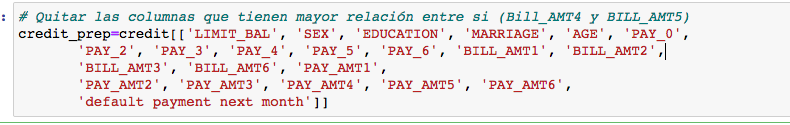
# Introducción

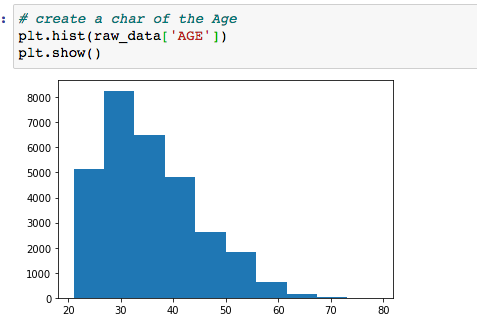
El presente documento corresponde al informe de lecciones aprendidas sobre el análisis de datos con Python y la preparación de la data correspondiente a los datos de default de clientes de tarjetas de crédito.

# Anális Realizado

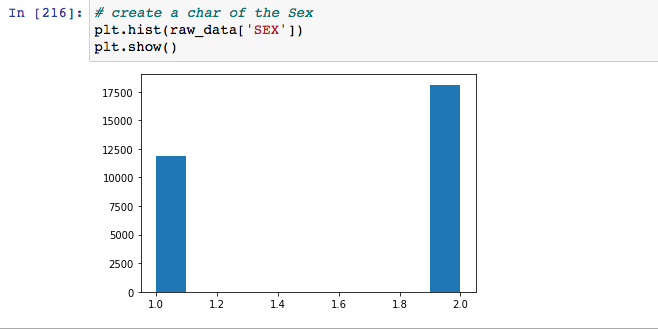
1. En trabajos anteriores se había trabajados sobre este mismo set de datos, por lo que con base en el análisis se determinó cuáles columnas tenían una alta correlación entre si y no tanto con la variable a evaluar "default payment next month". Con base en eso se decidió eliminar las columnas 'PAY\_AMT4' y 'PAY\_AMT5', dejando nuestro dataset con las siguientes columnas para evaluar



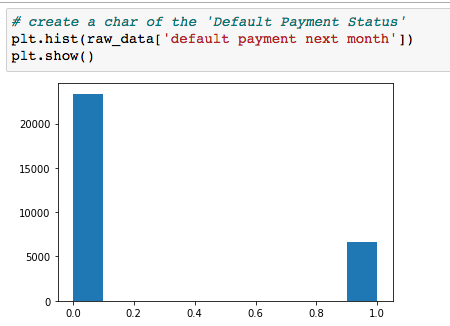
1. Una vez agregado las columnas deseadas se realizó el análisis con la data restante, se evalúa o se grafican algunas de las variables para medir el comportamiento de los datos, por ejemplo la distrubución por Edad, Sexo, Estado Civil



Se observa que la mayoría de los clientes están entre las edades de 20 a 50 años



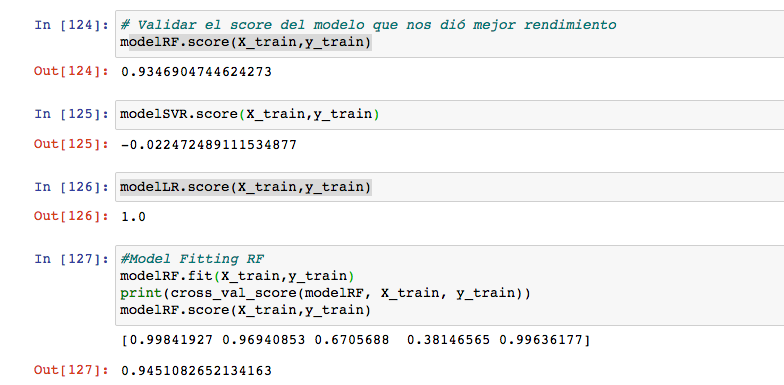
Se observa que existe una mayor cantidad de deudores hombres que mujeres



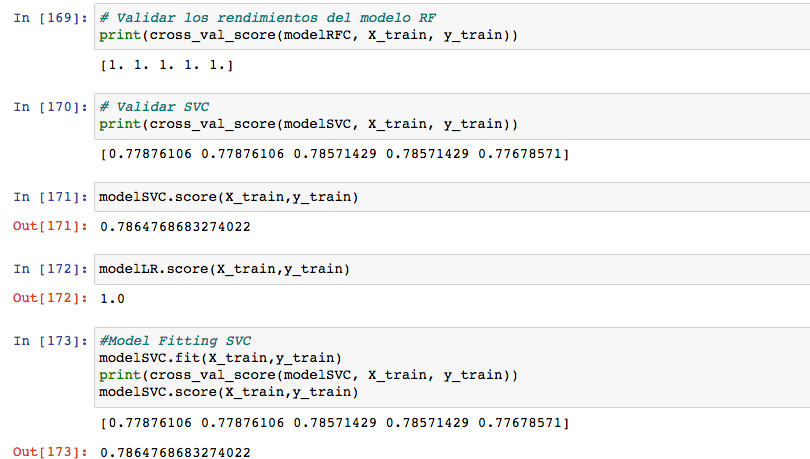
Alrededor del 20% de los clientes son los que entran en default para el próximo mes.

1. Se realizaron y aplicaron los diferentes modelos, primero para una variable no discreta usando los métodos de Regresión y posteriormente con la variable que deseamos evaluar y se determina que los modelos que funcionan para un tipo no necesariamente son los mismos que operan para otro, aunque se utilicen los mismos datos, así por ejemplo para los modelos de regresión se usó la variable dependiente "PAY\_AMT6", para lo cual el modelo con el mejor rendimiento fue el Random Forest mientras que para la variable 'Default Payment' con los modelos de clasificación el modelo con mejor dendimiento fue SVC

**Rendimiento de los modelos de Regresión**



**Rendimiento de los modelos de Clasificación**



# Conclusiones

1. Una vez realizado el análisis de los datos de los clientes se determinó que existen muchas variables que no parecen tener un efecto real sobre si el cliente puede o no entrar en default durante el tiempo de los créditos.
2. El modelado que vamos a utilizar depende de cada tipo de variable que vayamos a evaluar.
3. Existe una gran cantidad de características que si bien es cierto necesitamos conocer no necesariamente nos brindarán información sobre la posibilidad de un cliente de caer en default, ya que muchas de dichas variables no las podemos controlar.
4. Existen otra gran cantidad de modelos diferentes de los que hemos usado por y que podrían generar mejores rendimientos o más información.