UNIVERSIDAD
DE ANTIOQUIA

TRABAJO ANALÍTICA DE DATOS APLICADA EN RECURSOS HUMANOS

SANTIAGO GÓMEZ BERRÍO
DIEGO ANDRÉS LUNA PATERNINA
MARIA CLARA SALAZAR DUQUE

JUAN CAMILO ESPAÑA LOPERA
ANALÍTICA DE DATOS

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
MEDELLÍN

2023

Analítica de Datos Aplicada en RRHH

1. Diseño de la solución

Para diseñar una solución, primero se debe identificar el problema y su principal causa. En este caso de estudio, el problema es la alta tasa de retiro en una EPS en la India, lo cual genera diversos subproblemas.

Subproblemas:

- **S1:** Los proyectos en los que los empleados trabajan se atrasan y se pueden ver comprometidas las fechas establecidas y esto puede llegar a afectar la satisfacción de clientes y usuarios.
- **S2:** El trabajo del área de selección se aumenta y por lo tanto implica tener un área de mayor tamaño para el reclutamiento de las personas que se van.
- **S3:** Los empleados que permanecen tienen que trabajar más para cubrir las labores de las personas que se retiran y para la capacitación de las nuevas que llegan.
- **S4:** El conocimiento que tenían las personas y su experiencia se pierde y, debido a esto, el tiempo para ejecutar procesos aumenta.

Factores de priorización:

a. Nivel de gravedad - ¿cuánto daño ocasiona en la empresa? (45%).
b. Ejecutabilidad - ¿cuál es la posibilidad de solución teniendo en cuenta los datos? (35%).
c. Beneficio - grado de valor agregado que genera la solución (20%).

	Factor a	Factor b	Factor c	Total
S1	8	4	9	6.8
S2	8	6	7	7.1
S3	7	6	8	6.85
S4	9	5	7	7.2

Tabla 1. *Priorización de subproblemas*

Por otra parte se recurrió a la herramienta de los 5 porqués con el fin de identificar la raíz del problema, para posteriormente proponer una solución adecuada.

→¿Por qué los empleados se están retirando?
→ Están desmotivados ¿por qué?
→ Se sienten sobrecargados ¿por qué?
→ Por la alta rotación de personal ¿por qué?
→ No hay un correcto acompañamiento al empleado durante el cumplimiento de sus labores.

De acuerdo a la identificación de la causa del problema planteado, el área de RRHH es el responsable de realizar acompañamiento a los empleados para evitar su temprana deserción, haciendo un mayor aprovechamiento del conocimiento del personal que contratan sin llevarlos a un límite desgastante y asignándoles de manera adecuada sus labores dentro de la empresa. Para lograr esto, junto con el equipo de analítica diseñarán un modelo de clasificación, el cual determine para cada empleado su probabilidad de renuncia, mediante la identificación de los factores que influyen en esta problemática. El modelo predecirá en un periodo de tiempo anual con el fin de estar lo más equilibrado posible, y dichas predicciones se harán por lotes dado que el modelo se entrena con todos los datos disponibles antes de la predicción, y como los datos de

entrenamiento cambian (ya que reciben nuevos empleados), el modelo deberá ser reentrenado desde cero.

Toda la información obtenida y analizada se le entregará a los directivos de la EPS, por medio de un informe ejecutivo con las estrategias para cada segmento, con el fin de que estén atentos a la posible renuncia de algún empleado y puedan tomar las medidas necesarias para evitar que se retire, mitigando los costos que implica para la empresa. Para tener una visualización más clara de lo anteriormente dicho, véase la *figura 1*.

