**APLICACIONES DE LA ANALÍTICA: RECURSOS HUMANOS**

****

**POR:**

**DANIELA LÓPEZ ANAYA**

**VALENTINA MÚNERA PULGARÍN**

**DOCENTE:**

**JUAN CAMILO ESPAÑA**

**INGENIERÍA INDUSTRIAL**

**FACULTAD DE INGENIERÍA**

**UNIVERSIDAD DE ANTIOQUIA**

**2023**

# DISEÑO DE LA SOLUCIÓN

El diseño de la solución (observar Ilustración 1), consiste en un modelo de aprendizaje supervisado de tipo clasificación, el cual realizará la predicción si un empleado va a renunciar o no. En caso de que el modelo arroje como resultado empleados que van a renunciar, esta información será enviada al área de Recursos Humanos de la empresa, los cuales deben ponerse en contacto con estos empleados para constatar dicha situación y así mismo, diseñar estrategias que permitan la retención del personal. Seguidamente, el área de recursos humanos deberá enviar al área de Analítica, datos actualizados trimestralmente, con el fin de entrenar de modelo y realizar las nuevas predicciones para los siguientes tres meses.

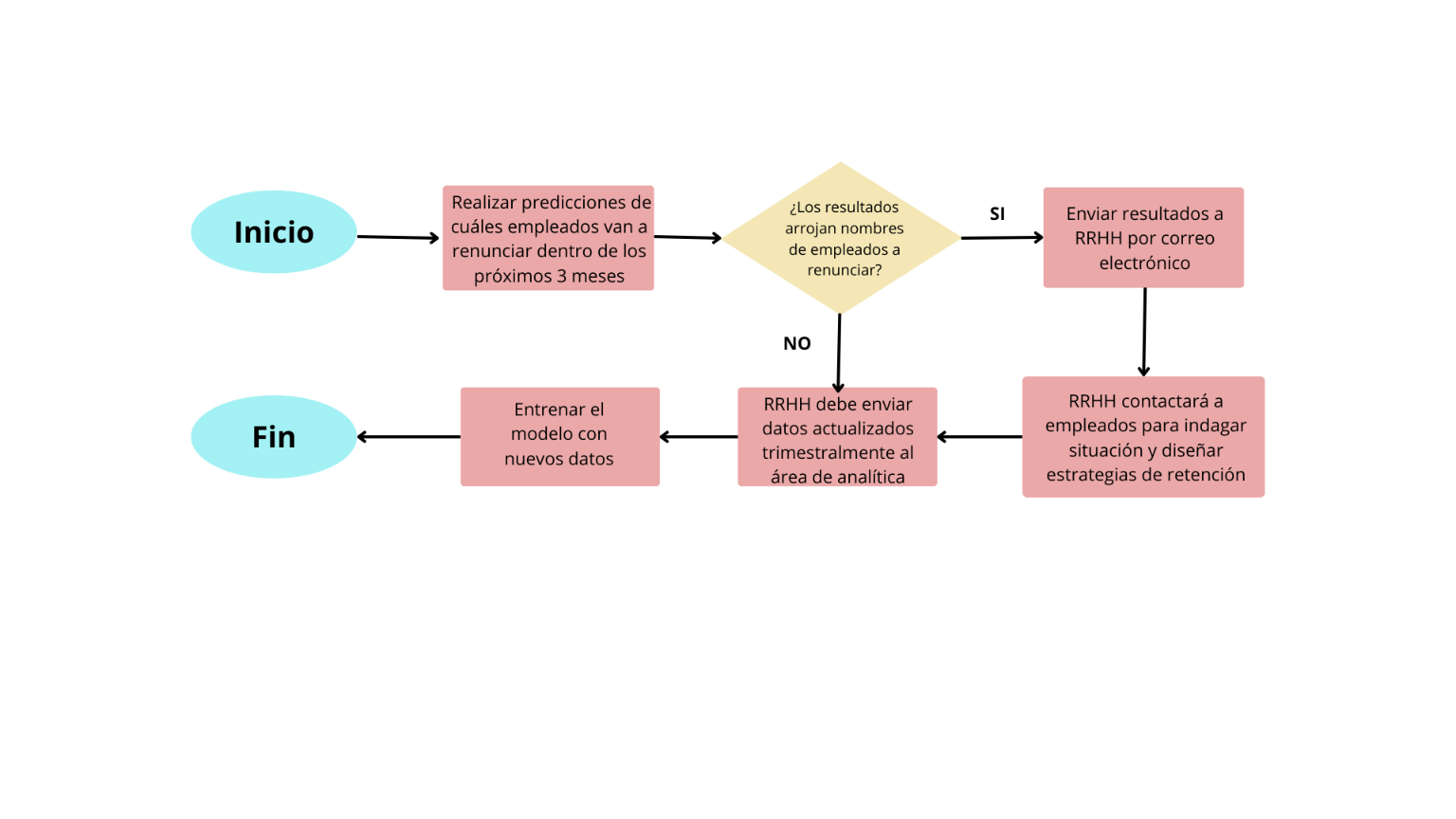


Ilustración 1. Diseño de solución

# LIMPIEZA Y TRANSFORMACIÓN

Inicialmente, fueron suministrados seis archivos en formato CSV para examinar. Tras su estudio, se determinó la pertinencia de seleccionar cuatro de ellos, los cuales fueron:

1. employee\_survey\_data.csv
2. general\_data.csv
3. manager\_survey\_data.csv
4. retirement\_info.csv

En un contexto general, los archivos #1, #2 y #3 contienen registros de 4410 de personas, independiente si se encuentran activos o no activos en la empresa. Se encontró información sobre nivel de satisfacción laboral, satisfacción del ambiente laboral, departamento al que pertenece, género, edad, años en la compañía, entre otros datos. Por otro lado, el archivo # 4 contiene información sobre 711 personas que se han retirado de la empresa, ya sea por despido o renuncia y contiene principalmente información como fecha de retiro, tipo de retiro y su razón.

