Análisis de Deserción de Laboral (Proyecto de aula: Deep Learning y Series de tiempo)

Claudia Marcela Caro Cortés
Universidad Sergio Arboleda
Bogotá, Colombia
claudia.caro01@usa.edu.co

***Abstract*—The abstract describes a summary of the paper. In some cases, the authors confuse the abstract with the introduction. In the abstract using of acronyms and abbreviations is undesirable.**

I. Introducción

El problema que se desea abordar es descubrir los factores que conducen a la deserción de los empleados utilizando el conjunto de datos ficticio creado por científicos de datos de IBM llamado "IBM HR Analytics Employee Attrition & Performance". [1]

A. Trabajos relevantes

1) El trabajo de Janio Martinez Bachmann con su publicación en kaggle del dataset "Attrition in an Organization - Why Workers Quit?" [2], aborda la problemática principal de la deserción de los empleados creando preguntas sobre el análisis de género, análisis por género y educación, el impacto de los ingresos en la deserción y el ambiente de trabajo. De acuerdo con el análisis realizado utilizando XGBOOST indica como razones principales de la deserción: la falta de horas extras (82% de importancia), el ingreso mensual (100% de importancia) y la edad (60% de importancia). Este trabajo es relevante ya que se encuentra con el mejor posicionamiento de entre los notebook de kaggle que trabajaron el conjunto de datos de estudio.
2) Por su parte Bhartiya [3] utiliza el data set en su publicación en la revista IEEE Xplore del artículo "Employee Attrition Prediction Using Classification Models", donde pretende detectar las causas clave de la deserción y cómo minimizarlos aplicando los modelos de machine learning: Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest y Naive Bayes y las predicciones fueron evaluadas utilizando tres métricas de rendimiento: Accuracy, Matriz de Confusión y la Curva ROC donde obtuvo los mejores resultados con Random Forest con un 83.3% de precisión. En cuanto a los análisis realizados de la deserción con respecto al campo de la educación obtiene que la mayor deserción es del área de recursos humanos con un 25,9%, en la deserción con respecto al género indica que no tiene una influencia significativa en la tasa de deserción y la deserción respecto a la tasa de rendimiento donde encuentra que la mayor deserción se presenta en los empleados con mayor rendimiento con un 25,9% y deserción 0% para los empleados con el peor rendimiento.
3) Abordando el tratamiento de los datos desbalanceados para el conjunto de datos de estudio, [4] utiliza técnicas de sobremuestreo como la técnica de sobremuestreo de minorías sintéticas (SMOTE) y el enfoque de muestreo sintético adaptativo (ADASYN) para mejorar los resultados de clasificación por Regresión Logística, Support Vector Machine (SVM), Random Forest y XGBoost donde se presenta una mejora representativa utilizando SMOTE con una mejora de aproximadamente del 15% en las métricas de Accuracy, Precision, Recall, F1-score y AUC.
4) El uso de técnicas de clasificación binaria para predecir si un empleado en particular deja una empresa o no, lo aborda [5] junto con regresión logística y evalúa el modelo con las métricas precisión, matriz de confusión, curva ROC obteniendo una precisión del 85%.

II. Método

El método de análisis de datos para este proyecto de machine learning es el modelo CRISP-DM el cual se desarrolla a continuación.

A. Comprensión del negocio

La deserción laboral es una problemática que cada vez más se ve en Colombia por diferentes factores que abarcan entre la relación del empleado con su entorno laboral como el escape de talentos hacia otras empresas donde ofrezcan mejores condiciones laborales y económicas.

B. Comprensión de los datos

Dentro del set de datos dispuesto por IBM, se encuentran diferentes variables que son importantes al evaluar la correlación con la deserción laboral.

Entre las variables con mayor correlación con la deserción laboral se encuentra "Overtime", el trabajo extra es la que tiene mayor relación con el desempleo, sin embargo a partir de esta se encuentran otras variables que permiten complementan el análisis del estudio.

En la sección "Visualización y análisis de los datos", se detallan los hallazgos.

C. Preparación de los datos

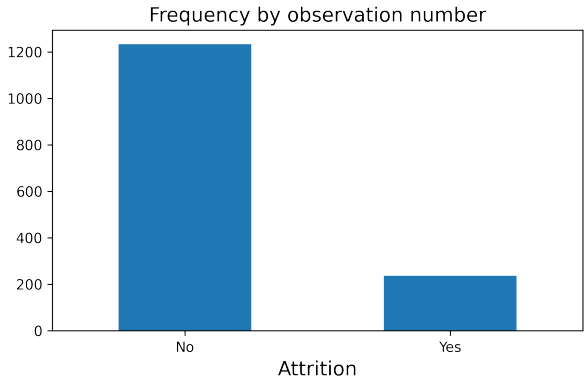
Los datos se prepararon de acuerdo con las características de los datos de cada variable con el uso de la instrucción de python employee.info(), donde se obtiene la siguiente información:

