**PROYECTO MACHINE LEARNING FRANCISCO QUINTERO**

**RESIGNATION PREDICTION**

**MEMORIAS**

**INTRODUCCION**

Nombre de la empresa = “X\_Recruit\_HR”

Años de servicio = 6

Nro de sucursales = 3

Empresa de reclutamiento de recursos humanos, con más de 6 años en el mercado y con 3 Sucursales en funcionamiento: Barcelona (Principal), Valencia,Madrid.

**PLANTEAMIENTO DEL PROBLEMA**

\*Considera que el número total de bajas con respecto al total de contrataciones es bastante alto, y concluyen que el 34% del total de sus contrataciones renuncian al termino de 2 años o antes.

\*Quieren enfocar los esfuerzos en predecir, al momento de tomar la decisión de contratar nuevos miembros, según ciertas características profesionales analizadas, si son potencialmente propensos a abandonar sus posiciones al termino de 2 años.

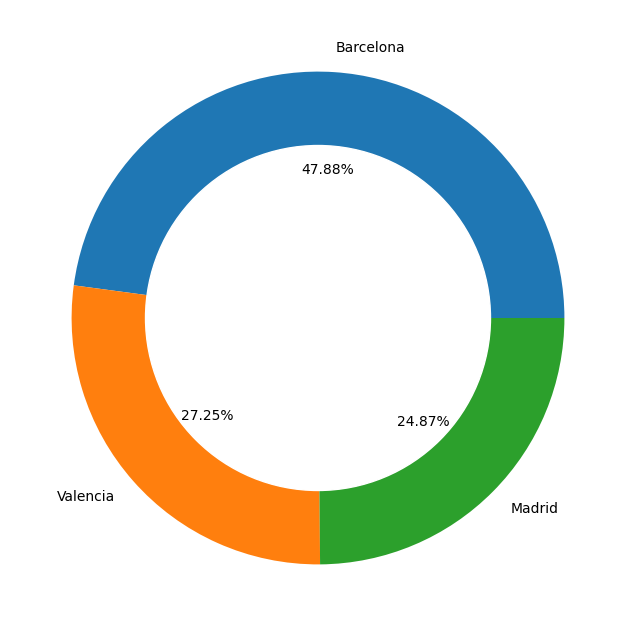
\*Estas predicciones ayudaran a tomar decisiones más acertadas a la hora de ejecutar las visiones profesionales según el perfil del individuo, así como mejorar las condiciones y estrategias de contratación.

**EDA**

**Datos Generales**

La empresa concluye que las siguientes variables son determinantes a la hora de predecir.:

