**MEMORIA TFM**

Isabel R. Rodríguez Robledo

# Introducción - ¿Qué es People Analytics?

**People Analytics** es un método de investigación basado en datos cuyo objetivo es estudiar a las personas que forman parte de una empresa, de tal modo que con los datos extraídos y un análisis inteligente de ellos se puede establecer conclusiones objetivas, válidas y fiables.

Se definen People Analytics como una**identificación y cuantificación sistemática de las personas que impulsan los resultados de un negocio con el objetivo de tomar mejores decisiones**. La finalidad última de People Analytics no es recoger datos, es ayudar a tomar mejores decisiones.

Cuando hablamos de **People Analytics en realidad nos estamos refiriendo a Big Data y Business Inteligence**aplicado al área de Recursos Humanos.

### Usos de People Analytics

Como hemos visto, el objetivo principal de HR Analytics es **ayudar a managers y otros directivos a gestionar mejor a los empleados** y tomar aquellas decisiones adecuadas que permitan impulsar su bienestar y, por lo tanto, el éxito de la empresa.

Dentro del departamento de Recursos Humanos, los usos de esta novedosa tecnología son tan variados cómo las preguntas que queramos realizar o los aspectos que queramos investigar de todo aquello relacionado con los empleados.

Aunque realmente el límite es casi infinito, Se podría definir en 5 grandes áreas: **reclutamiento, desarrollo, lealtad, compensación y organización.**

**En este Dataset vamos a realizar un estudio de lealtad de los trabajadores en una empresa.**

### Lealtad

People Analytics recoge datos sobre el estado de los empleados a tiempo real y permite identificar signos de desmotivación e intenciones de los empleados de dejar la empresa.

Esta monitorización al día ofrece la oportunidad de realizar acciones de retención del talento de forma preventiva antes de que sea demasiado tarde.

### Beneficios de People Analytics

Es posible que te preguntes qué ventajas tiene establecer un sistema de HR Analytics en una organización. Más allá de toda la utilidad que hemos comentado anteriormente, la implementación de esta metodología aporta grandes beneficios tales como:

1. Mejora la gestión de las personas a partir de datos reales y no de suposiciones.
2. Se puede hacer un seguimiento de las decisiones estratégicas que se toman para saber si funcionan o no y corregir a tiempo.
3. Permite descubrir áreas de mejora y crecimiento, potenciando el desarrollo de la empresa.
4. Permite dedicar más tiempo a las tareas importantes y de mayor valor, y no tanto a gestiones administrativas.
5. Reduce costes de tiempo y dinero.
6. Mejora el bienestar y satisfacción de los empleados, contribuyendo a mejorar los índices de productividad.

En este estudio vamos a ayudar a la empresa a determinar **qué candidatos** encajan mejor con la cultura de la empresa, identificando a los**patrones de rotación de personal**utilizando algoritmos de “machine learning” y ofreciendo la posibilidad de identificar qué empleados tienen mayor probabilidad de abandonar la empresa porque busquen un nuevo trabajo.

2 - Introducción al Estudio de mi TFM- Predecir la probabilidad de que un candidato busque un nuevo trabajo.

Una empresa activa en Big Data y Data Science quiere contratar científicos de datos

entre las personas que superan con éxito algunos cursos que imparte la empresa. Mucha gente se apunta a su formación. La empresa quiere saber cuáles de estos candidatos

realmente quieren trabajar para la empresa después de la formación o buscaran un

nuevo empleo porque ésto ayuda a reducir  costes y  tiempo, así como la calidad de la

formación o la planificación de los cursos y la categorización de los candidatos. La

información relacionada con la demografía, la educación y la experiencia está en manos de Recursos Humanos y la inscripción de los candidatos.

Este conjunto de datos está diseñado para comprender los factores que llevan a una

persona a dejar su trabajo actual para dedicarse también a investigaciones de recursos humanos. Por modelo (s) que utiliza las credenciales actuales, datos demográficos y

datos de experiencia, podrá predecir la probabilidad de que un candidato busque un

nuevo trabajo o trabaje para la empresa, además de interpretar los factores afectados en la decisión del empleado.

\* Nos encontramos ante un problema de Clasificación.

\* Desafíos de este Conjunto de datos:

  \* El conjunto de datos está desbalanceado.