De los archivos mencionados anteriormente, se realizó un proceso de limpieza y transformación, de la siguiente forma:

1. Análisis general de la cantidad de variables y registros.
2. Análisis de valores únicos por cada variable.
3. Estudio de valores nulos por variable y cálculo del porcentaje de nulos por archivo.
4. En caso de presencial de nulos, selección y aplicación de métodos para tratamiento de estos.
5. Observación del tipo de variable y transformación según conveniencia.
6. Para variables de tipo texto, se convirtieron a minúsculas, como también los nombres de variables a minúscula.

Cabe resaltar que, del archivo #4 se eliminaron los registros de los empleados de los cuales su tipo de retiro era por despido, dado que para el modelo que se plantea, dicho datos no son de interés.

Finalmente, Se realizó la unión de los cuatro archivos por medio de la variable ID Empleado, permitiendo así obtener un archivo consolidado.

# ANÁLISIS EXPLORATORIO

Seguidamente, tras lograr obtener un archivo consolidado, y teniendo como objetivo la construcción de un modelo de aprendizaje supervisado de tipo clasificación, se procedió a realizar el análisis exploratorio de los datos, encontrando lo siguiente.

En la Ilustración # 2, observamos que de los 641 registros de personas que han renunciado, el 28.9% tenía un nivel de satisfacción del ambiente laboral bajo, 18.4% medio, 27.1% alto y el 25.7% muy alto, acumulando así, las dos categorías más bajas un 47.3%.

