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*

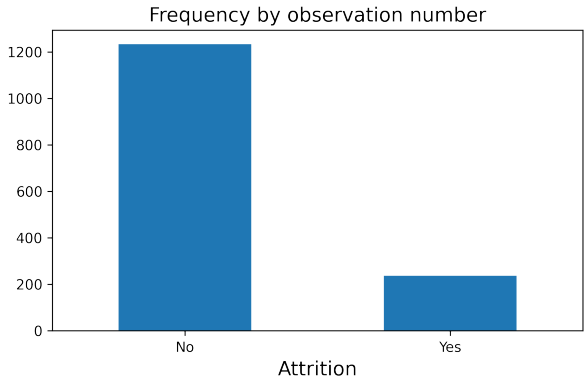
Los datos se prepararon de acuerdo con las características de los datos de cada variable con el uso de la instrucción de python employee.info(), donde se obtiene la siguiente información:

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

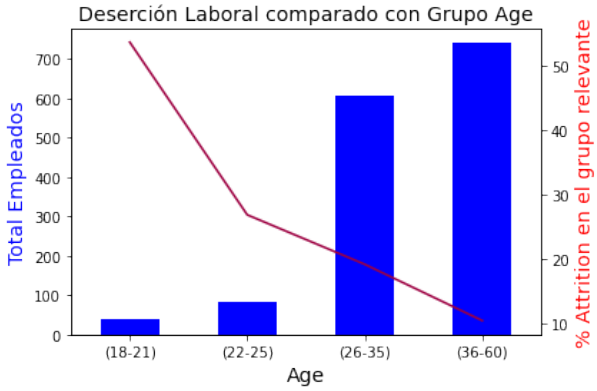


En el set de datos se observa muy pocos datos donde se registre la deserción laboral, con un porcentaje del 80.77% de datos marcados como Attribution = No.

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 Age

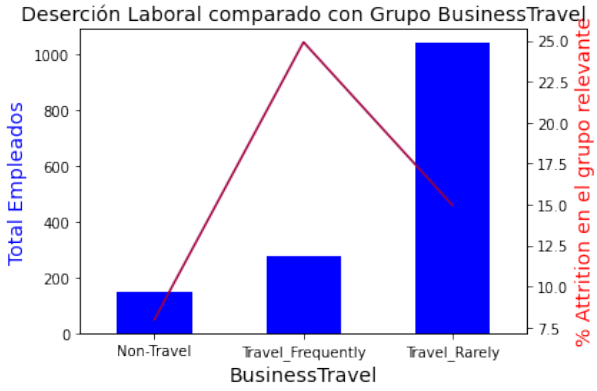
Fig. 3. Attrition Age Percent



El mayor porcentaje de deserción laboral se presenta en la población de 18-21 años con el 53.66%, la deserción disminuye con el paso de los años.

Característica BusinessTravel

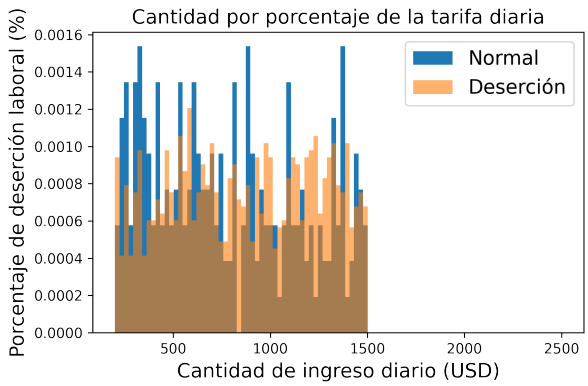
Fig. 4. Attrition BusinessTravel Percent



La mayor deserción se da en el grupo de trabajadores Viajeros Frecuentes con el 24.90%