  \* La mayoría de las características son categóricas (nominal, ordinal, binaria).

  \* La imputación faltante.

**\*\*Objetivo\*\***

Predecir la probabilidad de que un candidato trabaje para la empresa, despues de la formación o busque un nuevo trabajo.

Interpretar el (los) modelo (s) de tal manera que ilustre qué características afectan la decisión del candidato.

**Features**

* enrollee\_id : ID único del candidato.
* city: código de ciudad.
* city\_ development \_index : índice de desarrollo de la ciudad (escalado).
* gender: género del candidato.
* relevent\_experience: experiencia relevante del candidato.
* enrolled\_university: tipo de curso universitario inscrito, si lo hubiera.
* education\_level: nivel de educación del candidato.
* major\_discipline: disciplina principal de educación del candidato.
* experience: Experiencia total del candidato en años.
* company\_size: Número de empleados en la empresa del empleador actual
* company\_type : tipo de empleador actual.
* lastnewjob: diferencia en años entre el trabajo anterior y el trabajo actual,es decir, cuantos años lleva en esta empresa.
* training\_hours: horas de formación completadas.
* target: 0 - No busca un cambio de trabajo, 1 - Busca un cambio de trabajo.

Dataset tiene 19158 Filas

Dataset tiene 13 Columnas

# 3- TARGET

**Objetivo: probabilidad de que un afiliado busque un cambio de trabajo**

Target:

0 - No busco cambio de trabajo,

1 - Buscando un cambio de trabajo.

Tenga en cuenta que el conjunto de datos no está equilibrado en la variable. Para corregir este desbalance haremos un undersampling, ya que necesitamos corregir target =1.

El 75% de los inscritos no parecía un cambio de trabajo

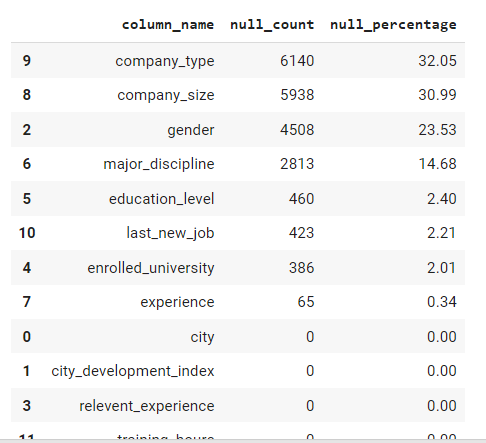
25% de los inscritos buscan un cambio de trabajo

Estamos ante un problema de Clasificación, ya que el target del problema es un conjunto de valores discretos binarios ( el empleado no busca/ Busca un nuevo empleo).

# 4- DETECCIÓN DE ANOMALÍAS.

# Missing values

8 de las variables tiene valores nulos. Trataremos cada variable una a una a continuación.



# Outliers

¿Qué es un outlier?

Se trata de un valor atípico dentro de nuestros datos. Un valor que se desvía mucho de las métricas estadísticas de centralidad (media, moda, mediana). En el análisis exploratorio de datos suele ser algo a estudiar, el por qué tengo algunos valores atípicos en los datos. No obstante, en machine learning los outliers en los datos implican penalizaciones en los modelos, sobretodo los que trabajan con distancias y con Gradient Descent. Por tanto, hay que lidiar con ellos.

¿Qué gráficas utilizo para visualizar los outliers?

Lo mejor son boxplots, histogramas, diagramas de densidad, scatter plots y count plots para categóricas.

¿Qué técnicas hay para detectar outliers?

1. **\*\*Gráficamente\*\***: Datos que se desvien mucho.

2. **\*\*Cuartiles\*\***: se ve en un diagrama de caja. Datos que caigan fuera del rango +/- 1.5\*IQR. Este 1.5 es muy restrictivo por lo que se suelen probar valores del 3 al 5. Dependerá mucho de cada feature.

3. **\*\*Desviación estándar\*\***: todo lo que caiga fuera de (media +/- N\*std) de la variable. Normalmente N es un valor de 3 a 5.

 ¿Qué hago con los outliers?

