Estrategias de aprendizaje supervisado para un problema de renuncia en una compañía.

Daniel Manco, Esteban Arcila, Astrid Daniela Giraldo.

Departamento de Ingeniería Industrial
Universidad de Antioquia
Medellín, Colombia

Introducción

La decisión de un trabajador de renunciar o abandonar su puesto en una empresa puede estar influenciada por una variedad de motivos. Entre ellos se incluyen un entorno laboral desfavorable, una remuneración insatisfactoria, limitadas oportunidades de crecimiento profesional o simplemente la falta de reconocimiento por su dedicación y esfuerzo.

Muchas empresas optan por diferentes estrategias para mitigar la renuncia de sus trabajadores y así tener una gestión en los recursos humanos ideal para cumplir con los objetivos de la organización. Uno de estas estrategias es hacer un estudio tanto cualitativo como cuantitativo acerca de todas las posibles causas que lleven al empleado a abandonar la organización

El objetivo de este estudio es analizar las posibles causas por las cuáles el 15% de los empleados en un año se retiran de la compañía por diferentes motivos conllevando implicaciones negativas a la empresa como;

- elevados costos de contratación
- los proyectos se ven afectados por retrasos , ocasionando insatisfacción del cliente y el usuario.
- sobrecargo laboral
- afectación en los procesos debido a que se pierde la experiencia y conocimiento.

Conociendo estas afectaciones negativas, la organización quiere tomar acciones necesarias para reducir el porcentaje de retiros y que estos no superen el 12%.

Diseño de la solución.

Teniendo en cuenta la problemática, información disponible y el proceso analítico y administrativo a realizar, se plantea un diagrama de procesos que contiene el esquema que guiará la implementación de la solución propuesta.

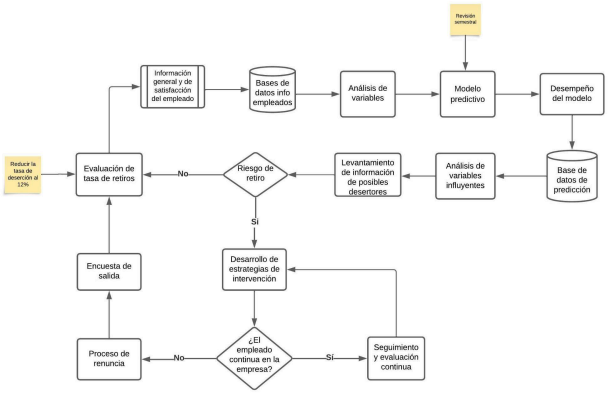


Imagen 1. Diagrama de proceso de diseño de la solución – Elaboración propia

Para profundizar en algunos procesos importantes del diseño de la solución, se realiza un esquema para mostrar qué actividades componen estos.

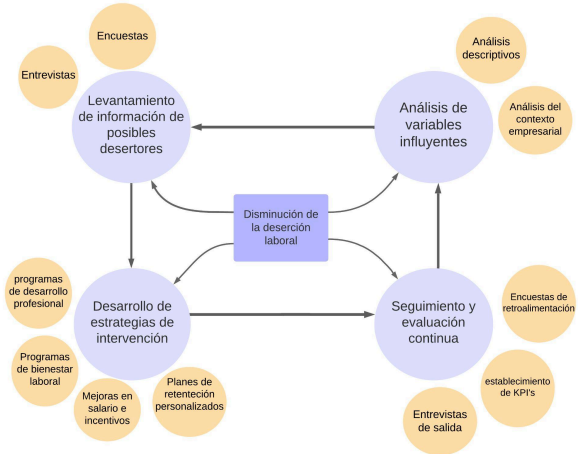


Imagen 2. Esquema componentes de diseño de la solución – Elaboración propia

A. Limpieza y transformación.

En este estudio se inició con la limpieza y transformación de bases de datos que contenían información relevante relacionada con el área de recursos humanos de una organización. En primer lugar, se llevó a cabo un exhaustivo proceso de entendimiento de los datos, examinado las variables consignadas en las bases de datos para entender la importancia para el problema en cuestión. esta fase evidencio la presencia de datos nulos los cuales no fueron

significativos, a lo cual se realizó un tratamiento adecuado para garantizar la utilidad de los datos, se implementa técnica de imputación de datos nulos.

