Reporte del Modelo de Employee Attrition

José Ángel González Rodríguez – angel.gonzalez.eco@outlook.com

I. INTRODUCCIÓN

La tasa de abandono es la tasa a la que los empleados de una compañía abandonan sus puestos de trabajo de forma voluntaria. Las compañías destinan recursos al proceso de contratación de su personal, por lo que es fundamental comprender las motivaciones de los empleados que deciden dejar la compañía. Esto se hace con la finalidad de retener la mayor cantidad de talento o, en su defecto, anticipar la búsqueda de nuevo personal. Por lo tanto, el objetivo de este reporte es presentar varios modelos de clasificación y medir su desempeño en función de diversas métricas. Esto se hace con el fin de identificar las principales características que llevan a los empleados a abandonar la compañía.

II. DATOS

La base de datos es proporcionada por la compañía en el *EmployeeAttrition* dataset.

```
['Age',
 'Attrition',
 'BusinessTravel',
 'Department',
 'DistanceFromHome(KMs)',
 'Education',
 'EducationField',
 'EnvironmentSatisfaction',
 'Gender',
 'JobInvolvement',
 'JobLevel',
 'JobRole',
 'JobSatisfaction',
 'MaritalStatus',
 'MonthlyIncome',
 'MonthlyRate',
 'NumberCompaniesWorked',
 'PercentSalaryHike',
 'PerformanceRating',
 'RelationshipSatisfaction',
 'TotalWorkingYears',
 'TrainingTimeLastYear(Weeks)',
 'WorkLifeBalance',
 'WorkedOverTime',
 'YearsAtCompany',
 'YearsInCurrentRole',
 'YearsSinceLastPromotion',
 'YearsWithCurrentManager']
```

Figura 1. Columnas de Employee Attrition dataset

El dataset contiene 1470 filas y 28 columnas, las cuales contienen datos numéricos, datos categóricos ordinales y datos categóricos nominales. La Figura 1 muestra las columnas del dataset.

III. ANÁLISIS EXPLORATORIO

La variable objetivo es 'Attrition', la cual es una variable categórica que contiene 'Yes' para los individuos que ahora son exempleados, mientras que 'No' es asignado a los empelados actuales. La distribución de abandono en la compañía se observa en la Figura 2. Se observa que alrededor del 16% de los individuos son exempleados, mientras que alrededor del 84% siguen en la compañía[1] (Son datos naturalmente desbalanceados pues, en general, habrán más personas dentro de la compañía que las que están renunciando).

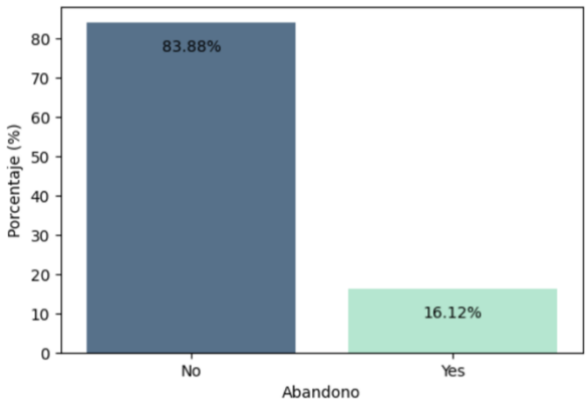

Figura 2. Porcentaje de abandono en la compañia

Además, fueron realizadas algunas gráficas donde se muestran distribuciones de variables, gráficas de caja y gráficas de barras con base en el abandono de empleados de la compañía.

[1] Los datos desbalanceados presentan un problema para el rendimiento en un modelo, por lo que en la sección de Metodología será gestionado este problema.