Tenemos varias opciones:

1. **\*\*Eliminarlos\*\***: es la técnica más habitual y sencilla

2. **\*\*No hacer nada\*\***: si no son exagerados. Los árboles de decisión y SVM (en este orden) son robustos frente a outliers.

3. **\*\*Transformaciones logarítmicas\*\***: elimina asimetría en las features, y por tanto reduce el efecto de los outliers.

4. **\*\*Binning\*\***: discretiza la variable en varios grupos. Esto me va a permitir incluir los outliers en un grupo donde haya otros datos no considerados como outiers (ver ejemplo abajo)

5. **\*\*Imputación\*\***: igual que con los missings, sustituir los valores. Esto tendrá sentido si los outliers están bien analizados, y desde el punto de vista de negocio conviene sustituirlos por un valor concreto.

6. **\*\*Tratamiento por separado\*\***: si es un número significativo de outliers quizá merezca la pena separar los datos y tratarlos como otro modelo aparte.

# Nuestras dos únicas variables numéricas hasta ahora cuentas con outliers. Simplemente he realizado una imputación a éstos.

* training\_hours: horas de formación completadas.
* city\_ development \_index : índice de desarrollo de la ciudad (escalado).

# 5- ANÁLISIS Y LIMPIEZA DE DATOS. Variable por variable.

* enrollee\_id : ID único del candidato.

Esta variable simplemente la eliminamos. Aparecen algunos valores nulos.

* city: código de ciudad.

Como es una columna que tiene demasiados valores (123) y no tiene información relevante, la eliminamos. Como la siguiente columna está relacionada con la geografía y nos aporta más valor, nos quedamos con la siguiente.

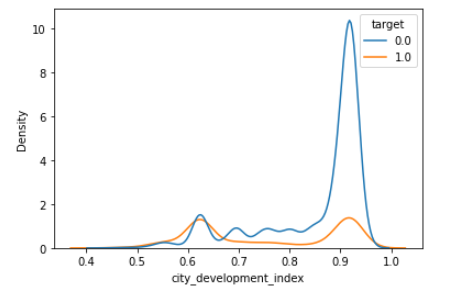
* city\_ development \_index : índice de desarrollo de la ciudad (escalado).

Comprobamos que el mayor numero de poblaciones tienes alrededor de un 90% de

Índice de desarrollo. Podemos ver que las personas que viven en ciudades desarrolladas tienen una menor

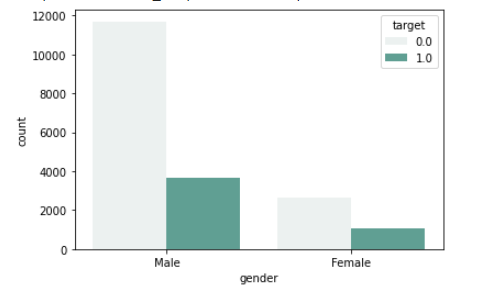
Probabilidad de cambiar de trabajo. Podemos ver que más personas de ciudades con un índice de desarrollo bajo están

Buscando un cambio de trabajo porque  están menos satisfechas con su trabajo actual o con el lugar en el que viven.

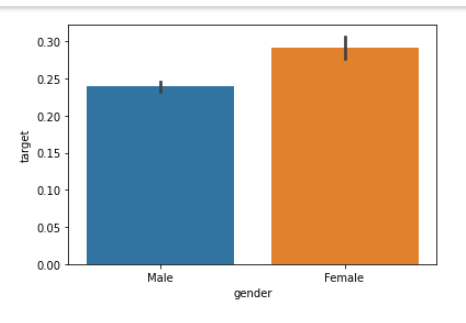


* gender: género del candidato.

En esta variable había bastantes valores nulos. He realizado la imputación de estos con una imputación aleatoria 50 % mujeres 50% hombres.



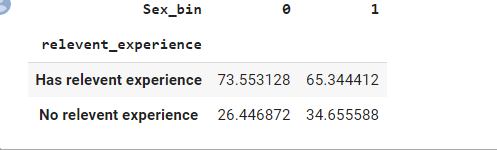
Mayor porcentaje de mujeres están buscando un nuevo trabajo.



* relevent\_experience: experiencia relevante del candidato.

Conclusión: La mayoria de personas que tiene experiencia relevante, NO  buscan un trabajo nuevo. Probablemente porque estarán más incentivadas por la empresa para retener este talento, ya sea a través de la remuneración, la formación o por mejores condiciones laborales.