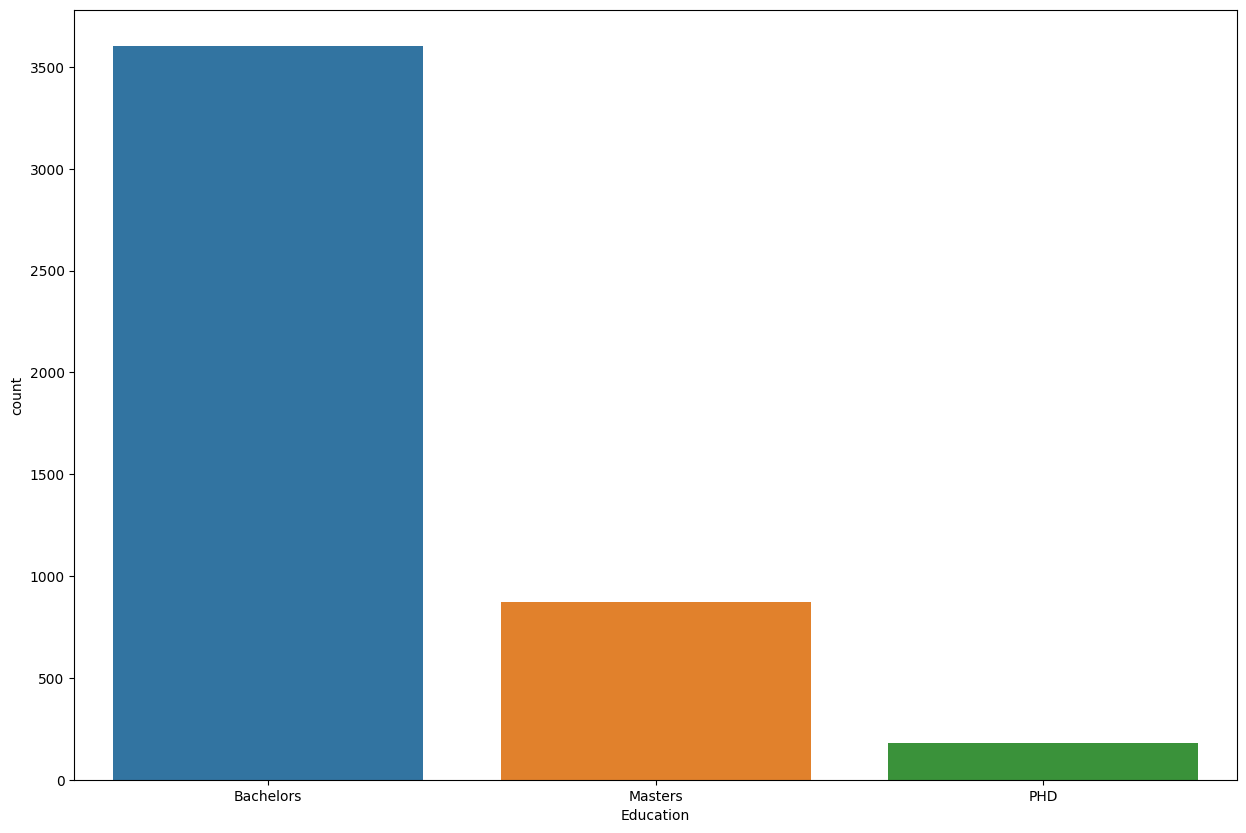
1. La empresa nos muestra la sucursal donde se producen la mayor cantidad de contrataciones:



\***Barcelon**a es la sucursal con más mayor número de contrataciones, con un **47.88%**

**\*Valencia 27.25%**

**\*Madrid 24.87%**

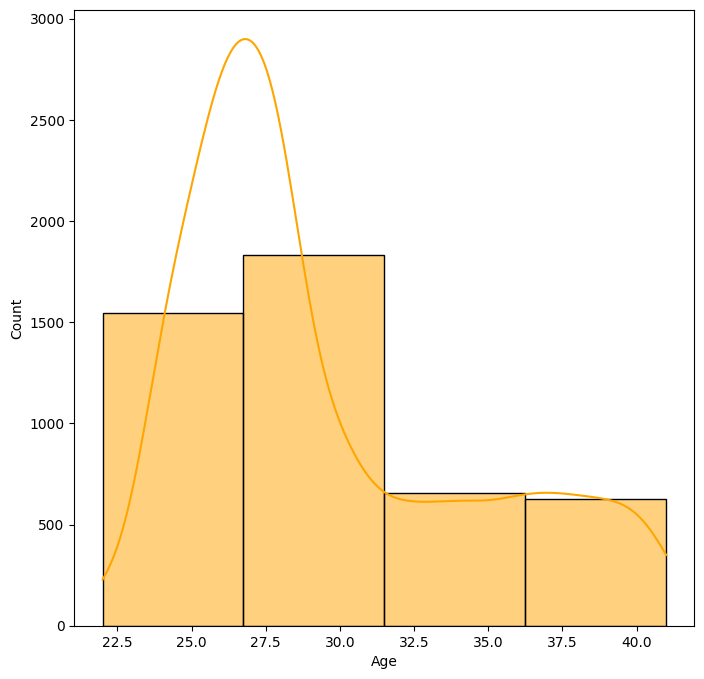
1. La empresa nos muestra el nivel de educación con más contrataciones.

