Modelo de predicción de la deserción de empleados en una empresa. (Junio de 2023)

Shirley Viviana Jiménez Osorio

Resumen – En este documento se hará una presentación de diversos experimentos ejecutados para predecir la deserción de un empleado en una compañía, siendo la variable de salida, 1 deserta, 0 permanece. Se encontrarán diversos trabajos citados al rededor del tema, así mismo varios modelos de clasificación con sus respectivas métricas.

Índice de Términos – Modelo, Machine Learning, clasificación, Attrition

I. INTRODUCCIÓN

En este documento se realizará el abordaje de un problema de clasificación en aprendizaje automático, específicamente se abordará el tema de la deserción de empleados en una compañía, identificando aquellos que se retiran y permanecen (1-0). Analizando diversas variables de entrada, tales como experiencia en años, salario, genero, ascensos, nivel de educación entre otras.

Los experimentos presentados se harán a través de modelos de clasificación tales como, Máquina de vectores de soporte, modelo de regresión lineal, Algoritmo Knn, modelo Bayesiano, Árboles de decisión, Random forest, y Neural Networks. De tal forma que podamos elegir el que mejor nos de resultados.

Todo esto con un la presentación de un tratamiento de datos previo, que permita obtener las mejores características, balancear las variables de salida y una optima limpieza de datos. Lo que permitirá que los experimentos ejecutados nos entreguen métricas de confianza que permitan la evaluación del aprendizaje y su éxito.

II. DESCRIPCIÓN DEL PROBLEMA DE PREDICCIÓN

A. Planteamiento del problema

La retención y permanencia del talento humano es uno de los principales desafíos que enfrentan las empresas en la actualidad; este se relaciona con aspectos fundamentales para el desarrollo adecuado de sus negocios y también como una arista importante en el ahorro de recursos financieros para la contratación. Es de esperarse que un empleado con alta permanencia, entienda y resuelva de forma más eficiente los desafíos y los problemas a los que se vea enfrentado, el conocimiento del negocio le agrega experiencia a la hora de desarrollar su trabajo de forma autónoma, así como en asuntos que requieran liderazgo. Por esta razón se ha vuelto relevante para las empresas determinar y predecir qué factores conllevan a la deserción de sus colaboradores, con el fin de tomar las medidas que permitan ampliar el tiempo de permanencia, esto por su puesto sin dejar de tener en cuenta que contratar un nuevo empleado implica incurrir en gastos de reclutamiento, exámenes médicos, capacitaciones, tiempos de ajustes, etc. Retener un empleado impacta también directamente sobre la carga que enfrentan áreas como gestión humana o formación.

Es por esta razón que se plantea el desarrollo de un modelo de Machine Learning, que permita predecir si un empleado podría darse de baja, teniendo en cuenta diferentes variables de su entorno personal y laboral, de esta manera podrían tomarse medidas preventivas que permitan retener al mismo.

B. Abordajes previos del problema

A través de cuatro modelos de Machine Learning Sierra Buriticá (2022) aborda el problema de deserción enfocado en en la industria de software en Colombia, con información de 1497 empleados y 19 variables, sus experimentos abordados a través de modelos como: (Niave Bayes, Random Forest, Decision Tree, Logistic Regression) le arrojan con mejores resultados y métricas un modelo de Decision Tree con 14 capas, cuyas métricas de evaluación superiores al resto de modelos fueron, su curva de aprendizaje y curva de ROC.

Rey Caldeyro (2021) plantea el ejercicio sobre este mismo tema realizando un análisis de predicción aplicado a la deserción de empleados. Con una base de datos publica de 1470 empleados determina que el mejor modelo para predecir la deserción es una regresión logística.

Reyes Huertas (2019), aborda el tema en cuestión a través de un modelo para predecir deserción con algoritmos genéticos y redes neuronales artificiales, como resultado de su ejercicio, obtiene que con un modelo de redes neuronales obtiene una precisión del 88,92% para identificar la deserción de un empleado.

Documento recibido el 20 de junio de 2023.