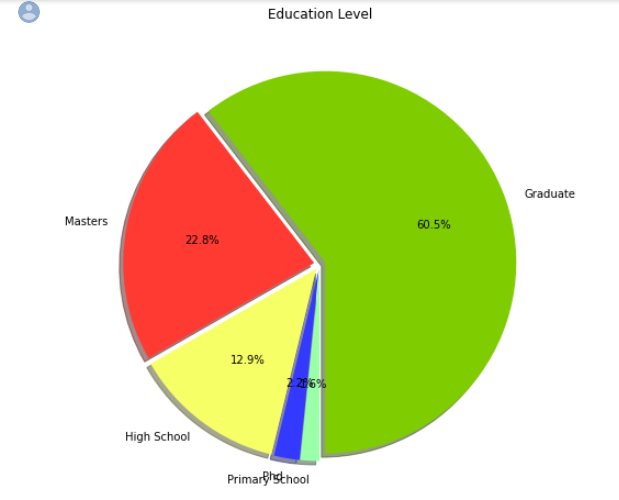
El número de candidatos masculinos tienen 73.5% de experiencia relevante y en el caso de las mujeres sólo un 65.3% cuenta con experiencia relevante.



* education\_level: nivel de educación del candidato.

Comprobamos que el mayor número de trabajadores poseen un Grado Universitario. El mayor número de personas buscando un nuevo empleo son Graduadas.

Los pocos valores nulos de esta variable (2.4%) los he imputado con “High School”, Uno de los valores minoritarios, para desbalancear.

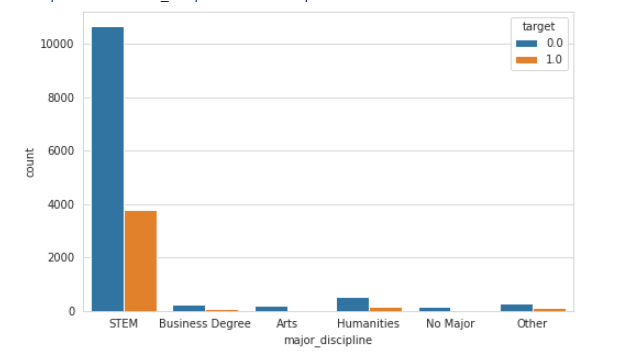


* enrolled\_university: tipo de curso universitario inscrito, si lo hubiera.

No le encuentro sentido a esta variable, ya que el mayor numero de empleados era graduados y aqui aparece  como no apuntandos a la universidad, asi que elimino esta columna. Por lo tanto también estoy eliminando los pocos valores nulos.

* major\_discipline: disciplina principal de educación del candidato.

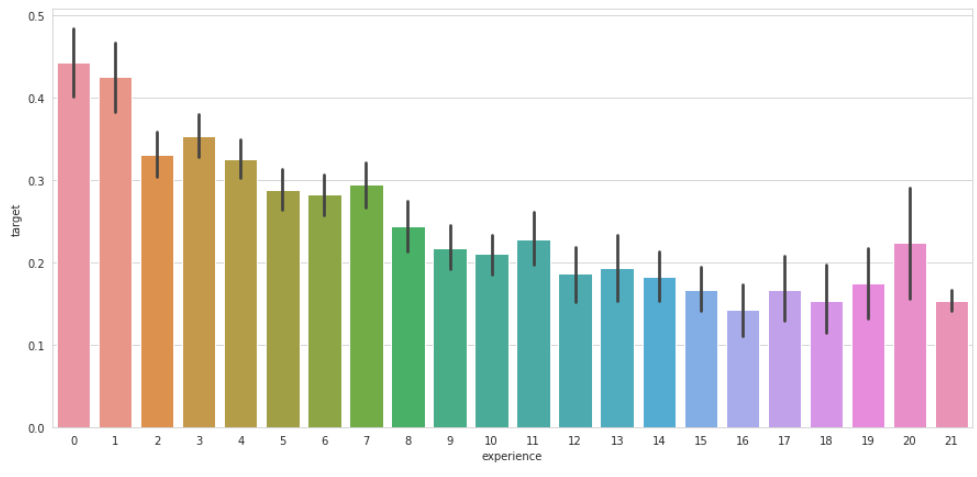
El mayor número de trabajadores de esta empresa tiene estudios de la rama de Ciencias. Elimino esta columna porque no me aporta demasiado valor. Realice comprobaciones posteriores y por eso tome esta decisión. Por eso no hago nada con el 14.68% de los valores nulos.



