.7126
Random Forest	0.95	0.99	0.92	0.983
Random Forest Balanceado	0.94	1.00	0.89	0.984
XG Boost	0.97	0.98	0.95	0.979
XG Boost Balanceado	0.96	0.98	0.95	0.978

Tabla comparativa 1

Se selecciona el modelo de XG Boost sin balanceo, ya que:

Tiene el F1 Score más alto (0.97), lo que indica un buen equilibrio entre precisión y recall. Esto es importante, dado que el F1 score captura tanto la capacidad del modelo para identificar correctamente los retiros (recall) como para minimizar falsos positivos (precision).

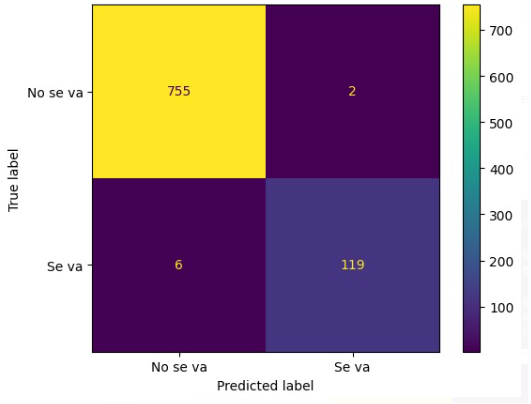
Su **Precision** Es alta (0.98), lo que significa que, de las predicciones de retiros la gran mayoría son correctas. **Recall**: Tiene un valor también muy alto (0.95), indicando que el modelo está capturando casi todos los casos de retiros reales. **AUC ROC**: Está cerca de 1 (0.979), lo que significa que el modelo tiene una muy buena capacidad de distinguir entre quienes se retiran y quienes no.

Afinamiento de Hiperparámetros

Después del análisis anterior se realiza el tuning del modelo **XG Boosting Classifier** sin balanceo y vemos que no se obtiene una mejora considerable, esto debido a que se encontraban en un punto óptimo para nuestro conjunto de datos. En algunos casos, el ajuste de hiperparámetros (tuning) puede llevar a un sobreajuste, donde el modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad de generalización en los datos de prueba.

Reporte de clasificación:

	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	757
1.0	0.98	0.95	0.97	125
accuracy			0.99	882
macro avg	0.99	0.97	0.98	882
weighted avg	0.99	0.99	0.99	882



CONCLUSIONES

El modelo **XGBoost** sin ajuste de hiperparámetros ofreció el mejor rendimiento en la predicción de los empleados que se retirarán. A pesar de haber explorado diversas técnicas de ajuste, como el balanceo de clases y la optimización de hiperparámetros, el modelo sin ajustes logró un balance óptimo entre precisión, recall y F1-score, especialmente en la clase de empleados que se retiran (1.0), lo que sugiere que los valores predeterminados fueron adecuados para el conjunto de datos utilizado.

ESTRATEGIAS

Desarrollo de programas de retención personalizados: Dado que el modelo ha identificado características clave que influyen en el retiro de empleados, la empresa podría desarrollar programas de retención enfocados en los factores más relevantes. Por ejemplo, si se ha detectado que la satisfacción laboral o el equilibrio entre vida y trabajo son predictores importantes, se podrían implementar programas para mejorar el bienestar de los empleados en estas áreas, como flexibilización laboral, oportunidades de desarrollo profesional, o mejoras en el ambiente de trabajo.

Monitoreo proactivo de empleados con alto riesgo de retiro: Utilizando el modelo predictivo, la empresa puede implementar un sistema de alerta temprana para identificar a los empleados con mayor probabilidad de retirarse. Al tener esta información, los gerentes podrían intervenir de manera proactiva mediante conversaciones individuales, planes de carrera personalizados, o incentivos para retener a estos empleados, antes de que tomen la decisión de abandonar la empresa.

Enlace Repositorio : https://github.com/juantoro5/Analitica-para-RH-E7.git