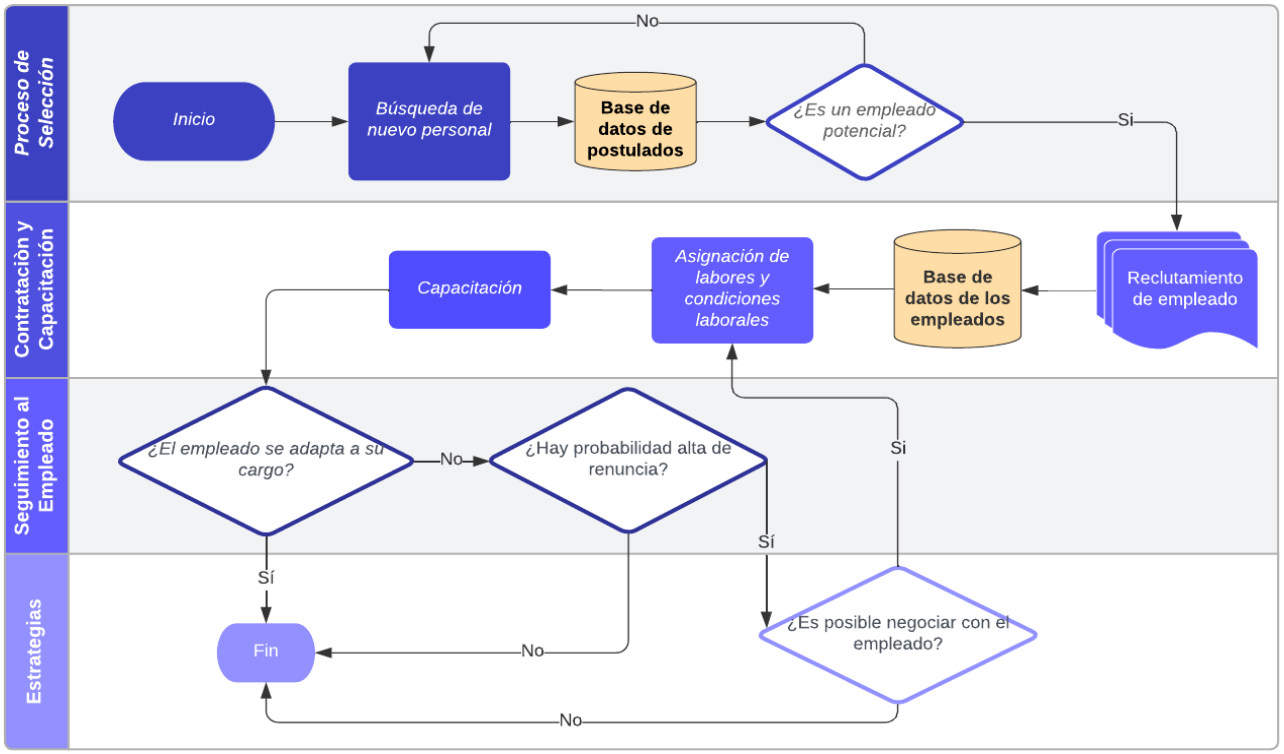

Figura 1. *Diagrama del diseño solución*

2. **Limpieza y transformación:**
   Se realizó un dataframe uniendo las bases de datos de la información general de los empleados junto con la encuesta realizada a los empleados sobre satisfacción laboral, y la encuesta de desempeño de los empleados realizada por parte de los jefes. Posteriormente, se verificó que no se presentaran datos duplicados y la base de datos de los empleados retirados se tuvo en cuenta aparte para no generar datos nulos. Adicionalmente, para manipular mejor los datos, se optó por usar un código que pusiera todos los nombres de las variables en minúscula. Por otra parte, como la variable de salida o interés, es retirados, se dividió en dos grupos: los empleados que están activos y que fueron despedidos, frente a los que renunciaron, con el fin de delimitar el estudio.

3. **Análisis Exploratorio:**
   Se plantearon una serie de preguntas con el fin de identificar qué información se puede obtener de los datos y cómo se comportan las variables con respecto a la renuncia de los empleados de la EPS.

   Es así cómo se deducen las siguientes conclusiones gracias al análisis de los datos:
   - Las razones por las que más se retiran los empleados de la EPS son motivos diferentes al salario o al estrés, los cuales se considera que pueden ser debido a mejores oportunidades en otras compañías, y valoración de su conocimiento.

- En el departamento de RRHH se evidencian mayores retiros, debido a que es muy pequeño en relación a los demás, es por ello que probablemente sientan un nivel de trabajo alto y decidan retirarse de la compañía por estrés, salario u otras razones.
- Se observa que muchos de los retirados ascendieron rápidamente en la compañía durante sus primeros años, algunos incluso sin tener el año de trabajo en la compañía y en general ascienden con una media de entre 2 y 3 capacitaciones por año.
- Se puede observar que existe una relación inversa entre nivel de trabajo y el nivel de satisfacción, ya que a medida que aumenta el nivel de trabajo el nivel de satisfacción de los empleados va disminuyendo.
- La edad del empleado no es un factor influyente con respecto a los ascensos; sin embargo, las mayores renuncias se presentan para las personas con 28, 29 y 31 años. Cosa contraria a lo que sucede con las personas con más años de edad, ya que tienden a quedarse en la compañía.
- En general, se observa que en todas las áreas el nivel de satisfacción promedio está por encima de 2 casi 3 lo que se supone es bueno, solo que en los empleados que están muy insatisfechos si se debe tener cuidado y generar alerta para predecir posibles renuncias, en especial en los campos de educación de medicina y ciencias de la vida.
- Se evidencia claramente cómo mes a mes les pagan más a los hombres que a las mujeres, aunque hay que tener cuidado ya que en la compañía hay más empleados hombres que mujeres.
- Se observa como las capacitaciones recibidas no se ven reflejadas en mayor salario mensualmente, incluso empleados con una capacitación ganan lo mismo o más que personas que han presentado seis capacitaciones.