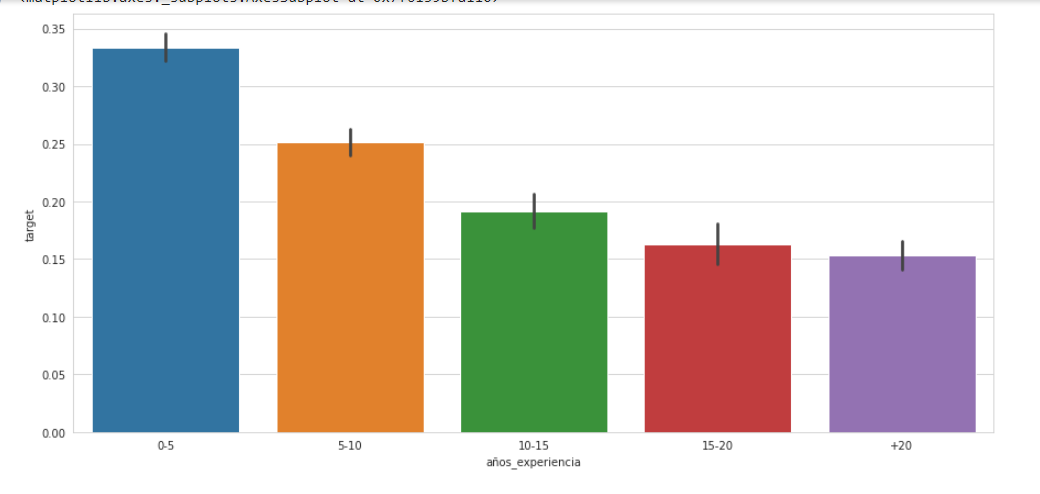
* experience: Experiencia total del candidato en años.

Esta variable contaba con pocos valores nulos, simplemente le he imputado el valor que más se repetía. Los años de experiencia van desde 0 a +20 años .

Comparando con la variable objetivo podemos ver la tendencia de que a mayor años de experiencia menor es la probabilidad de buscar un nuevo trabajo. Los trabajadores se acomodan en su empresa por diferentes motivos.



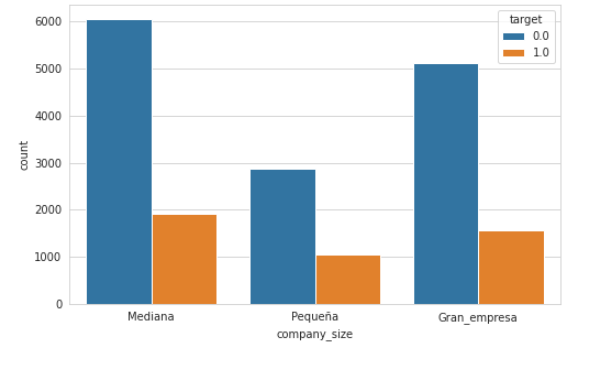
Realice una imputación a los valores agrupándolos la experiencia de 5 en 5 años para reducir los valores.



* company\_size: Número de empleados en la empresa del empleador actual

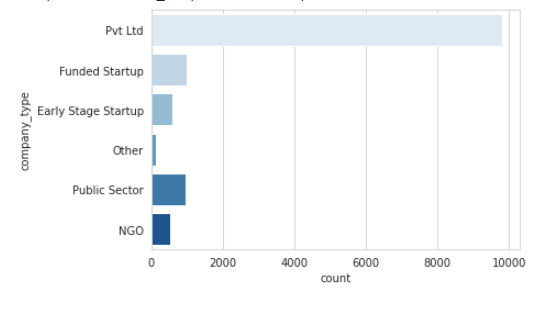
Esta variable contada con demasiados valores nulos, realice la imputación de los mismos con los diferentes métodos de “ffill” y “bfill”. Copia el valor de la celda de delante o la de detrás.