Característica DailyRate

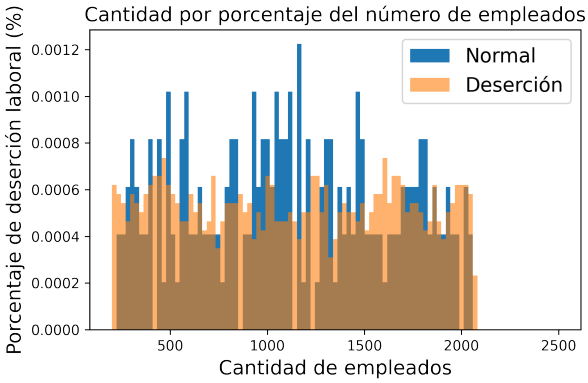
Fig. 5. Dailyrate vs Attrion



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

Característica EmployeeNumber

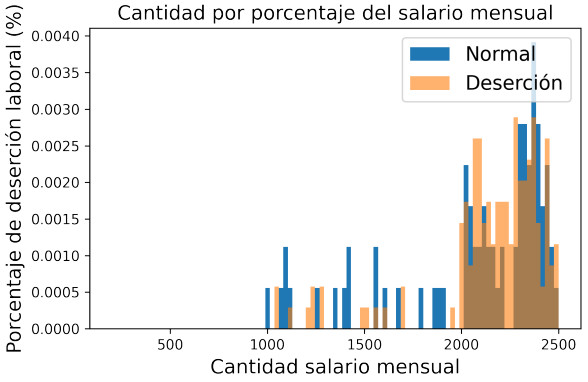
Fig. 6. EmployeeNumber vs Attrion



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

Característica MonthlyIncome

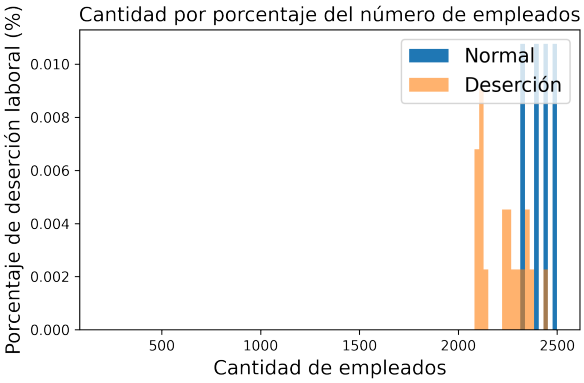
Fig. 7. MonthlyIncome vs Attrion



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

Característica MonthlyRate

Fig. 8. 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 directamete en el notebook del proyecto.

El siguiente paso es la transformación de las variables categóricas y numéricas en la fase exploratoria con las cuales se construye el arbol de decisión para obtener las variables más representativas con mejor correlación con la variable objetivo:

Fig. 9. Código árbol de decisión

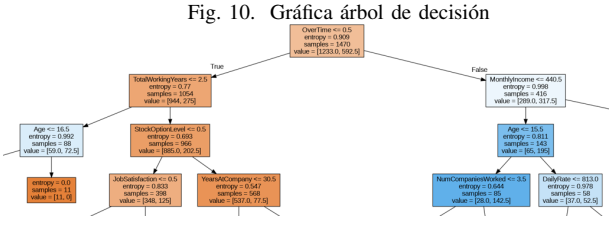
```
import pydotplus
import matplotlib.image as mpimg
from sklearn import tree
from subprocess import check_call
from IPython.display import Image as PImage
%matplotlib inline

# 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. 10. Gráfica árbol de decisión

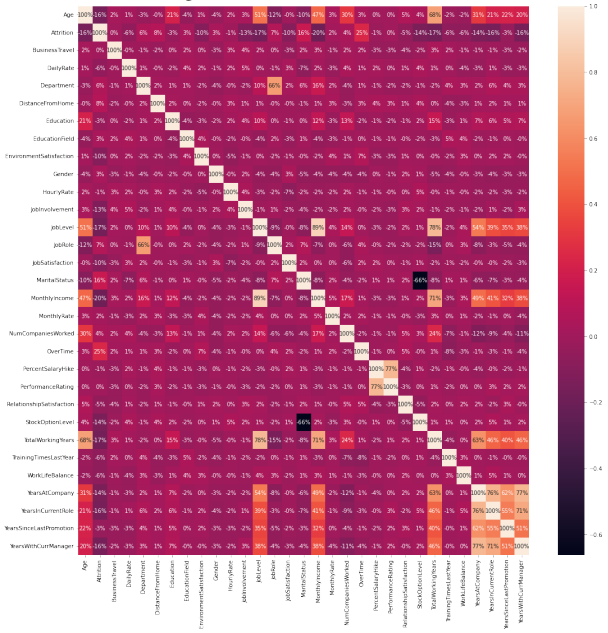


En la imagen, se muestra la raiz del árbol de decisión, donde se aprecia la variable que tiene mejor correlación con

la deserción laboral.

Este misma exploración se realizó utilizando la matriz de correlación:

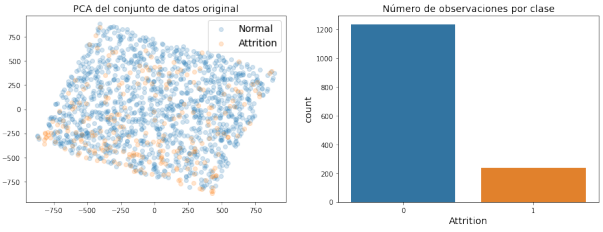
Fig. 11. Gráfica matriz de correlación



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.

Con PCA (análisis para componentes principales) se fusiona las características para la exploración de datos, este es un metodo no supervisado.

Fig. 12. Gráfica de observaciones por clase



Se puede visualizar la representación de la distribución de las característica y el desbalanceo de la variable objetivo ”Attrition” o deserción laboral.

En la fase de Preprocesamiento se construyen las funciones de one_hot_encoder, simple_imputer y preprocesamiento, y se ejecuta con una sola instrucción el preprocesamiento de los datos.

Apartir del proceso posterior de correlación se eliminan del modelo de datos a procesar las variables: JobRole_Sales Executive, Department_Human Resources, MonthlyRate, PerformanceRating, JobRole_Research_Scientist, HourlyRate, PercentSalaryHike, EducationField_Other, las cuales no aportaban relación mayor a 0.02% (sea en valores positivos o negativos).

Por último para finalizar el preprocesamiento, se crea el set de entrenamiento y pruebas, con un porcentaje 80 a 20.

D. Modelado

Se selecciona los modelos de Regresión logística para los datos sin balancear y se evaluaron los modelos. Se utiliza Grid search para ajustar los hiperparámetros.

Se seleccionan otros modelos para realizar el entrenamiento como: RandomForest, K-Means.

E. Evaluación

Se evalua los modelos de balanceo y se comparan con respecto al modelo de Regresión Logística.

Se evalúa las reducciones de dimensionalidad para determinar si pueden ser aplicadas al set de datos.

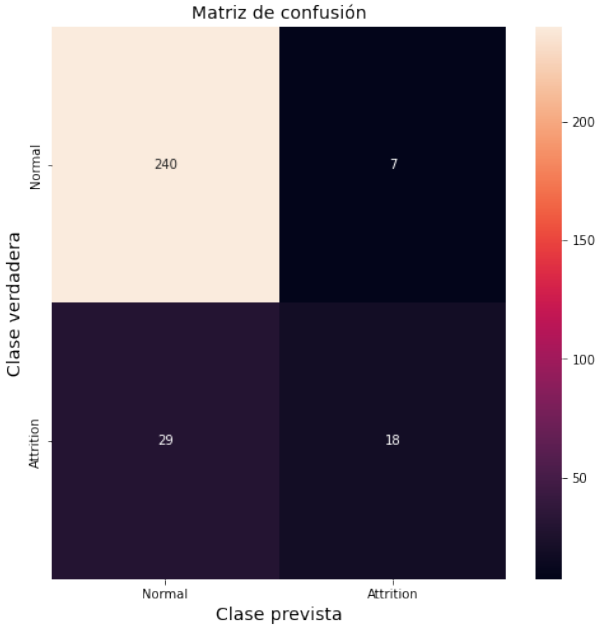
III. Visualización y análisis de los datos

A. Regresión logística

La regresión logística es un tipo de algoritmo de aprendizaje supervisado que se utiliza comúnmente para la clasificación de datos.

Se entrena los datos sin balancear para validar el nivel de precisión del modelo.

Fig. 13. Gráfica matriz de confusón con regresión logística