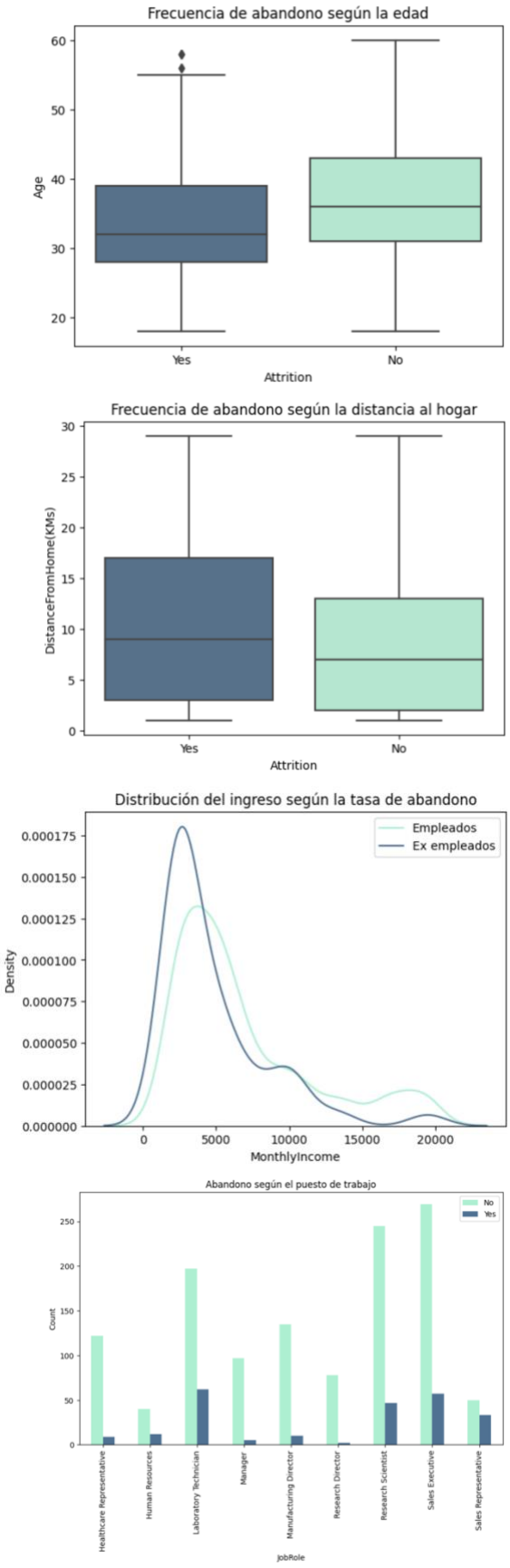

Figura 3. Gráficas análisis exploratorio

En la Figura 3 se puede observar cómo, en promedio, los empleados que tienen alrededor de 28 años son más propensos de renunciar que los empleados de mayor edad. Además, se observa que los empleados que más alejados de la compañía son más propensos en renunciar a sus empleados. La siguiente gráfica muestra que los empleados que reciben salarios más bajos también tienden a abandonar la compañía. Finalmente, la ratio de abandono en la compañía es mayor en los empleados con el cargo de Sales Representative que en otros cargos.

La matriz de correlación en la figura 4 muestra el grado de relación entre las variables. Destaca que 1) el ingreso mensual está correlacionado con la edad y con el tipo de cargo, 2) el aumento salarial está correlacionado con el desempeño de los empleados y 3) el total de años trabajados está correlacionado con los años trabajando con el mánager actual, años desde la última promoción, años trabajando en el mismo cargo y años trabajados en la compañía.