Además, se identificó la necesidad de segmentar las bases de datos con base al año de los registros, lo cual facilita un análisis más detallado. para llevar a cabo se utilizó como lenguaje de programación SQL y Python. Mediante consultas de SQL se extrajeron y dividieron los datos en bases diferentes, una base de datos contiene información correspondiente al año 2015 y otra del 2016, preparándose así para análisis posteriores.

Las bases de datos utilizadas para este trabajo fueron:

general_data: Contiene información general de los empleados.

employee_survey_data: resultados de encuesta realizada a los empleados respecto a su nivel de satisfacción con su empleo actual.

manager_survey_data: resultados obtenidos por los empleados en su última evaluación de desempeño.

retirement_info: Contiene información sobre la renuncia de los empleados.

B.Análisis exploratorio

Después de completar el proceso de limpieza y transformación de los datos, se llevó a cabo un análisis exploratorio que fue fundamental para comprender la naturaleza de los datos antes de realizar cualquier suposición. Este análisis nos permitió identificar patrones y detectar valores atípicos. Para lograrlo, se procedió a graficar tanto las variables categóricas como las numéricas, lo que nos permitió observar y comprender su comportamiento de manera más clara y detallada. En esta gráfica se muestran las variables de la base de datos del 2015, ya que esta se utilizará como base para la creación de los modelos de este caso de estudio.

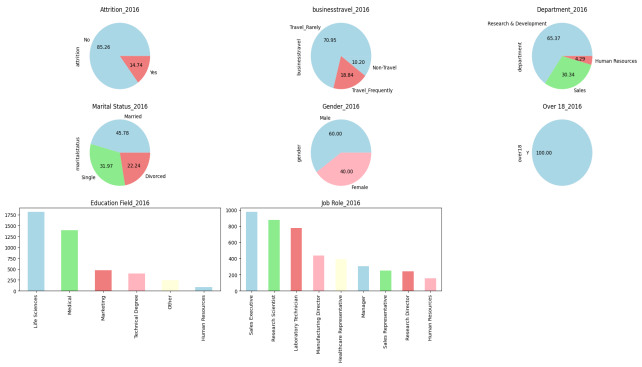


Imagen 3. Barras y tortas para variables categóricas – Elaboración propia

Se puede notar que, en uno de los gráficos de torta, la totalidad del mismo está dada por un solo valor lo que significa que la variable (Over18), toma el mismo valor en cada una de sus filas y se opta por eliminarla de la base de datos ya que se cree que no va a representar ninguna importancia dentro de los modelos posteriores.

A continuación, se realiza una visualización gráfica de las variables numéricas, con el objetivo de analizar su comportamiento y clasificarlas dentro de variables categóricas según fuese el caso más adelante.

La base de datos a la cuál se aplica este análisis es la del año 2015.

Imagen 4. Histogramas para variables numéricas – Elaboración propia

Lo que se puede notar es que hay variables que presentan una distribución completamente uniforme dentro de un solo valor, lo que indica que toma el mismo valor en cada una de sus filas y se opta por eliminar las que se comporten de esta manera (employeecount y standarhours).

Finalmente se grafica la variable objetivo (atrittion) de las base de datos del año 2016, para saber cómo se distribuyen sus datos.

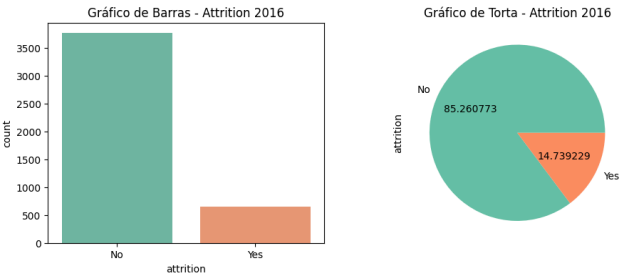


Imagen 5. Barras y torta para variable objetivo – Elaboración propia

Se puede notar que se presenta el No en una mayor proporción con aproximadamente un 84%, mientras que las personas que sí renunciaron el último año tienen un porcentaje de 15%. Esto es muy importante al momento de crear el modelo.