Se presenta la matriz de confusión, y en particular se enfoca en la clase 2 (Attrition), que es la que se quiere detectar.
Se observan 29 falsos negativos y 18 verdaderos positivos, lo que resulta en un recall de 0.38, un valor que se desea mejorar. Es importante destacar que el modelo no es capaz de detectar correctamente los casos de fraude.

Los resultados que se obtienen son los siguientes:

Fig. 14. Gráfica resultados con regresión logística

	precision	recall	f1-score	support
0	0.892193	0.971660	0.930233	247.000000
1	0.720000	0.382979	0.500000	47.000000
accuracy	0.877551	0.877551	0.877551	0.877551
macro avg	0.806097	0.677319	0.715116	294.000000
weighted avg	0.864666	0.877551	0.861454	294.000000

B. Validación del balanceo de clases

Luego de aplicar las estrátegias de balaneo de clases: Penalización, NearMiss Subsampling, Random Oversampling, Smote Tomekm y Ensemble se obtienen los siguientes resultados:

Fig. 15. Gráfica resultados estrategias de balanceo

	algorithm	precision	recall	overall
1	Penalización	0.93	0.68	0.805
2	NearMiss Subsampling	0.92	0.62	0.770
3	Random Oversampling	0.92	0.60	0.760
5	Ensemble	0.90	0.51	0.705
4	Smote Tomek	0.90	0.49	0.695
0	Regresion Logística	0.89	0.38	0.635

La estrategia con la que se obtiene mejor resultado es la Penalización con una precision del 0.93 y un recall del 0.68, lo que indica que es capaz de clasificar correctamente tanto las instancias positivas como las negativas de la clase minoritaria.
El modelo 2 obtuvo una precision parecida sin embargo fué mucho menor el recall con respecto al modelo 1. Los modelos 3, 4 y 5 mantuvieron una presición del 0.9 sin embargo el recall iba disminuyendo, sin embargo, es importante destacar que todas las técnicas aplicadas logran mejorar el modelo inicial de Regresión Logística, que solo alcanzaba un 0.38 de recall para la clase de Attrition (deserción laboral). Cabe recordar que el conjunto de datos presenta un desbalanceo considerable entre las clases.

C. Grid search

La búsqueda en cuadrícula o Grid search es un método para realizar la optimización de hiperparámetros que permite encontrar la combinación óptima.

Los mejores hiperparámetros que se obtienen con los datos de entrenamiento son:

'C': 2, 'gamma': 0.1767766952966369

Mejor score: 0.8767039307609087

Obteniendo como mejor resultado con los datos de test:

0.8767039307609087

D. RandomForest

Este modelo es fácil de interpretar, los árboles pueden ser visualizados. El costo computacional del uso del árbol para predecir la categoría de un ejemplo es mínimo comparado con otras técnicas.

Luego del entrenamiento se obtuvieron como mejores parámetros de entrenamiento para el árbol de desciones:

max_features: 0.699752894357787
n_estimators: 669

Con estos parámetros se obtuvo un accuracy de 0.8571438874864767 para los datos de entrenamiento y 0.8707482993197279 para los datos de test.

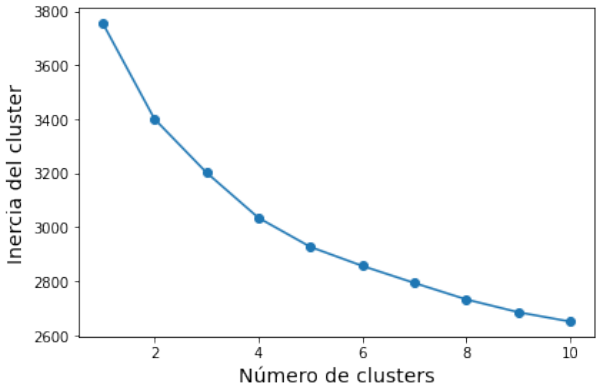
De igual forma, bajo este método se obtienen las mejores característica que tiene mayor relación con la deserción laboral:

Importancia de características:
Característica MonthlyIncome
Característica Age
Característica MonthlyRate
Característica TotalWorkingYears
Característica OverTime
Característica DistanceFromHom
Característica HourlyRate
Característica YearsAtCompany
Característica PercentSalaryHike
Característica NumCompaniesWorked
Característica YearsWithCurrManage
Característica EnvironmentSatisfaction

E. K-Means

Método de aprendizaje no supervisado como algoritmo de agrupamiento.

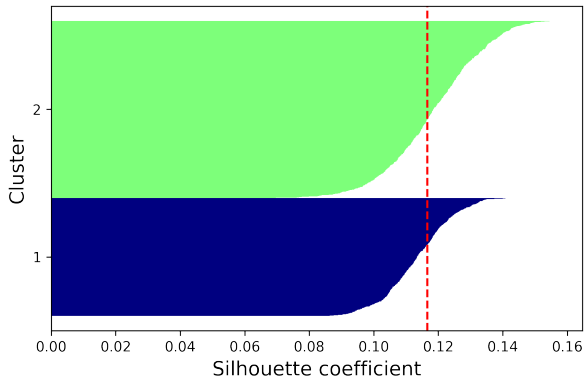
Fig. 16. Gráfica inercia vs número de clústeres



De la gráfica anterior se puede apreciar que para k=2 hay una buena opción de agrupamiento de datos

Al entrenar el modelo con un k=2, se obtiene el siguiente coeficiente de silueta:

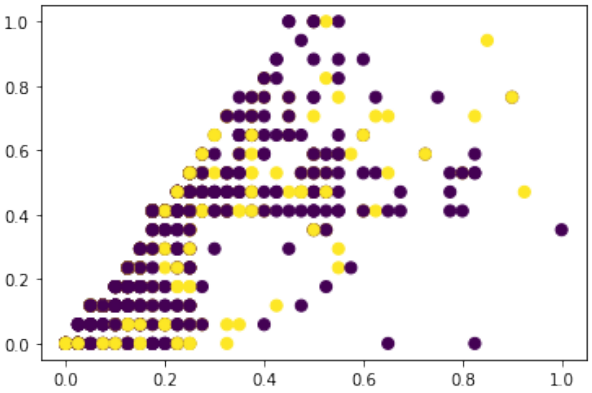
Fig. 17. Gráfica inercia vs número de clústeres



Resultado: 651 de 1470 muestras se etiquetaron correctamente. Accuracy score: 44.29 %

Al aplicar el método SpectralClustering se obtiene la siguiente gráfica;

Fig. 18. Gráfica KMeans ScatterSpectralClustering

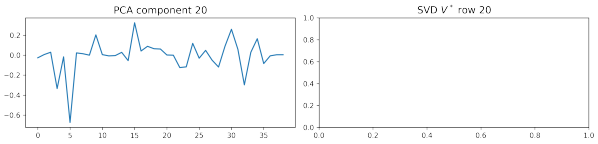


Resultado: 964 de 1470 muestras se etiquetaron correctamente. Accuracy score: 65.58 %

F. Reducción de dimensionalidad

Se evaluen los métodoes de reducción: Análisis de Componentes Principales (PCA) que es una técnica de reducción de dimensionalidad lineal y el método de Singular Value Decomposition (SVD).

Fig. 19. PCA vs SVD



Por PCA se obtienen los siguientes resultados:

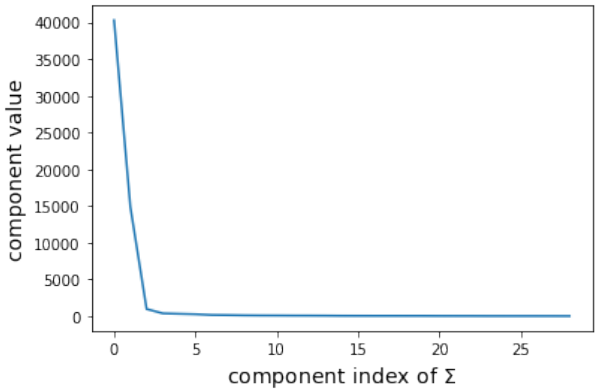
0.6557823129251702
0.8428571428571429

Por SVD se obtienen los siguientes resultados:

Keeping 7 components

Por análisis de componentes se obtiene:

Fig. 20. Gráfica de análsisi de componentes

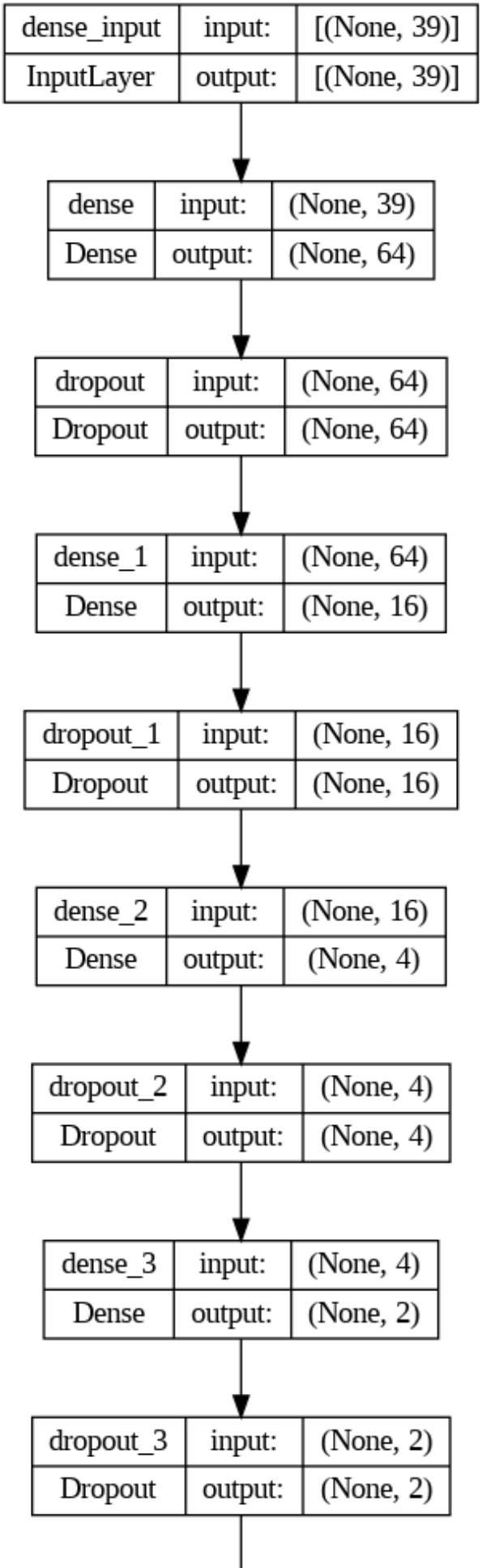


Donde se obtine 1 solo componente como resultado.

G. Redes Neuronales NNs

Se crea una red neuronal de clasificación con las siguientes capas:

Fig. 21. Gráfica modelos de la red neuronal



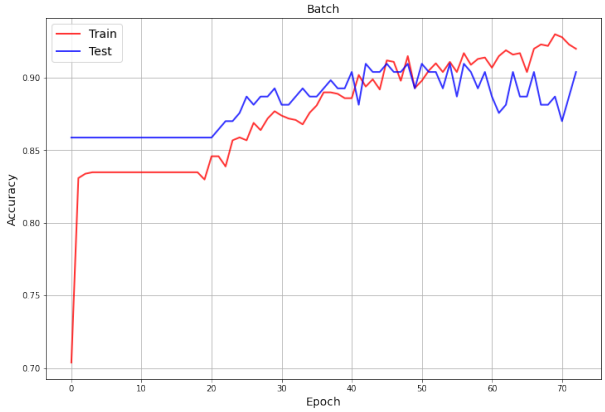
Al utilizar tf.keras.callbacks.EarlyStopping se logró evaluar el mejor momento de para del entrenamiento, una vez se detiene ajusta los pesos del modelo:

108/125 [=========================¿.....] - ETA: 0s - loss: 0.2213 - accuracy: 0.9248Restoring model weights from the end of the best epoch: 43.