4. **Selección de algoritmos:**
   Debido a que lo que se busca con el modelo es predecir si es alta la probabilidad de que un empleado renuncie, se tuvo en cuenta el modelo Linear SVC, una red neuronal y una regresión logística, los cuales lograrían adecuarse al caso de estudio, ya que podrían aprender y modelar las relaciones entre los datos de entrada y salida, categorizando los datos de manera ideal para obtener predicciones certeras. Adicionalmente, se hizo uso del Random forest, puesto que es fácil de aplicar e interpretar, es estable y por lo general presenta buenas coincidencias, se usa en modelos de regresión o clasificación, y presenta muchas ventajas en comparación con otros algoritmos de datos. Para los modelos se tuvo en cuenta una sola base de datos de la cual el 33% fueron para la prueba y el resto (67%) para entrenar al modelo.

5. **Selección de variables:**
   Se realizó una matriz de correlación, la cual buscaba identificar qué variables tenian mayor correlación entre sí, o sí dado el caso, observar cuáles presentaban alta colinealidad para eliminarlas. Se pudo notar que al prescindir de las variables “over18” y “StandardHours”, no se afectaría el modelo a desarrollar, ya que estas no aportan información relevante con respecto a la variable de interés (retirados). Posteriormente, se usó dummies para separar las variables por cada una de sus categorías y poder escalar, estandarizar y normalizar los datos mediante un arreglo en el cual X contiene todas las variables y Y solo la variable target. Después, se recurrió al método K best que calificaba a cada una de las variables asignándoles una puntuación, de allí se sacaron las de mayor puntuación (superior a 20), y de esta manera se obtuvieron 10 variables.

6. **Comparación y selección de técnicas:**
   Para la selección del modelo lo que se hizo fue tener en cuenta el rendimiento del modelo, como se muestra en la siguiente tabla:

	Modelo Random Forest	Modelo Regresión logística	Neural Network	Linear SVC
Accuracy	97,11%	95,5%	96,7%	91%
Recall	0: 0.98 1: 0.90	0: 0.97 1: 0.91	0:	0: 0.99 1: 0.85

Tabla 1. *Modelos*

Adicionalmente, se realizó una confusion matrix para cada modelo, donde se puede observar que el modelo de Random Forest tuvo mayores aciertos al momento de predecir la renuncia de los trabajadores. Como se muestra en la *figura 2.*

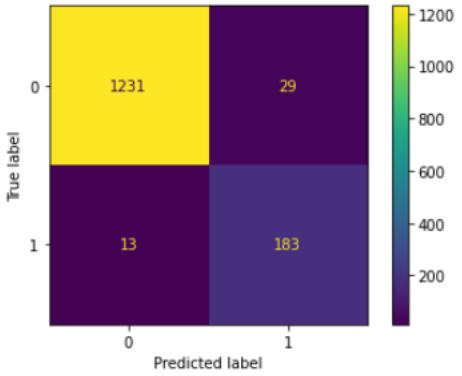

Figura 2. *Confusion matrix del modelo Random Forest*

7. **Afinamiento de hiperparámetros:**
   Debido a que el algoritmo escogido fue Random Forest, dentro de este mismo se pudo verificar si el modelo se encontraba desbalanceado, para ello se hizo uso del class_weight='balanced' y el class_weight='balanced_subsample'; sin embargo, no se vio afectado en gran medida, ya que los hiperparametros más importantes seguían siendo los mismos, y el accuraccy (exactitud) disminuyó un poco al aplicar el 'balanced_subsample'.

8. **Evaluación y análisis del modelo:**
   De acuerdo con las métricas de evaluación que se utilizaron para evaluar el modelo, es decir, confusion matrix, accuracy y clasification report (f1-score, recall, precision), se muestra un resultado adecuado dando un porcentaje de precisión del 97%. Esto refleja un apropiado comportamiento del modelo, acertando en 1231 casos en los que las personas no renunciaron y un total de 183 personas que sí lo hicieron, a su vez, tiene un desacierto con 42 personas (totalizando falsos positivos y falsos negativos).