C.Selección de algoritmos y técnicas de modelado.

Para la selección de los modelos esperados a implementar, nos basamos en la evidencia obtenida en el estudio de un caso anterior similar al presente. En el cual los algoritmos de *Regresión Logística* y *Random Forest* fueron los modelos con mayor rendimiento teniendo en cuenta la naturaleza y

comportamiento de los datos.

Además, estos modelos también serán probados bajo la aplicación de selectores de características de las variables incluidas en el modelo.

D.Selección de variables

Lo primero que se hace es separar la variable objetivo (atrittion) del resto de variables que son las predictoras.

- **Variables numéricas iniciales**: age, distancefromhome, education, joblevel, monthlyincome, percentsalaryhike, stockoptionlevel, traingintimelastyear, yearsatcompany, yearssincelastpromotion, yearswithcurrmanager, jobinvolvement, performancerating.

Luego de analizar estas variables por medio de histogramas, se llegó a la conclusión de que hay algunas que pueden ser de naturaleza categórica por lo que se procede a codificarlas y agregarlas en una base aparte de variables categóricas. Estas variables fueron: stockoptionlevel, performancerating, jobinvolvement, joblevel y education.

Ahora se procede a analizar las variables tipo *object.*

- **Variables object iniciales:** businesstravel, department, educationfield, gender, jobrole, maritalstatus, resignationreason y retirementtype..

Para este conjunto de variables se realizaron gráficos de torta donde se encontró que 4 variables se comportan como categóricas por lo que también se agregan a la base de datos aparte. Las variables fueron: maritalstatus, gender, department y businesstravel.

Finalmente se analizan las variables tipo *float.*

- **Variables float iniciales:** numcompaniesworked, totalworkingyears, enviromentsatisfaction, jobsatisfaction, worklifebalance, mean_time.

En este caso se encontraron 3 variables de naturaleza categórica. Las variables son: enviromentsatisfaction, jobsatisfaction y worklifebalance.

Finalmente se convierten en *dummies* todas aquellas variables que se identificaron como tipo categóricas ya que los modelos sólo admiten variables numéricas.

El último paso antes de modelar es unir las variables *float, object* e *int* en una sola base de datos totalmente limpia que es con la que se va a trabajar de ahora en adelante. Esta base se llama *X_total2015.*

Selector de variables Lasso: Al implementar este selector de variables, de 55 variables selecciona nueve a las cuales asigna un valor de coeficiente como lo se muestra en la siguiente imagen.

```
Coeficientes del estimador Lasso: [-0.01899644 -0.         -0.         0.         -0.         -0.
 -0.         -0.02270421  0.         -0.01035558 -0.         -0.
 -0.         -0.         -0.         -0.         -0.         -0.
 -0.         -0.         -0.          0.08036828 -0.         -0.
  0.         -0.         -0.          0.         -0.         -0.
  0.         -0.         -0.         -0.          0.0220281  -0.
 -0.         -0.          0.         -0.         -0.         -0.
  0.          0.0402899  -0.         -0.         -0.          0.02675902
  0.         -0.         -0.02365644  0.         -0.         -0.00197569
  0.        ]
```

Imagen 6. Coeficientes del estimador Lasso – Elaboración propia

Estas variables son posteriormente almacenadas para la aplicación de este selector en un modelo de *Regresión Logística.*

E.Selector de variables basados en importancia de características

Este algoritmo estima la importancia de las características de todas las variables aplicándolas a modelos de predicción como *Random Forest*. Los valores de importancia se traducen en la participación de estas variables en el modelo.