Como esta variable tiene muchos valores agrupe el tamaño de la empresa en Pequeña, mediana grande. Podemos observar la proporción en el siguiente empresa correspondiendo a que trabajadores buscan o no un nuevo empleo dentro de las diferentes empresas.



* company\_type: tipo de empleador actual.

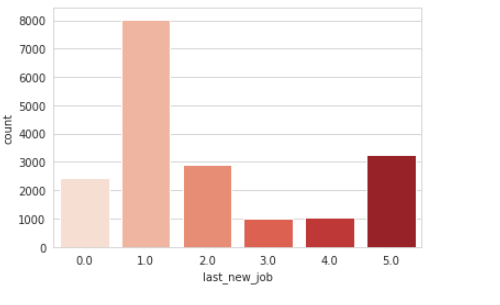
En esta variable podemos visualizar el tipo de empresa. Vemos que la mayoría son Sociedades de responsabilidad limitada. Pero como esta variable cuanta con un 32% de valores nulos, he decidido eliminarla.



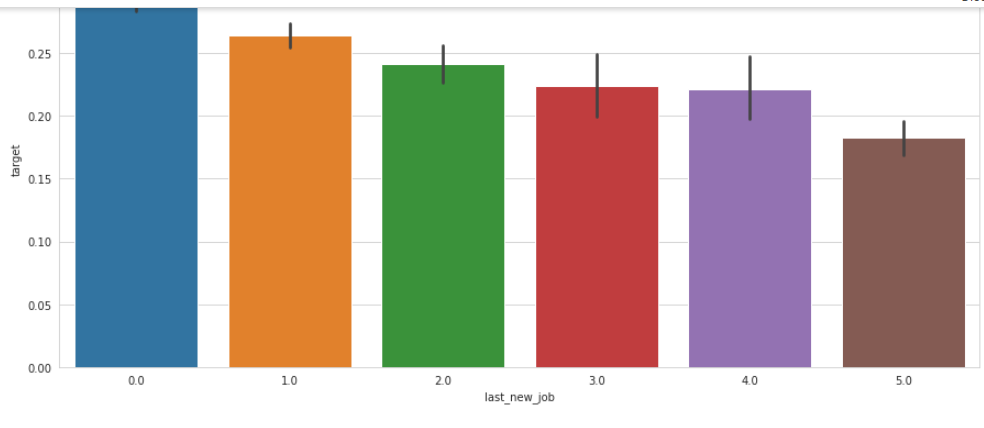
* Last\_new\_job: diferencia en años entre el trabajo anterior y el trabajo actuales decir, cuántos años lleva en esta empresa.

Realizo la imputación de los missing values(2.21%) con uno de los valores minoritarios para descompensar el desajuste de los datos.

La mayoría de trabajadores formados tienen alrededor de un año de experiencia.



A más años de experiencia en la empresa, menor es el ratio de abandono.



* training\_hours: horas de formación completadas.

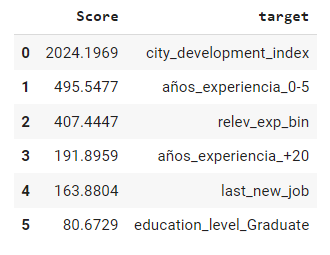
La media de la formación se encuentra en 50 horas por trabajador. A mayor formación menor ratio de abandono. Mayor satisfacción de los empleados.

* target: 0 - No busca un cambio de trabajo, 1 - Busca un cambio de trabajo.

# 6- PREPROCESAMIENTO DE LOS DATOS.

1. Estandarizamos las variables numéricas (‘training\_hours’ y ‘last\_new\_job’) con el método preprocessing de la librería Sklearn. Estandarizando nos asegurarnos que implementaremos la uniformidad necesaria para mantener la eficacia de sus análisis.
2. Binarizamos las variables No numéricas por el proceso de get dummies.
3. Estudiamos la correlación de las variables con la variable objetivo a través de una matriz de correlación.
4. Para no sobre cargar el modelo de ML, seleccionamos solo las mejores variables con el método SELECTKBEST de la librería de sklearn.

Este fue el resultado. Luego están mis variables independientes para el modelo .