Observamos el nivel:

**Bachelors** es el perfil profesional con mayor número de contrataciones, con un **77.39%.**

**Master’** **18.76%**.

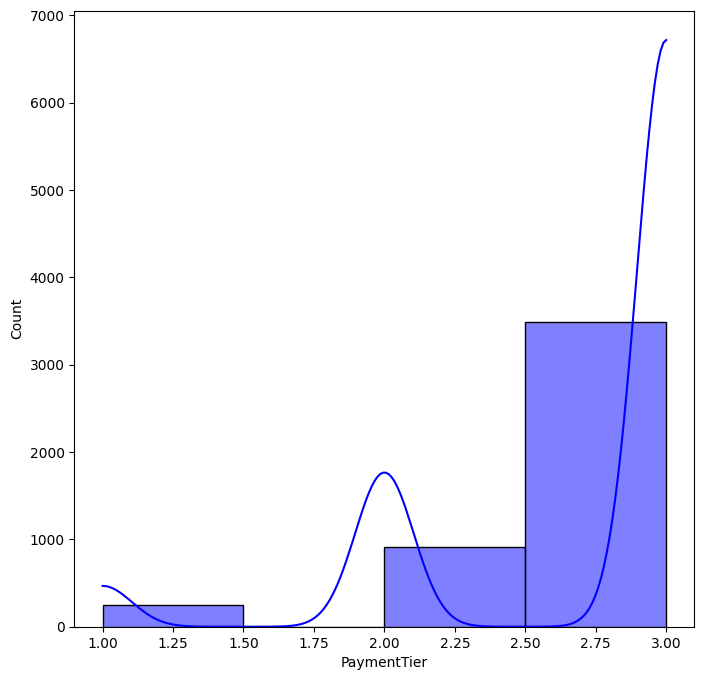
**PHD** **3.85%.**

1. Nos indican que el promedio de edad de las contrataciones es de 29 años
2. Nivel de sueldos, donde:

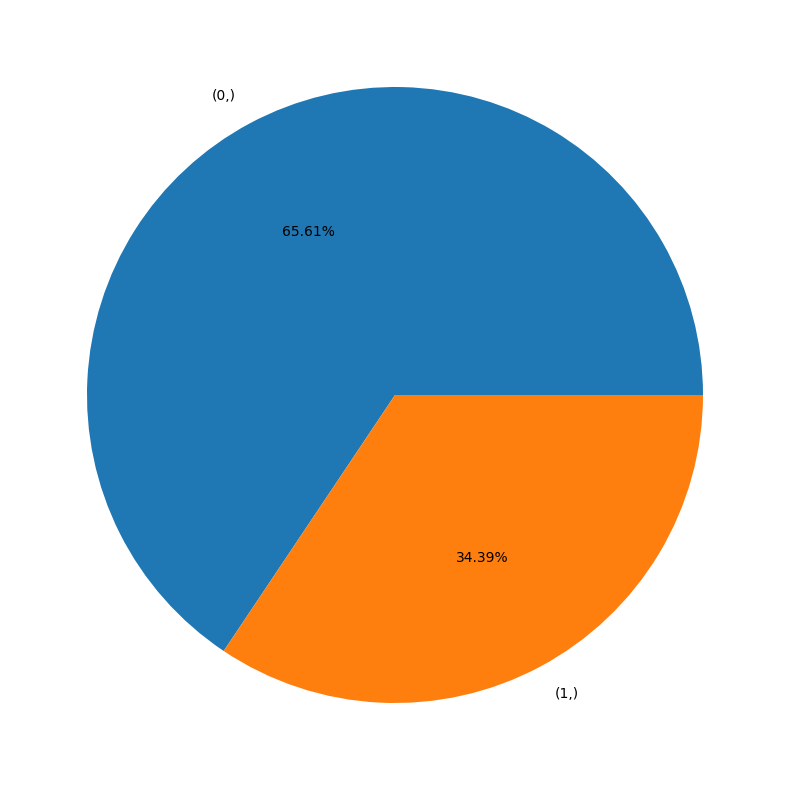
1=Alto

2=Medio

3=Bajo



1. Nos muestran que tienen un 34.39% de bajas totales con respecto al total de contrataciones.



**TRATAMIENTO Y PREPARACION DE LOS DATOS**

\*Recibimos un DataSet con 4653 registros limpios, con las siguientes columnas que representan las características generales de cada miembro contratado (Features):

1. ‘Education’-Nivel de educación profesional
2. ‘JoiningYear’-Muestra el año de contratación
3. ‘City’-Ciudad que gestiona la contratación
4. ‘PaymentTier’-Rango de sueldos
5. ‘Age’-Edad del empleado
6. ‘EverBenched’
7. ‘ExperienceInCurrentDomain’-Experiencia en el cargo
8. ‘LeaveOrNot’-Columna de decisión final Se va o No.

\* Una vez analizado los datos, exploramos que tipos de datos teníamos, si sus valores estaban dentro de una estructura homologada, y que tan relevantes eran para montar un modelo.

Procedimos a hacer los siguientes tratamientos:

1. Dividimos en Train y Test las muestras.
2. tratamos con las variables categóricas aplicando ‘getdummies’, ya que las categorías en cada una de ellas no superaban la cantidad de 3, así que no se agregarían tantas columnas a nuestros datos.