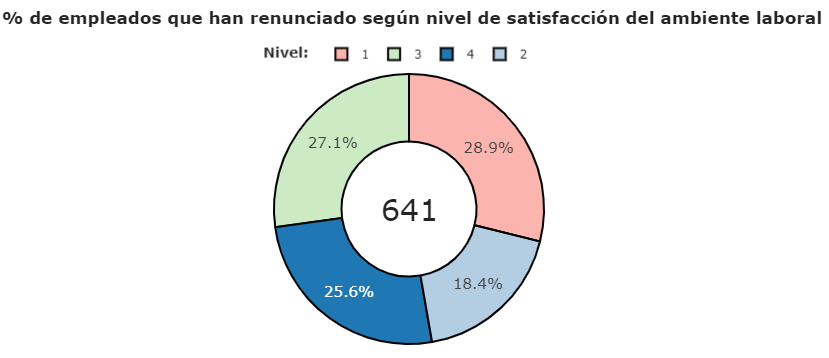


Ilustración 2. Porcentaje de empleados que han renunciado según su nivel de satisfacción del ambiente laboral.

En cuanto a la satisfacción laboral, se encontró que para el 27.5% del total de personas que renunciaron su nivel fue bajo, 19.7% medio, 31.7% alto y 21.2% muy alto. Teniendo así, entre las dos categorías más bajas un total de 47.2%.

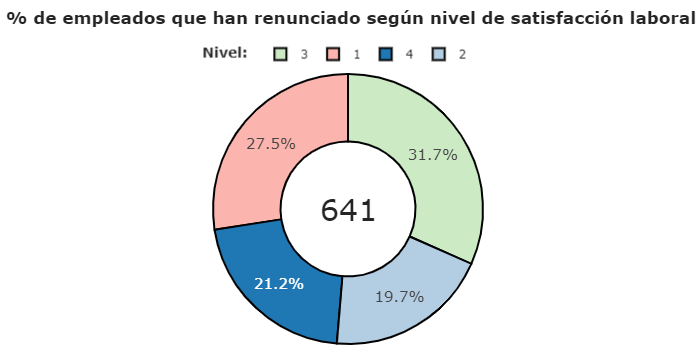


Ilustración 3. Porcentaje de empleados que han renunciado según su nivel de satisfacción laboral.

Por otro lado, se exploró acerca de la cantidad de personas que han renunciado por género y el total de registros, observar Ilustración 4. Encontrándose que, de 1745 registros de mujeres renunciaron 251 de ellas, equivalentes al 14.4%. De forma similar, de 2595 registros de hombres renunciaron 390 de ellos, equivalentes al 15%.

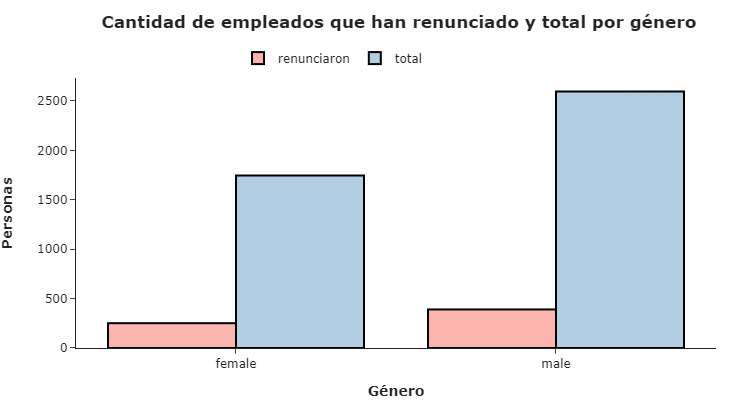


Ilustración 4. Cantidad de empleados que han renunciado y total por género.

Referente a la razón de retiro, observar Ilustración 5, se encontró que el 29.5% de las personas que se retiraron su motivo fue relacionado con el salario, el 20.1% estrés y el 50.4% se atribuyen a otros motivos.

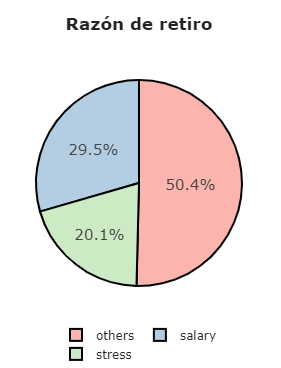


Ilustración 5. Razón de retiro.