III. ENTRENAMIENTO Y EVALUACIÓN DE LOS MODELOS

A. Base de datos usada

Se usó una base de datos de Kaggle con originalmente 19.104 instancias, que luego de una preparación de datos, y eliminación de nulos que no podían ser imputados, termina con 18.064 datos y 13 columnas, cuatro variables cuantitativas y 9 cualitativas.

Se obtiene la variable 'retention_days' a través de la resta de diversas variables de fechas que se tenía en la base original, esta variable permite obtener los días de permanencia en la empresa que tiene un trabajador. Para balancear la base de datos se realiza el sobre muestreo para las variables de entrada y salida.

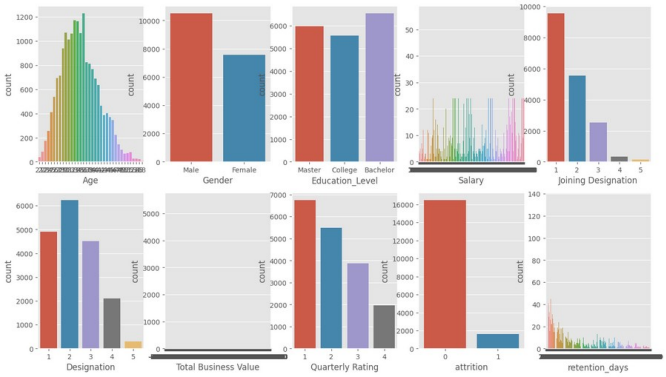


Fig. 1. Frecuencias por variable. Elaboración propia

La variable Join Designation resume el puesto en el que fue asignado la persona cuando ingresó, y Designation el rango con el que terminó, es decir que a través de esta calculamos la variable Promoted, es decir, si la persona alguna vez tuvo un ascenso.

B. Experimentos

El primer experimento consiste en correr siete modelos de Machine Learning, con los datos normalizados y usando para entrenamiento el 80% de los datos:Máquina de vectores de soporte, modelo de regresión lineal, Algoritmo Knn, modelo Bayesiano, Árboles de decisión, Random forest, y Neural Networks.

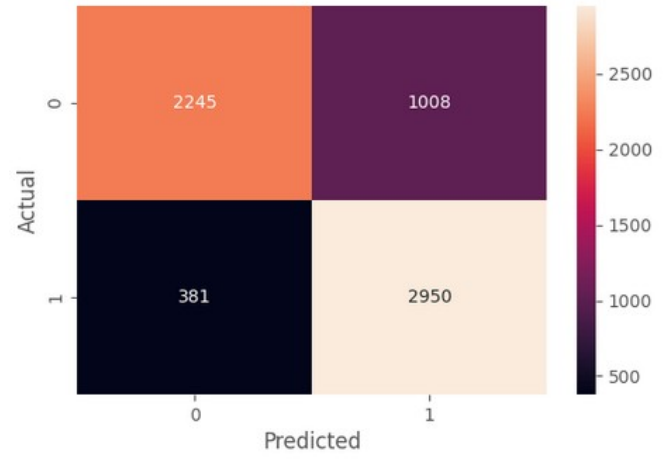
La exactitud arrojó como resultado el siguiente resumen:

```
===== Accuracy de los modelos =====

modelSVC       : 0.7867557715674363
modelLR        : 0.6591737545565006
modelknn       : 0.7890340218712029
modelComplNB   : 0.7506075334143378
modelTreeClas  : 0.5451093560145808
modelRanForest: 0.7091433778857837
ModelNN        : 0.5147326852976913
```

Fig.2.Resultados exactitud de los siete modelos, elaboración propia.

El modelo que mejor nos arroja exactitud es el algoritmo Knn, a este se le ha obtenido la siguiente matriz de confusión con los siguientes resultados:



	precision	recall	f1-score	support
0	0.85	0.69	0.76	3253
1	0.75	0.89	0.81	3331
accuracy			0.79	6584
macro avg	0.80	0.79	0.79	6584
weighted avg	0.80	0.79	0.79	6584

Fig. 3. Matriz confusión, entrenamiento con el 80% de los datos.