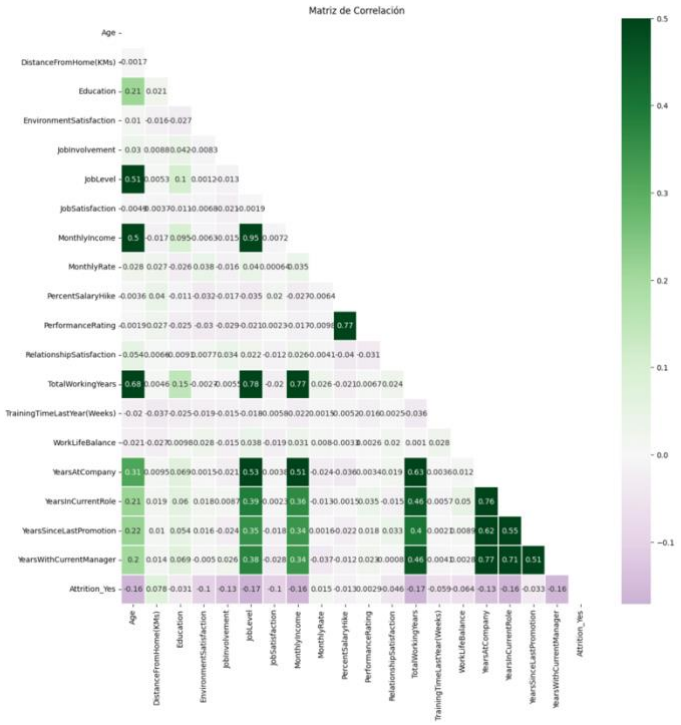

Figura 4. Matriz de correlación

IV. PREPROCESAMIENTO DE DATOS

1. Feature Selection

Feature selection es el proceso de elegir al conjunto de variables relevantes para entrenar el modelo, con la finalidad de eliminar las variables innecesarias o redundantes. Una técnica para identificar a las variables es la prueba de correlación que se muestra en la figura 4. Se observa una alta correlación entre los años trabajados y las variables antes mencionadas, por lo que eliminaré esas columnas para evitar problemas de multicolinealidad. Además, hay variables como MonthlyRate y PerformanceRating que tienen una correlación muy baja con la variable objetivo, por lo que pueden ser descartadas.

2. Feature Engineering

Feature Engineering es el proceso de crear nuevas características a partir de los datos originales con la finalidad de mejorar la capacidad predictiva del modelo. Para ello, fueron aplicadas las técnicas de One Hot Encoding y estandarización.

2.1 One Hot Encoding

Esta técnica transforma las variables categóricas a un formato numérico, el cual es esencial para entrenar los modelos de Machine Learning dado que estos algoritmos requieren que las variables de entrada estén en formato numérico para poder aprender y realizar predicciones de manera efectiva. La técnica fue aplicada a todas las variables categóricas nominales.

2.2 Estandarización

Es un proceso en el que los valores de una variable son normalizados, es decir, que tengan media cero y una desviación estándar de uno. Este paso es fundamental para eliminar las diferencias en las escalas de las variables, lo cual permite la comparación justa entre las características, pues elimina la influencia de la magnitud de variables con valores grandes en el modelo.

V. METODOLOGÍA

Fueron entrenados y evaluados cinco modelos de clasificación de aprendizaje automático supervisado con la finalidad de comparar y elegir el de mejor desempeño. Fue desarrollado el modelo de Regresión Logística que clasifica a una instancia a alguna clase según la probabilidad de pertenecer a ella; Decision Tree, que crea un árbol de reglas de decisión basado en características de clasificación de datos; Vecinos más cercanos (KNN), que clasifica según la mayoría de las clases entre sus K vecinos más cercanos, Support Vector Machines (SVM), que encuentra un hiperplano óptimo para separar clases en el espacio de las características y Random Forest, que crea conjuntos de árboles de decisión para mejorar la precisión y evitar el sobreajuste.[2] Posteriormente, el desempeño de los modelos fue evaluado a través de cuatro métricas fundamentales: Accuracy, Precision, Recall y F1-Score las cuales proporcionan una visión general de la capacidad predictiva de los modelos.

Balanceo de datos

La variable objetivo *Attrition* cuenta con un problema de desbalance pues las personas exempleadas conforman aproximadamente el 16% de las instancias, por lo que es importante balancear los datos para generar modelos donde la influencia de la clase mayoritaria (personas actualmente empleadas) no reduzca la importancia de la clase minoritaria (exempleados). De este modo, son realizados 15 modelos de clasificación. Los cinco modelos son sometidos a las técnicas de balanceo por undersampling y oversampling; además de que son incluidos los modelos con datos en desbalance.

[2] Danny Varghese, 2018, Comparative Study on Classic Machine learning Algorithms, https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

VI. RESULTADOS

Los resultados de los modelos de clasificación son presentados en las tablas 1, 2 y 3.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.717	0.731	0.720	0.725
Decision Tree	0.915	0.872	0.981	0.923
KNN	0.864	0.817	0.953	0.880
SVM	0.713	0.725	0.720	0.723
Random Forest	0.972	0.966	0.981	0.973

Tabla 1. Modelos de clasificación con datos balanceados (oversampling)

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.716	0.762	0.653	0.703
Decision Tree	0.621	0.618	0.694	0.654
KNN	0.663	0.707	0.592	0.644
SVM	0.726	0.756	0.694	0.723
Random Forest	0.716	0.762	0.653	0.703

Tabla 2. Modelos de clasificación con datos balanceados (undersampling)