Finalmente, se realizó una exploración acerca de la evolución de la cantidad de retiros por mes durante el año 2016, hallando que la menor cantidad de retiros ocurrió cerca del 30 de septiembre y la mayor cantidad en julio.

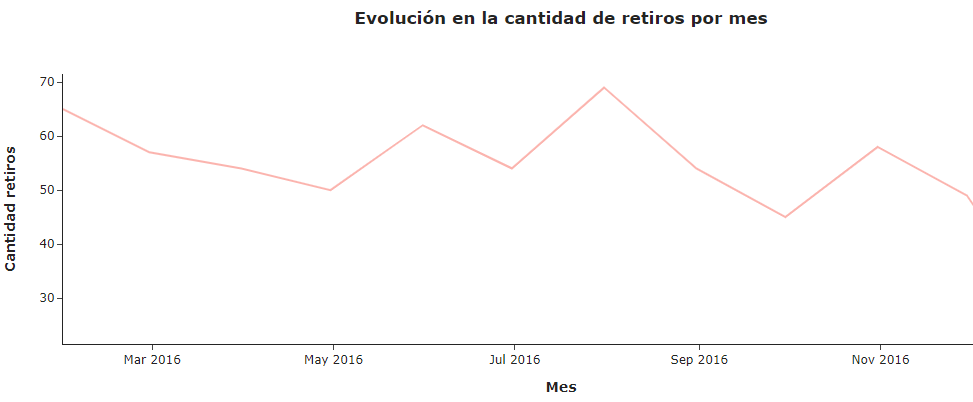


Ilustración 6. Evolución en la cantidad de retiro por mes durante el año 2016.

# SELECCIÓN DE ALGORITMOS

Teniendo en cuenta las características del problema y solución planteada de un modelo de predicción de aprendizaje supervisado de tipo clasificación y siguiendo la guía de Scikit Learn (Scikit Learn, 2023), se seleccionó dos tipos de modelos para aplicar: SGD Classifier y Kernel aproximation. La ruta seguida para seleccionar los anteriores modelos se puede visualizar en la Ilustración 7.

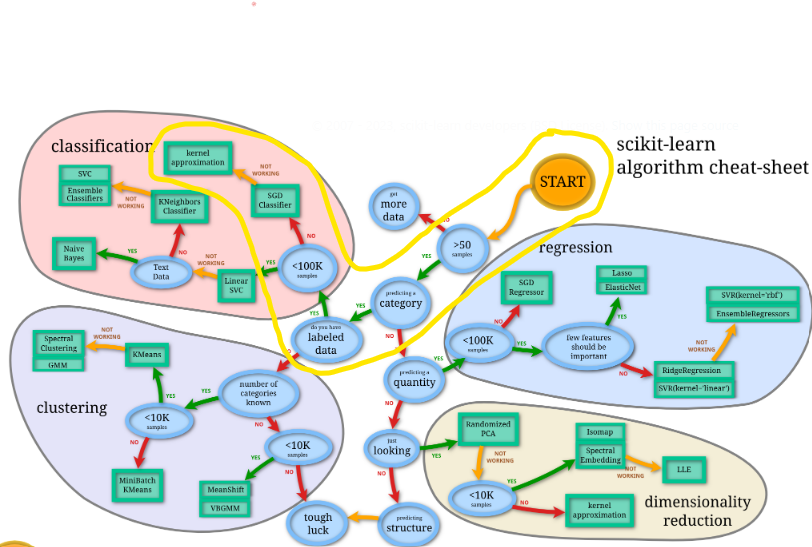


Ilustración 7. Ruta Scikit Learn.