Este experimento arroja una precisión del 75% de los retiros y un 79% de exactitud.

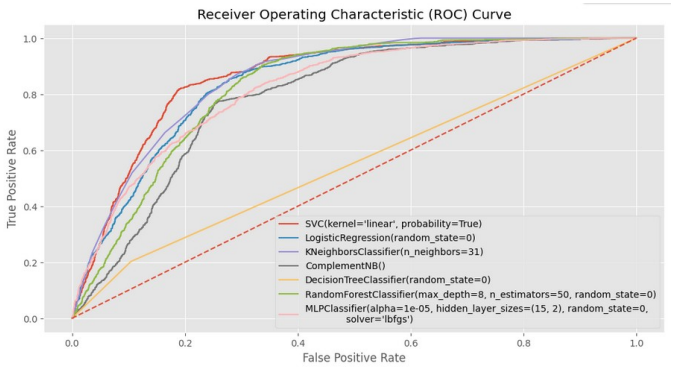


Fig. 4. Curva ROC, entrenamiento con el 80% de los datos.

A continuación se realizó un experimento con validación cruzada, es decir, utilizando todos los datos de la base, se ejecutaron 10 pliegues, y se evaluaron cuatro métricas, 'accuracy', 'precision', 'recall', 'f1'.

Estos fueron los resultados por modelos para el f1 fueron:

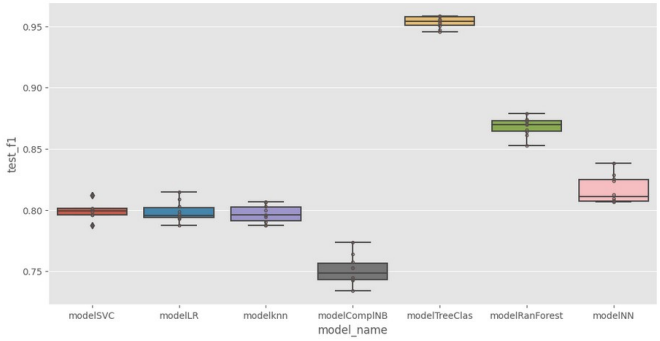


Fig. 4. f1-score de los modelos con validación cruzada

C. Resultados y conclusiones

Corriendo los siete modelos con el 80% de los datos para entrenamiento se obtiene como mejor modelo el algoritmo Knn Este experimento arroja una precisión del 75% de los retiros y un 79% de exactitud.

Sin embargo como esta metodología no utiliza el 100% de los datos, se decide probar con validación cruzada para observar el comportamiento de los modelos con todos los datos. Para este ejercicio se evalúan cuatro métricas, ‘accuracy', 'precision', 'recall', 'f1'. El modelo que mejor métricas obtuvo fue el árbol de decisión, con accuracy y f1 de 95%, precisión del 91% y recall de 1.

REFERENCES

[1] (Journal Online Sources style) Sierra Buriticá,(2022), ,”Análisis y predicción de la deserción de empleados. Un caso de estudio en la industria de software colombiana” *articulo de internet, Universidad Eafit. Disponible en: https://repository.eafit.edu.co/bitstream/handle/10784/32155/ElianaMarcela_SierraBuritic%C3%A1_2022.pdf?sequence=2&isAllowed=y*

[2] (Journal Online Sources style) Rey Caldeyro, (2021), ,” Análisis de predicción aplicado a la deserción de empleados.” *articulo de internet, Universidad Complutense Madrid.. Disponible en: https://eprints.ucm.es/id/eprint/68423/1/rey-caldeyro-mar%C3%ADa-ema-tfm.pdf*

[3] (Journal Online Sources style) Reyes Huertas (2019), ,”Predicción de deserción laboral utilizando algoritmos genéticos y redes neuronales artificiales ”*articulo de internet, revista Dialnet. Disponible en:https://dialnet.unirioja.es/servlet/articulo?codigo=7425952*

[4] (Journal Online Sources style) Base de datos tomada de Kaggle: https://www.kaggle.com/datasets/pavan9065/predicting-employee-attrition