Para seleccionar las variables con más participación en el modelo, se extraen las variables que tengan una importancia mayor a 0.01, obteniendo las variables que se observan en la siguiente imagen. Estas variables serán utilizadas para la implementación del modelo *Random Forest.*

	Feature	Random Forest Importance
0	age	0.085114
1	totalworkingyears	0.075246
2	monthlyincome	0.073373
3	yearsatcompany	0.058173
4	distancefromhome	0.054739
5	percentsalaryhike	0.052838
6	numcompaniesworked	0.045066
7	yearswithcurrmanager	0.044399
8	yearssincelastpromotion	0.037492
9	traningtimeslastyear	0.034429
10	maritalstatus_Single	0.021202
11	environmentsatisfaction_1.0	0.018542
12	jobsatisfaction_1.0	0.017699
13	businesstravel_Travel_Frequently	0.016245
14	worklifebalance_3.0	0.014212
15	educationfield_Life Sciences	0.013588
16	environmentsatisfaction_4.0	0.013366
17	jobsatisfaction_4.0	0.013185
19	worklifebalance_1.0	0.012097
20	jobsatisfaction_2.0	0.011914
22	jobrole_Sales Executive	0.011570
35	resignationreason_no aplica	0.009442
37	department_Human Resources	0.008705
38	retirementtype_no aplica	0.008542
39	retirementtype_Resignation	0.008078
41	businesstravel_Non-Travel	0.007008
46	educationfield_Marketing	0.005794
47	educationfield_Technical Degree	0.005574

Imagen 7. Variables seleccionadas por importancia de características – Elaboración propia

F. Comparación y selección de técnicas

RESULTADOS Y ANÁLISIS DE MODELOS

A continuación, se mostrarán los resultados de los modelos realizados.

Modelo de regresión logística (RL)

El primer paso es separar el conjunto de datos de entrenamiento y testeo en donde se obtuvieron los siguientes resultados:

```
Tamaño del conjunto de entrenamiento. X:  (3528, 46)  Y:  (3528,)
Tamaño del conjunto de validación. X:  (882, 46)  Y:  (882,)
```

Imagen 8. Tamaño de conjuntos del modelo de RL – Elaboración propia

Luego se procede a realizar el modelo y su respectiva matriz de confusión.

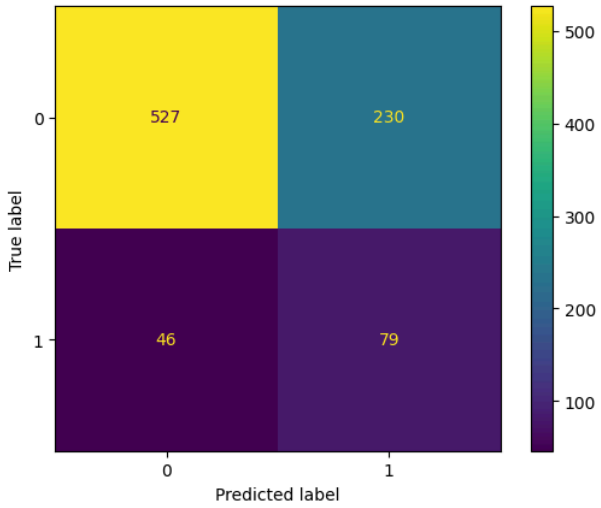


Imagen 9. Matriz de confusión modelo de RL – Elaboración propia

Cuando se analiza la matriz de confusión se puede notar que los falsos positivos son muy altos y son los que determinan la poca calidad del modelo. Estos falsos positivos lo que dicen es que el modelo no está logrando predecir correctamente las personas que de verdad NO van a renunciar a la empresa y lo que está haciendo es decirnos que van a renunciar cuando realmente NO lo van a hacer. Lo mismo pasa con los falsos negativos, pero en menor medida; lo que dice es que una persona no va a renunciar cuando realmente sí lo va a hacer.

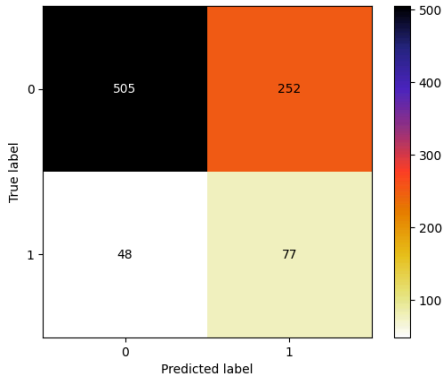
```
Precision: 0.255663430420712
Recuperacion: 0.632
F1-score: 0.3640552995391705
Especificidad: 0.6961690885072656
```

Imagen 10. Métricas del modelo de RL – Elaboración propia

Los resultados anteriores se ven reflejados en las métricas donde dicen que la precisión del modelo es de tan solo el 25.5% y el *f1-score* es de poco más del 36% lo que significa que uno de los problemas del modelo puede radicar en el clasificador de las variables. Además, cabe resaltar que, aunque todas las métricas son muy importantes, en este contexto es mucho más importante predecir quién se va a ir (recuperación) para tomar acciones a tiempo, que saber quién no se va a ir (precisión).