   Además, se realizó un gráfico donde se pueden ver las variables que contiene la importancia de las características del modelo, con sus respectivas categorías. De esta manera, se puede observar que las variables con mayor grado de importancia para el modelo, son: "Attrition", seguido de "totalworkingyear", "age" y "yearsatcompany" como se muestra en la *figura 3.*

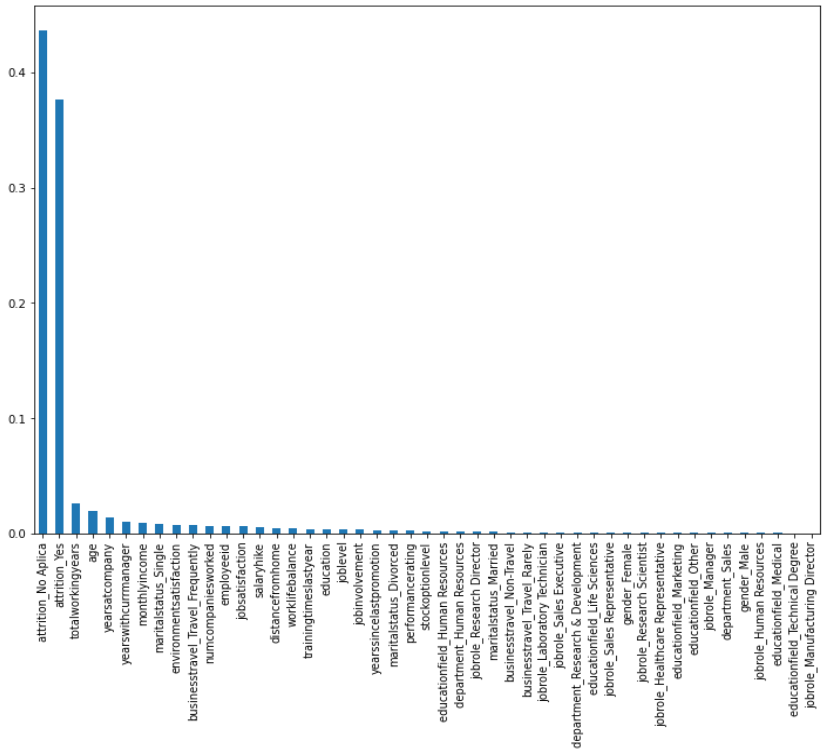

Figura 3. *Variables más relevantes para el modelo*

9. Despliegue del modelo:

Una vez seleccionado el modelo de Random Forest, el cual está compuesto de los features anteriormente mencionados, se procedió a verificar cuál era el comportamiento del modelo, donde de 1456 datos que tiene de entrenamiento, es capaz de predecir con un 97% de aciertos cuántas personas están próximas a renunciar, para así alertar a la compañía y tomar medidas preventivas.

En la *figura 4* se puede evidenciar el comportamiento del modelo, con los datos entrenados y los datos que predice, demostrando que siguen casi la misma distribución o comportamiento en ambos casos, lo cual va en coherencia con el resultado del 97% de rendimiento.

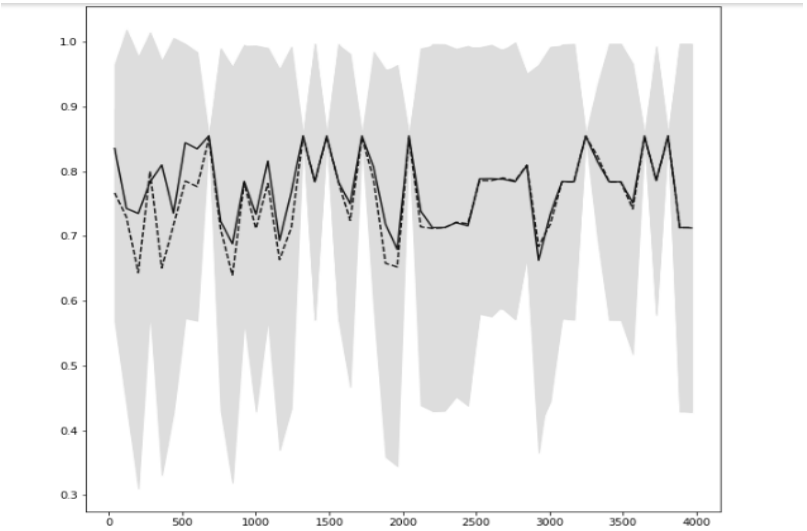

Figura 4. *Predicciones vs entrenamiento del modelo Random Forest*

Además de esto, se logró observar que del total de datos usados para que el modelo realice sus predicciones, 212 empleados están próximos a renunciar, lo cual significa casi el 15% de los datos, y va en concordancia con el problema actual de la compañía.