# 7- MACHINE LEARNING.

Antes que nada, he realizado varias versiones del modelo:

1. Sin undersampling:
   1. Sin hyperparámetros
   2. Con hyperparámetros.
2. Con Undersampling:
   1. Sin hyperparámetros
   2. Con hyperparámetros.

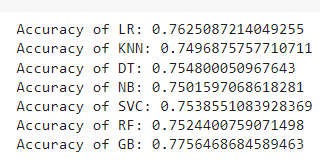
Empezamos dividiendo los datos en datos de entrenamiento y datos de prueba (80%, 20%) (X\_train, X\_test, y\_train, y\_test)

Entrenamos los datos con diferentes algoritmos a través del modelo de **validación cruzada**.

***Validación cruzada*** es un procedimiento de remuestreo que se utiliza para evaluar modelos de aprendizaje automático en una muestra de datos limitada. ... Es decir, usar una muestra limitada para estimar cómo se espera que  funcione el modelo en general cuando se usa para hacer predicciones sobre datos no usados ​​durante el entrenamiento del modelo.

La validación cruzada es una técnica muy útil para evaluar la eficacia de su modelo, especialmente en los casos en los que necesita mitigar el sobreajuste.

Elegimos el modelo más preciso a través del método **Accuracy**: La exactitud es una métrica que nos indica cómo de preciso es un modelo de Machine Learning a la hora de hacer predicciones. Formalmente, se define como el ratio entre las predicciones correctas respecto al número total de ejemplos. Cuanta más alta sea la exactitud, es decir más cercano a 1, más acertará nuestro modelo y, por lo tanto, será más preciso.



Estos fueron mis resultados. Por lo tanto, para no sobrecargar el notebook me quedo con el algoritmo GradientBoostingClassifier, porque tiene la exactitud más alta. Entrenamos el modelo GB.

También calculamos la Curva Roc y su Area bajo la curva y la medida de F1 para ver cuánto de preciso es nuestro modelo.

\* Curva Roc: Nos dice qué tan bueno puede distinguir el modelo entre dos cosas. Mejores modelos pueden distinguir con precisión entre los dos, mientras que un modelo pobre tendrá dificultades para distinguir entre los dos.

An ROC curve (receiver operating characteristic curve) is a graph showing the

performance of a classification model at all classification thresholds. This curve plots

two

parameters: True Positive Rate. False Positive Rate.

Es el area bajo la curva ROC: Una situación ideal es cuando la curva ROC es igual a 1, es

Decir, cuando las dos curvas no se superponen en absoluto.

La peor situación es cuando AUC es igual a 0.5, es decir,el modelo no tiene capacidad

Para distinguir entre clase positiva y clase negativa.

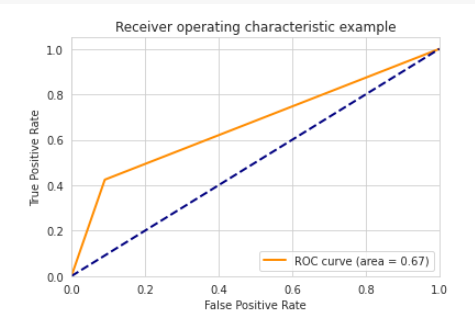
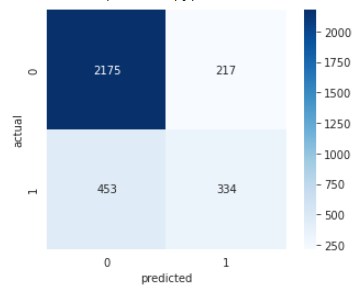
Cuando AUC es aproximadamente 0, el modelo en realidad está correspondiendo las

. Significa que el modelo predice la clase negativa como positiva y viceversa.

Vuelvo a realizar los mismos pasos esta vez con los hyperparámetro de GradientBoostingClassifier. Observamos que los resultados casi no cambian.

Comprobamos que tan bien predice nuestro modelo con la matriz de confusión.

Visualizamos también el Área debajo de la curva.



Volvemos a realizar los mismos pasos pero esta vez con el Método Undersampling de la librería Sklearn.

# 8- CONCLUSIÓN

Sorprendentemente, estos fueron mis resultados. Después de hacer varias versiones, los resultados casi no variaron. Mi mejor predicción fue sin hacer undersampling y sin hacer tuning.