A. *Modelo de regresión logística con selector Lasso -Alpha de 0.01*

Inicialmente se observa y analiza la matriz de confusión de este modelo.



A pesar de que se aplica un selector de variables que supone seleccionar las variables con más influencia en el modelo, vemos que los falsos-positivos aumentan, por tanto, el modelo no realiza una buena predicción.

Luego, se observan sus métricas de desempeño:

```
Precision: 0.23404255319148937
Recuperacion: 0.616
F1-score:  {0.3392070484581498}
Especificidad: 0.667107001321004
```

Imagen 11. Métricas del modelo de RL con selector Lasso con Alpha 0.01 – Elaboración propia

Se puede notar que ninguna de las métricas bajaron respecto al modelo anterior, por tanto, no se considera este modelo como el apropiado para la solución del caso de estudio.

B. Random Forest classifier.

-Hiperparametros estándar.

Para este modelo, se utilizan las variables seleccionadas por sus características dadas en el selector de variables descrito anteriormente y se somete a un modelo de Random Forest son hiper parámetros estándar. .

```
Train - Accuracy : 1.0
Train - classification report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3003
           1       1.00      1.00      1.00       525

    accuracy                           1.00      3528
   macro avg       1.00      1.00      1.00      3528
weighted avg       1.00      1.00      1.00      3528

Test - Accuracy : 0.9727891156462585
Test - classification report :
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       757
           1       0.89      0.92      0.91       125

    accuracy                           0.97       882
   macro avg       0.94      0.95      0.94       882
weighted avg       0.97      0.97      0.97       882
```

Imagen 12. Métricas Random Forest Classifier – Elaboración propia

A continuación, se observa la matriz de confusión de este modelo.

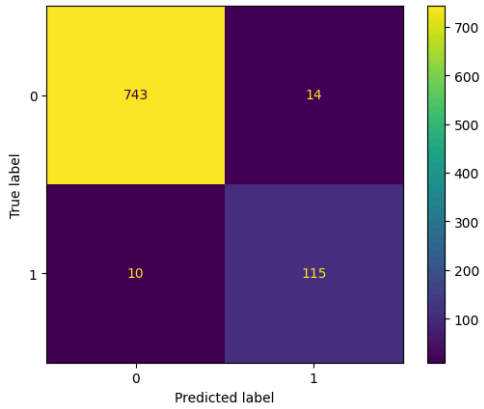


Imagen 13. Matriz de confusión del modelo Decision Tree Classifier - Elaboración propia

El modelo ha logrado una precisión perfecta en el entrenamiento, lo que podría indicar un posible sobreajuste. Pero el rendimiento en el conjunto de prueba también es muy bueno, con una precisión general del 97.27% lo que puede indicar que el modelo no aprendió particularidades de los datos y sí aprendió las características generales de los datos.

- *Tuneo de hiperparámetros.*

El primer paso es definir todos los hiperparámetros del *Random Forest Classifier* de acuerdo a la documentación. Los hiperparámetros que mejor se ajustan al modelo son:

```
RandomForestClassifier(class_weight='balanced', criterion='entropy',
                       max_depth=30, max_features=0.5, random_state=42)
```

Imagen 14. Mejores hiperparámetros del modelo de Random Classifier – Elaboración propia

Luego de saber cuáles son los mejores hiperparámetros para el modelo, se procede a calcular la matriz de confusión y las métricas para saber cómo se está comportando.