Fig. 1. tipo de datos

#	Column	Non-Null Count	Dtype
0	Age	1470 non-null	int64
1	Attrition	1470 non-null	object
2	BusinessTravel	1470 non-null	object
3	DailyRate	1470 non-null	int64
4	Department	1470 non-null	object
5	DistanceFromHome	1470 non-null	int64
6	Education	1470 non-null	int64
7	EducationField	1470 non-null	object
8	EmployeeCount	1470 non-null	int64
9	EmployeeNumber	1470 non-null	int64
10	EnvironmentSatisfaction	1470 non-null	int64
11	Gender	1470 non-null	object
12	HourlyRate	1470 non-null	int64
13	JobInvolvement	1470 non-null	int64
14	JobLevel	1470 non-null	int64
15	JobRole	1470 non-null	object
16	JobSatisfaction	1470 non-null	int64
17	MaritalStatus	1470 non-null	object
18	MonthlyIncome	1470 non-null	int64
19	MonthlyRate	1470 non-null	int64
20	NumCompaniesWorked	1470 non-null	int64

Luego se realiza la revisión del desbalanceo de la variable objetivo:

Fig. 2. balanceo

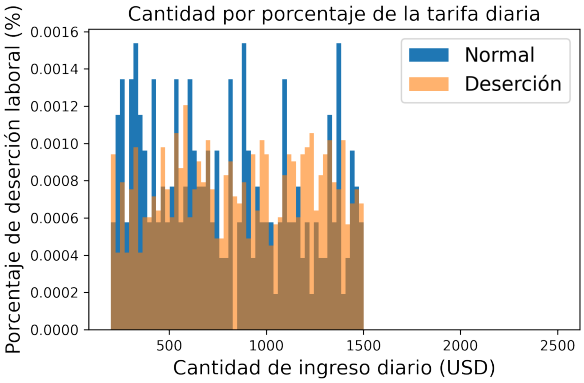


En el set de datos se observa muy pocos datos donde se registre la deserción laboral, con un porcentaje del 80.77% de datos marcados como Attribution = No.

Lo siguiente son pruebas de relacionamiento directo entre la variable objetivo Attribution y las características del dataset, en esta primera exploración se grafica toda variable que posee una relación directa, aunque con porcentaje muy bajo:

Característica DailyRate

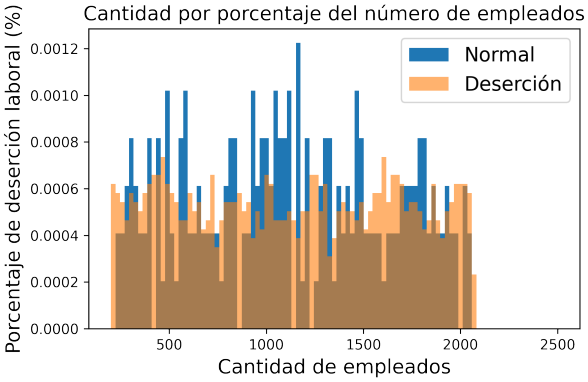
Fig. 3. Dailyrate vs Attrion



No se observa una tendencia característica entre el aumento del ingreso laboral y la deserción laboral.

Característica EmployeeNumber

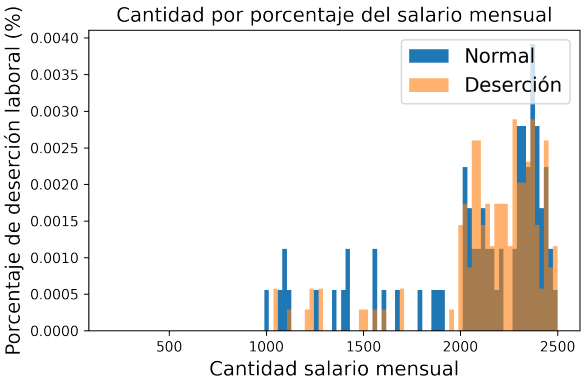
Fig. 4. EmployeeNumber vs Attrion



No se observa una tendencia característica entre el aumento del número de empleados y la deserción laboral.

Característica MonthlyIncome

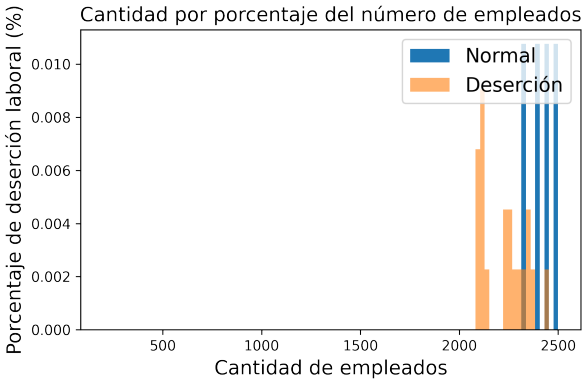
Fig. 5. MonthlyIncome vs Attrion



Se observa una tendencia entre el aumento del ingreso mensual con el nivel de deserción laboral.