125/125 [==============================] - 0s 4ms/step - loss: 0.2376 - accuracy: 0.9199 - val_loss: 0.3265 - val_accuracy: 0.9040 Epoch 73: early stopping

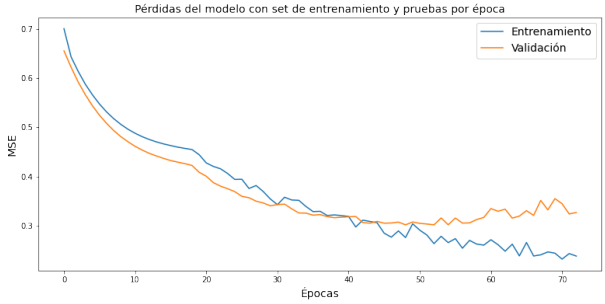
En la gráfica del Accuracy:

Fig. 22. Gráfica de accuracy en los Bacth



En la gráfica de la Pérdida:

Fig. 23. Gráfica de pérdida en los Bacth

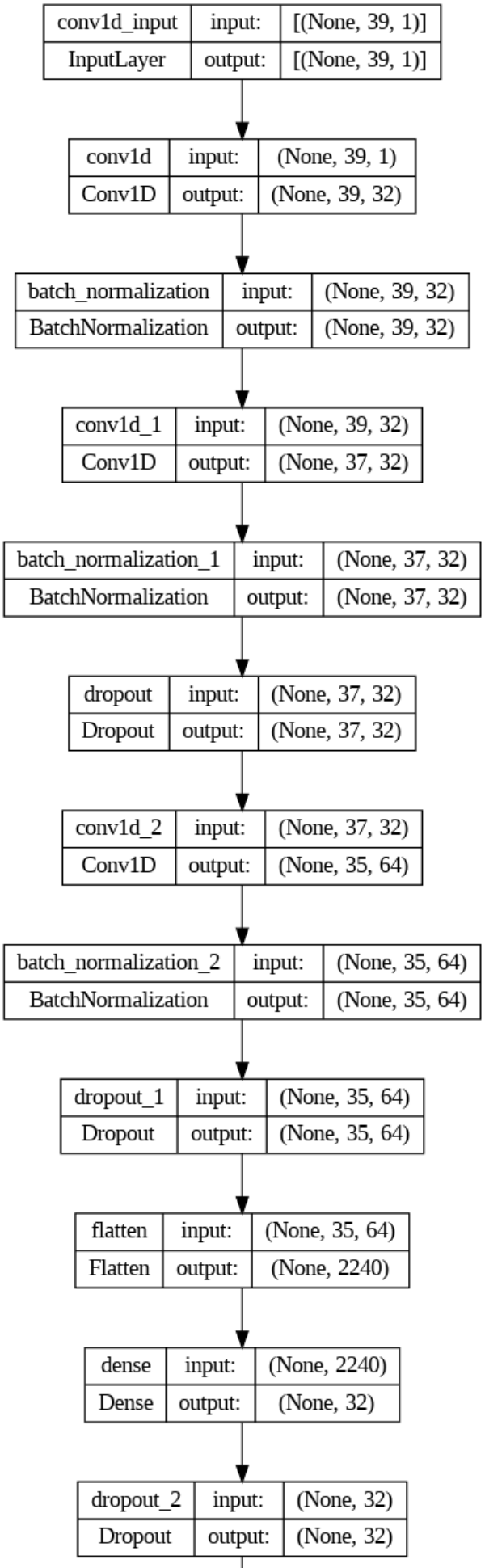


Con una precisión en los datos de prueba: 0.8775510191917419

H. Redes Neuronales Convolucionales CNNs

Se crea una red neuronal convolucional que posee como entrada capas convolucionales para obtener la mayor parte de caracteríticas de aprendizaje y hacia la salida capas de clasificación, la red convolucional utilizada para el dataset de Decersión Laboral es la Convolucional 1D (Convolution1D):

Fig. 24. Gráfica modelos de la red neuronal convolucional



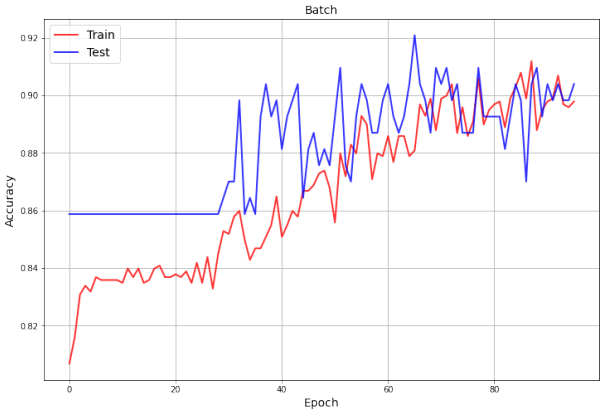
Al utilizar tf.keras.callbacks.EarlyStopping se logró evaluar el mejor momento de para del entrenamiento, una vez se detiene ajusta los pesos del modelo:

62/63 [==========================¿.] - ETA: 0s - loss: 0.2615 - accuracy: 0.8972Restoring model weights from the end of the best epoch: 66.

63/63 [==============================] - 1s 13ms/step - loss: 0.2609 - accuracy: 0.8979 - val_loss: 0.3761 - val_accuracy: 0.9040 Epoch 96: early stopping

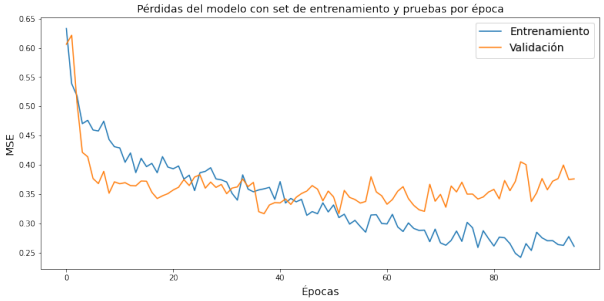
En la gráfica del Accuracy:

Fig. 25. Gráfica de accuracy en los Bacth CNN



En la gráfica de la Pérdida:

Fig. 26. Gráfica de pérdida en los Bacth CNN



Con una precisión en los datos de prueba: 0.8809523582458496

IV. Resultados

Al comparar todos los modelos de entrenamiento, se obtiene los siguientes resultados:

Fig. 27. Tabla de resultados

	algorithm	accuracy
7	CNN	88.095236
0	Regresion Logística	87.755102
6	Redes Neuronales	87.755102
1	Grid Seach	87.414966
2	RandomForest	87.074830
5	PCA	84.285714
4	SpectralClustering	65.578231
3	K-Means	44.285714

Como se puede apreciar, se obtuvo como mejor modelo de entrenamiento las Redes Neuronales Convolucionales (CNN) con un Accuracy del 88.095% superando el modelo de Regresión Logítica cuyo modelo fué el más implementado entre los autores citados en algunos casos con los mejores resultados. El modelo de redes neuronales de clasificación no fueron sufientes para superar la regresión logística pero lograron un Accurary muy similar. Es algo interesante que el método de Grid Search no haya logrado mejorar el modelo de Regresión Logística, es posible que sea por que se decidió trabajar con los datos sin balancear, ya que al balancearlos con el mejor médodo de balanceo la red neuronal no lograba entrenar, y los Accurary generales no superaban del 84%.