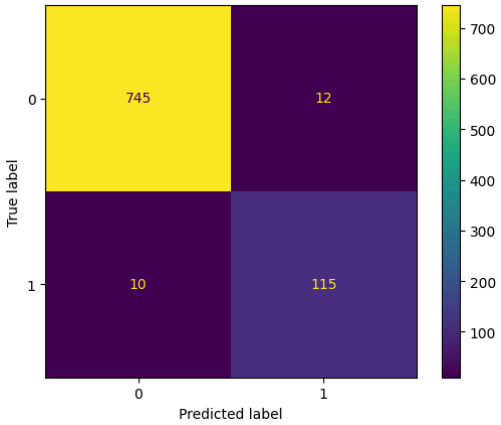


Imagen 15. Matriz de confusión del modelo Decision Tree Clasifier con tuneo de hiperparámetros- Elaboración propia

El modelo ha tenido nuevamente una precisión perfecta en el entrenamiento como era de esperarse, lo que podría indicar por segunda ocasión un posible sobreajuste. El rendimiento en el conjunto de prueba sigue siendo muy bueno, incluso presentando un pequeño aumento, con una precisión general de más del 98% lo que puede indicar que el tuneo de hiperparámetros sirvió para que los posibles errores en la predicción sean mínimos y así el modelo logre ayudar a tomar decisiones en el año 2017.

G. Variables influyentes.

A través de la siguiente tabla, se muestra el top 10 de variables que mejor explican la variable objetivo.

	columns	importances
2	monthlyincome	0.107729
0	age	0.090570
9	totalworkingyears	0.076656
1	distancefromhome	0.071779
5	yearsatcompany	0.059215
7	yearswithcurrmanager	0.058966
3	percentsalaryhike	0.058095
8	numcompaniesworked	0.049847
4	trainingtimeslastyear	0.047023
6	yearssincelastpromotion	0.046460

Imagen 16. Top 10 variables influyentes en el modelo - Elaboración propia

Conclusiones

A continuación, se postula la conclusión general de este estudio.

Entre las variables más significativas, destaca en primer lugar el ingreso mensual. Un análisis exploratorio revela que la mayoría de los empleados se encuentran en los rangos salariales más bajos. En segundo término, la edad emerge como otro factor relevante. Según el análisis exploratorio, esta variable exhibe una distribución aproximadamente normal, con una concentración destacada entre los 30 y 40 años. Esto sugiere que las personas que se retiran están en una edad óptima para trabajar, ya que poseen experiencia considerable y aún conservan relativa juventud.

Además de estos factores, se repiten con frecuencia los años de servicio del empleado, tanto en la empresa actual como en empleos anteriores, así como los años desde la última promoción. Esta repetición indica que, con el tiempo, los empleados tienden a retirarse por diversas razones.

Un hallazgo sorprendente es la duración del empleado bajo el mando del actual gerente.

En cuanto a las propuestas, la primera sugerencia es revisar los salarios de los empleados y establecer un plan de méritos basado en habilidades, trayectoria y experiencia. Esto podría mejorar la percepción de la remuneración laboral y aumentar la satisfacción en el trabajo.

Otra propuesta para el año en curso implica combinar salario y edad para identificar a aquellos empleados de entre 30 y 40 años con mayor antigüedad en la empresa y experiencia. Se podrían ofrecer incentivos monetarios, días de descanso adicionales y otros beneficios para aumentar su satisfacción laboral, considerando que están en una etapa de vida donde buscan alcanzar metas profesionales y personales.

Finalmente, resulta preocupante que el liderazgo del actual gerente esté vinculado a los retiros de empleados. Se recomienda una auditoría periódica para evaluar su desempeño tanto profesional como personal. Identificar si los empleados perciben un trato inadecuado o deficiente desempeño por parte del gerente permitirá tomar medidas para mejorar la situación, o en última instancia, considerar su retiro de la empresa.

Referencias

[1] Amazon, «AWS Amazon,» Amazon Web Services, 1 Enero 2022. [En línea]. Available: https://aws.amazon.com/es/what-is/python/. [Último acceso: 25 Septiembre 2022].

[2] A. S. Alberca, «Aprende con Alf,» Aprende con Alf, 14 Junio 2022. [En línea]. Available: https://aprendeconalf.es/docencia/python/manual/pandas/. [Último acceso: 25 Septiembre 2022]
[3] scikit-learn, « scikit-learn Machine Learning in Python» Diciembre 2022. [En línea]. Available: https://scikit-learn.org/stable/

Anexo

- *imagen 17. Gráfica de las 10 variables influyentes en el modelo*

- *Repositorio github*