Característica MonthlyRate

Fig. 6. MonthlyRate vs Attrion



Se observa una pequeña tendencia entre el aumento de la tasa mensual con el nivel de deserción laboral, pero no se considera representativa.

Luego, se realiza las graficas de proporción de datos para cada una de las variables categóricas, las cuales se pueden observar directamente en el notebook del proyecto.

Se retiran las variables que no tienen variación en los datos a lo largo del dataset:

Fig. 7. Variables eliminadas

```
employee = employee.drop(columns=['Over18'])
employee = employee.drop(columns=['StandardHours'])
employee = employee.drop(columns=['EmployeeCount'])
employee = employee.drop(columns=['EmployeeNumber'])
```

En las anteriores, solo existía una sola clasificación de la variable, por lo que no influye en los resultados del análisis.

Para obtener las variables más representativas con mejor correlación con la variable objetivo, se recurre a los árboles de decisión, para esto se transformaron las variables categóricas en numéricas y se pasaron como parámetros al clasificador:

Fig. 8. Código árbol de decisión

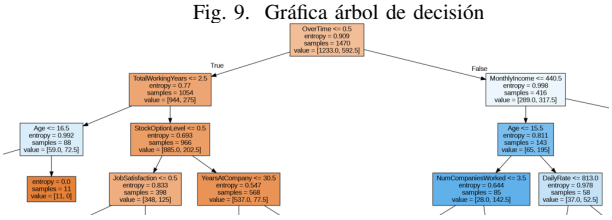
```
import pydotplus
import matplotlib.image as mpimg
from sklearn import tree
from subprocess import check_call
from IPython.display import Image as PImage
%matplotlib inline

# Crear Arbol de decision con profundidad = 5
decision_tree = tree.DecisionTreeClassifier(criterion='entropy',
                                            min_samples_split=20,
                                            min_samples_leaf=5,
                                            max_depth = 5,
                                            class_weight={1:2.5})

decision_tree.fit(X, y)
```

El resultado del árbol es el siguiente:

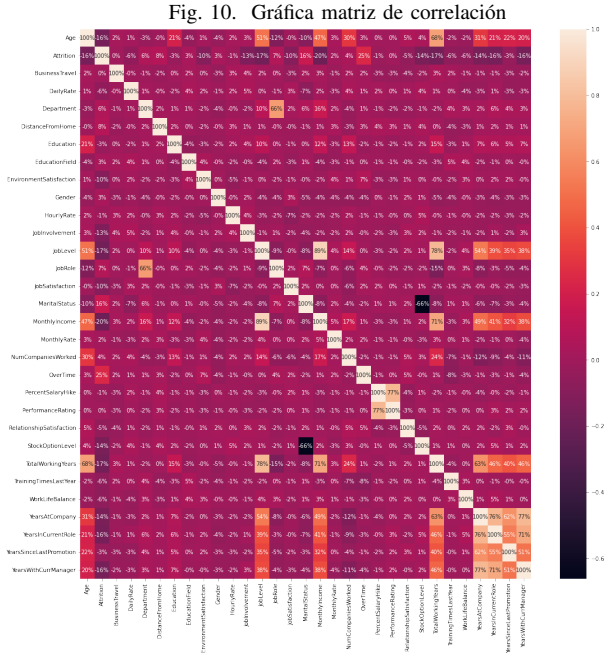
Fig. 9. Gráfica árbol de decisión



En la imagen, se muestra la raiz del árbol de decisión, donde se aprecia la variable que tiene mejor correlación con la deserción laboral.

Este misma exploración se realizó utilizando la matriz de correlación:

Fig. 10. Gráfica matriz de correlación



En esta se obtienen los mismos resultados que en el árbol de decisión, por lo que se validan los hallazgos de la

exploración de los datos.

D. Modelado

Se aborda en la entrega número 2.

E. Evaluación

Se aborda en la entrega número 3.

III. Visualización y análisis de los datos

aaa

IV. Conclusiones

aaa

V. Repositorio

https://github.com/claudiamarcelacaro/Maestria_IA/tree/main/Disenio_software_inteligente

References

[1] I. data scientists, "Ibm hr analytics employee attrition performance — kaggle.com," https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv, 2017.

[2] J. M. Bachmann, "Attrition in an organization —— why workers quit? — kaggle.com," https://www.kaggle.com/code/janiobachmann/attrition-in-an-organization-why-workers-quit, 2019.

[3] N. Bhartiya, S. Jannu, P. Shukla, and R. Chapaneri, "Employee attrition prediction using classification models," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 2019, pp. 1–6.

[4] R. A. Danquah, "Handling imbalanced data: A case study for binary class problems," 2020. [Online]. Available: https://arxiv.org/abs/2010.04326

[5] S. Ponnuru, G. Merugumala, S. Padigala, R. Vanga, and B. Kantapalli, "Employee attrition prediction using logistic regression," *Int. J. Res. Appl. Sci. Eng. Technol*, vol. 8, no. 5, pp. 2871–2875, 2020.