# SELECCIÓN DE VARIABLES

Aplicando los métodos de Select Kbest y Recursive Feature Elimination, se seleccionaron 9 features, las cuales son:

* Businesstravel
* Jobsatisfaction
* Environmentsatisfaction
* Department
* Worklifebalance
* Trainingtimeslastyear
* Numcompaniesworked
* yearssincelastpromotion

# COMPARACIÓN DE MODELOS

Inicialmente, se platearon dos modelos de tipo SGD Classifier y un modelo para Kernel aproximation, obteniendo los siguientes scores (observar Tabla # 1).

|  |  |
| --- | --- |
| **Modelos** | **Score** |
| modelo1 = SGDClassifier(loss="perceptron", penalty="l2", max\_iter=500) | 0.8548 |
| modelo2 = SGDClassifier(loss="huber", penalty="l2", max\_iter=500) | 0.8771 |
| modelo3 = SGDClassifier(max\_iter=250, random\_state=0, tol=1e-3) | 0.8609 |

Tabla 1. Comparación de modelos

Dado los resultados obtenidos para el score de cada modelo, se eligió el mayor de ellos correspondiente al modelo # 3.

# AFINAMIENTO DE HIPERPARÁMEROS

Para la fase de afinamiento de hiperparámetros del modelo # 3, se realizó una función la cual realiza la evaluación de cada hiperparámetro con diferentes valores y finalmente, selecciona los que presenten mayor recall, obteniendo como resultado los siguientes valores:

* 'max\_iter': 150
* 'random\_state': 9
* 'tol': 0.01

# EVALUACIÓN Y ANÁLISIS DEL MODELO

Aplicando los métodos de evaluación de desempeño de K Fold - Cross Validation y Repeat Random Test, se puede observar en la Tabla # 2 como los resultados son superiores al 80%.

Referente al método de KFold - Cross Validation, el cual es una técnica con baja varianza y mayor precisión, dado que el algoritmo se entrena y evalúa varias veces con diferentes datos se obtuvo como resultado un rendimiento del 82%.

Adicionalmente, en el método Repeat Random Test, el cual es una variación del método explicado anteriormente, caracterizado por su velocidad y reducción de la varianza en el rendimiento, se obtuvo un resultado de 84%.

|  |  |
| --- | --- |
| **Método de evaluación de desempeño** | **Resultado** |
| KFold - Cross Validation | 0.82 |
| Repeat Random Test | 0.84 |

Tabla 2. Evaluación de desempeño.

Ahora bien, en cuanto a las métricas de desempeño se aplicó Classification Accuracy, la cual es la métrica más común para problemas de clasificación, obteniendo un resultado del 85.5%, sin embargo, este resultado no es confiable, ya que nuestra variable Target no tiene un número igual de observaciones para cada clase.

Así mismo, se utilizó la métrica de desempeño de Matriz de Confusión (Ilustración # 8), de la cual se puede observar que el modelo predijo certeramente que 1000 personas permanecerán en la empresa cuando realmente lo harán, sin embargo, el modelo también predijo que 120 personas permanecerán en la empresa cuando realmente renunciaran. Adicionalmente, el modelo también predijo que 60 personas renunciaran cuando realmente lo harán y 122 personas renunciarían cuando realmente no lo harán. Lo anterior puede explicarse, ya que existen mayor cantidad de datos de personas que permanecen en la empresa en contraposición con las personas que han renunciado.

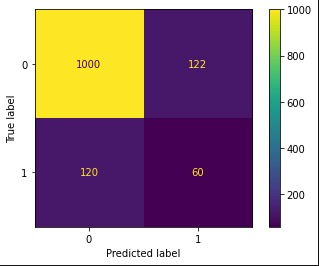


Ilustración 8. Matriz de Confusión.

# DESPLIEGUE DEL MODELO

El despliegue del modelo se realizó por medio de la librería Pickle, la cual permite un proceso por de una jerarquía de objetos Python los cuales se convierten en un flujo de bytes.

# REFERENCIAS

Scikit Learn. (7 de Marzo de 2023). *scikit-learn.* Obtenido de scikit-learn: https://scikit-learn.org/stable/tutorial/machine\_learning\_map/index.html