V. Conclusiones

El mayor porcentaje de deserción laboral se presenta en la población de 18-21 años con el 53.66%, la deserción disminuye con el paso de los años.

La mayor deserción se da en el grupo de trabajadores Viajeros Frecuentes con el 24.90%.

La mayor deserción laboral se presenta en los trabajadores que menos ganan a diario, los que ganan entre 1-500 dólares poseen una deserción del 19.25%. La deserción disminuye entre mayor sea su tarifa diaria.

El departamento de Investigación y Desarrollo (Research & Development) es el que posee menor tasa de deserción del 13.84% y el de mayor deserción el departamento de Ventas (Sales) con el 20.62%.

Poseen una deserción laboral más alta del 20.95% las personas que viven a más de 10 millas del lugar de trabajo. La deserción aumenta con la distancia entre la casa y el trabajo.

Los empleados que cursan niveles inferiores de escolaridad son los que tienen mayor Deserción Laboral con el Bachelor del 17.30% y Bellow College con el 18.23% de deserción laboral, entre ambos representan el 35.53% de la deserción de la empresa, siendo muy representativo.

Los empleados que cursan niveles inferiores de escolaridad son los que tienen mayor Deserción Laboral con el Bachelor del 17.30% y Bellow College con el 18.23% de deserción laboral, entre ambos representan el 35.53% de la deserción de la empresa, siendo muy representativo.

Con una satisfacción con el ambiente laboral baja se obtiene el mayor porcentaje de deserción laboral, del 25,35%

Se tiene una mayor deserción laboral de los hombres respecto a las mujeres con el 17%.

Dentro del grupo de JobInvolvement la característica con mayor relevancia en porcentaje de nivel de deserción laboral fue el nivel bajo con un 33.73%, sin embargo, la cantidad de empleados no es representativo. La calificación Hight es la más representativa a nivel compañía con un 14.4% de deserción pero que representa el 52.74% de deserción a nivel empresa.

La deserción disminuye con el aumento del nivel de trabajo. Se tiene el mayor porcentaje de deserción en el Nivel-1 con 26.33%. Los Técnicos de Laboratorio poseen el mayor nivel de deserción a nivel de la compañía con el 26.16% seguido del Ejecutivo de Ventas con un 24.05%.

Es sorpresivo que el mayor porcentaje de Deserción Laboral se presente con un nivel alto (High) de satisfacción laboral con un 30.80% a nivel de toda la compañía.
Los empleados Solteros son los que tienen mayor porcentaje de deserción laboral a nivel de la empresa con un 50.63%.