3. Verificamos y Balanceamos el número de muestras en Train respecto a nuestro Target para evitar SESGAR nuestro algoritmo desde el entrenamiento. Esto lo hicimos mediante un SMOTE() dándole relevancia a las clases minoritarias, tomando las más importantes matemáticamente.

**ELIGIENDO EL MODELO DE MACHINE LEARNING**

Para este proyecto, probaremos varios modelos **SUPERVISADOS**, tanto clasificadores como regresores, y nos quedaremos con el de mejor ‘Accuracy’ (GSCV)

¿Por qué modelo Clasificatorio?: Deseo predecir 2 clases, (Si o No)

¿Porque modelos Regresores?: Deseo predecir 2 números, (0 o 1)

Ambos casos me funcionan.

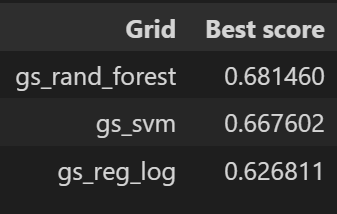
Probaremos 3 modelos en concreto.

1. Regresión logística
2. Vector Soporte Clasificador
3. Random Forest Clasificador

Para el preprocesado, en cada modelo fijamos pipeline con StandardScaler(escalar/estandarizar) y SimpleImputer(rellenar datos con mean,modar,most frequent) antes de cada modelo.

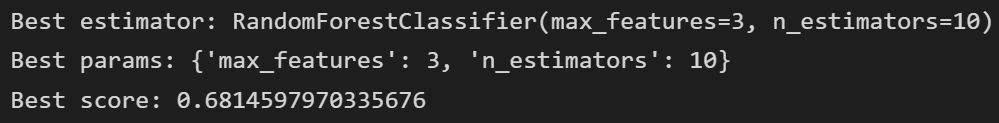
Cada uno tendrá hyperparametros que nos ayudaran a identificar el modelo que mas encaja con nuestra búsqueda de predicción.

Por último, ya con el pipeline y los hyperparametros definidos aleatoriamente, aplicamos el GSCV para cada uno de los modelos y así encontrar el mejor accuracy.



Con un 0.68 de accuracy, el modelo que mejor encaja con el proyecto del cliente es un Random Forest Clasificador.

Los mejores parámetros, estimadores y score son:



¡Gracias!