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.816	0.750	0.103	0.182
Decision Tree	0.741	0.275	0.190	0.224
KNN	0.803	0.500	0.155	0.237
SVM	0.803	0.000	0.000	0.000
Random Forest	0.810	0.600	0.103	0.176

Tabla 3. Modelos de clasificación con datos no balanceados

De acuerdo con los resultados de los modelos de clasificación, el mejor método para obtener mayor precisión de predicción es utilizar la técnica de oversampling. La ventaja de esta técnica es que evita la pérdida de información y mejora la detección de la clase minoritaria, la cual nos interesa predecir con exactitud.[3]

Así, se observa que Random Forest bajo la técnica de balanceo de datos por oversampling es el modelo que tiene la mayor capacidad predictiva, con una calificación de accuracy del 97%, precisión del 97%, Recall del 98% y F1 Score del 97%. La matriz de confusión de este modelo se puede observar en la figura 5.

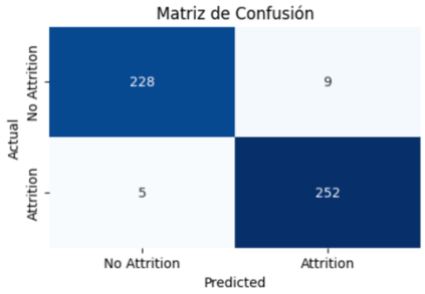

Figura 5. Matriz de confusión de Random Forest (oversampling)

Por esta razón, Random Forest balanceado (oversampling) es el modelo propuesto para que la compañía utilice para clasificar a sus empleados de acuerdo con la probabilidad de que abandonen la compañía.

VII. FEATURE IMPORTANCE

En la Figura 6 se presentan las principales razones que inciden en la decisión de los empleados de abandonar la compañía. Los ingresos mensuales destacan como el factor principal, ya que los salarios más bajos inducen a los empleados a abandonar la compañía. La edad, la experiencia laboral total y dentro de la empresa explican el abandono pues los empleados jóvenes y menos experimentados muestran una tendencia a no permanecer durante largos períodos en una sola empresa. La distancia al hogar también es relevante, ya que vivir más lejos tiende a llevar a las personas a considerar otras oportunidades. Además, factores como el aumento porcentual del salario, la satisfacción laboral y en el entorno, la antigüedad en el puesto actual y otros elementos, constituyen las principales variables que explican la tasa de abandono de los empleados.

[3] Cabe destacar que una desventaja de esta técnica es que puede llevar a un sobreajuste (overfitting) lo cual puede dificultar la generalidad del modelo. Por lo tanto, es vital tomar a consideración las demás técnicas de balanceo.

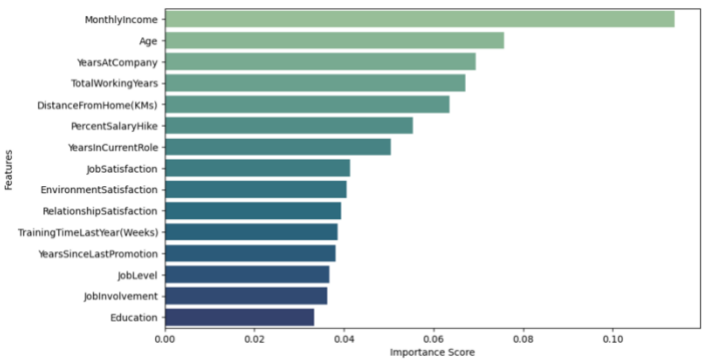

Figura 6. Top características más importantes

VIII. CONCLUSIÓN

En conclusión, el objetivo de este reporte fue presentar diversos modelos de aprendizaje automático con la finalidad de determinar cual tiene mejor rendimiento. Random Forest balanceado por la técnica de oversampling demostró ser el modelo con el mejor desempeño, con una puntuación de accuracy del 97%, precisión del 97%, Recall del 98% y F1 Score del 97%. Recomendamos que la empresa utilice este modelo para anticipar la tasa de abandono de los empleados. Además, se demostró que el ingreso, la edad, la distancia a casa, los años totales trabajados y años trabajados en la compañía son las principales características de porque los empleados deciden abandonar la compañía. En consecuencia, es esencial desarrollar estrategias destinadas a prevenir la salida de los empleados de la compañía.

REFERENCIAS

Varghese, D. (2018). Comparative Study on Classic Machine Learning Algorithms. Towards Data Science. https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222