La mayor deserción laboral se da en los rangos salariales de 2501-3000 dólares con un 40% a nivel de toda la empresa.

Se observa que el grupo con mayor deserción laboral es el de 20001-27000 con un 28.29% a nivel de la empresa.

A nivel de la empresa, quienes tienen una sola experiencia anterior son los que tienen mayor nivel de deserción laboral con un 41.35% a nivel de la empresa.

OverTime o tiempo extra de trabajo del horario laboral tiene un porcentaje alto de deserción laboral con el 53.58% a

nivel de la empresa. Es muy representativa.

La mayor deserción laboral se da en los aumentos salariales más bajos, 32% a nivel de la compañía para los aumentos entre 13-15 dólares y del 31.22% a nivel de la compañía a los aumentos entre 10-12 dólares.

A nivel de la compañía las persona con un rendimiento Excelente son las que tienen mayor tasa de deserción laboral con un 84.38% a nivel de la empresa, lo cual por la cantidad de empleados en esa clasificación es muy representativa.

Es sorpresivo que a nivel de la empresa los empleados que indicaron un nivel de satisfacción Alto (High) sean los que tiene la mayor tasa de deserción laboral del 29.95% a nivel de la empresa.

Los trabajadores sin acciones son los que tiene mayor nivel de deserción laboral con el 64.97% a nivel de la empresa.

Se observa una disminución de la deserción laboral a medida que se tiene mayor tiempo de trabajo en la misma empresa. Sin embargo, a nivel de la empresa, se tiene la mayor deserción con una experiencia de 6-10 años con un 38.39%.

Los empleados que tuvieron 2 y 3 entrenamientos el año anterior tienen el mayor número de deserciones las cuales componen el 70.5% de todas las deserciones en la empresa.

A nivel de la clasificación se tiene la mayor deserción con un nivel malo de equilibrio con un 31.25%, sin embargo a nivel de la empresa se tiene mayor deserción las personas que indicaron tener un equilibrio mejor (Better) con un 53.58%.

La mayor deserción laboral se encuentra en los empleados del 2-5 años con un 36.70% a nivel de toda la compañía. En el grupo de clasificación la mayor deserción es con 1 año de experiencia con el 34.5%.

La mayor deserción laboral ocurre en persona que han durado en el mismo cargo entre 2-4 años con un 41.77% a nivel de la empresa.

Las personas recién egresadas con las que tienen mayor porcentaje de deserción laboral con el 20.67%. a nivel de la compañía.

La mayor deserción laboral de produjo cuando trabajaban con el mismo jefe durante 1 año con el 21.09% a nivel de la empresa.

VI. 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.