Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

PROYECTO

TRABAJO FINAL DE ESPECIALIZACIÓN

Riesgo Crediticio. Aplicación de Métodos Predictivos y Análisis Multivariado.

AUTOR: ERIKA BEATRIZ MUÑOZ

[DICIEMBRE 2019]

Resumen

El presente trabajo, se desarrolla con el fin de determinar una metodología eficiente que pueda realizar una evaluación crediticia contemplando diversas características de la persona en cuestión, con lo cual se pueda determinar si incurrirá en un incumplimiento del pago de sus créditos; buscando además encontrar variables explicativas o bien una correlación entre ellas.

Para ese fin, se realizó una investigación tanto de los conceptos y el marco teóricos que abarca el riesgo crediticio, como también la aplicación de técnicas de minería de datos y machine learning o aprendizaje automático, con el fin de evaluar distintas características de las personas que nos den como resultado una probabilidad de incumplimiento, y que esto pueda contribuir a que las entidades financieras puedan reducir la tasa de morosidad, la cual es muy importante a la hora de asegurar la liquidez y futuro de dicha institución.

Estructura

Introducción

La gestión del riesgo crediticio es una de las decisiones más importantes que deben tener en cuenta las entidades financieras, ya que estas decisiones afectarán su liquidez y sus resultados a corto y mediano plazo. Por esta razón, es relevante que las entidades tengan un buen sistema de control que determine la consistencia de los sistemas tanto internos como externos en los modelos que se apliquen a sus efectos.

Para establecer el "riesgo crediticio" se analizará al prestatario, y su condición para hacer frente a las obligaciones; esto implica, un análisis riguroso del flujo de fondos futuros y las obligaciones preexistentes que demostrarán la probabilidad de que el mismo incumpla con las obligaciones contraídas.

El objetivo de este trabajo es poder establecer una metodología que permita determinar tanto una relación entre las variables que se analizan como una probabilidad de incumplimiento por parte del tomador de crédito; para esto, se utilizarán distintas técnicas de minería de datos y aprendizaje automático, y serán analizadas distintas variables, ya sea cuantitativas como cualitativas y no sólo el flujo de fondos futuro.

Mediante este análisis, se podrá determinar el perfil del cliente, ya sea para reducir la morosidad en las entidades financieras con un análisis riguroso de determinadas características, como para establecer un patrón y poder determinar a qué grupo de personas con similares características pertenece y posteriormente poder ofrecer distintos productos en base a eso; sin embargo, este último análisis más enfocado al marketing.

En el presente trabajo, el primer capítulo se establecen los antecedentes del riesgo crediticio, cómo surgió, su relación con el análisis estadístico y la importancia de la aplicación de distintas técnicas de minería de datos en la era del Big Data.

En el segundo capítulo, se realizará un análisis de los datos, aplicando distintas técnicas, para determinar el perfil de los clientes en nuestra base de datos, cómo podrían agruparse y la relación entre las distintas variables. Por último, por medio del análisis multivariado se buscará establecer las variables que sean más determinantes en las conclusiones y poder eliminar aquellas que nos brindan información redundante o no relevante para la problemática en cuestión, ya que este análisis se centra en la investigación simultanea de dos o más variables y es sumamente importante a la hora de determinar qué variables deben tenerse en cuenta y cuáles no, al momento de analizar la información. Cabe aclarar, que con este último método lo que se busca además de reducir la cantidad de variables es perder la menor cantidad de información posible.

El tercer capítulo, tendrá como objetivo aplicar modelos predictivos, mediante técnicas estadísticas como Random Forest, Regresión Logística y Decision Tree para determinar una metodología que luego deberá ser analizada al momento de establecer el otorgamiento del crédito. En este proceso, se determinará además la estimación del modelo y la performance que obtendremos con cada técnica.

Finalmente, a raíz de los resultados obtenidos y las técnicas aplicadas se expondrán las conclusiones que se han obtenido a lo largo del trabajo.

1. El riesgo crediticio y su marco teórico

El riesgo de crédito es la posibilidad de una pérdida económica debida al incumplimiento del prestatario en las obligaciones asumidas por el mismo mediante contrato, esto implicará además de la pérdida de liquidez en el caso de que la tasa de morosos sea alta, también un costo administrativo para el posterior recupero de la misma.

Cabe mencionar que el riesgo crediticio no es sólo utilizado por bancos o entidades financieras, sino también por Compañías de Seguro, las cuales utilizan metodologías similares en cuanto al análisis de riesgo, pero teniendo en cuenta los siniestros; sin embargo, son distintas las maneras de encarar las pérdidas, ya que en las compañías de seguro se trata de erogaciones relacionadas a los siniestros y las entidades financieras harán frente a las mismas mediante capital o reservas previamente calculadas.

1.1 Los antecedentes del Riesgo Crediticio

El origen de las calificaciones crediticias se remonta a 1860 de la mano de Henry Varnum Poor que, con el libro "Historia de Ferrocarriles y canales en los Estados Unidos", publica información completa sobre el estado financiero de compañías financieras estadounidenses. En 1906, Luther Lee Blake realiza publicaciones de empresas no relacionadas con ferrocarriles, para luego unir su compañía "Standard Statistics Bureau" con la de Henry Poor, formando Standard & Poor´s Corp. considerada una de las tres grandes agencias de calificación de crédito ubicada en EEUU, Nueva York.

Se entiende por crédito, la inversión que es realizada a un cliente, ligada a la venta de un producto o servicio, y la principal razón de otorgar crédito, es generar e incrementar las ventas. (Ross, Westerfield & Jaffe, 2012).

Según Puertas & Martí en "Análisis de Credit Scoring" se denomina Credit Scoring a "todo sistema de evaluación crediticia que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito. Riesgo que estará en función de la solvencia del deudor, del tipo de crédito, de los plazos, y de otras características propias del cliente y de la operación, que van a definir cada observación, es decir, cada solicitud de crédito. Únicamente no existirá riesgo en una operación de crédito cuando la entidad que los instrumenta actúe como mediadora o intermediaria, o bien cuando el crédito se conceda con la garantía del Estado".

Para las entidades financieras es de suma importancia poder reducir el riesgo, ya que está en juego su patrimonio, y hasta hay registros de bancos que están pasando una delicada situación debido a la creciente tasa de morosidad; por lo que es imperante encontrar un sistema automatizado que permita gestionar eficientemente la concesión de creditos, con el que se aseguren que el cliente podrá hacer frente a las obligaciones crediticias (Puertas & Martí, 2013).

Los créditos incumplidos al vencimiento no sólo generarán costos financieros para las entidades, sino también administrativos, por la gestión que deberá realizarse para su posterior recupero (Puertas & Martí, 2013).

Se mantendrá un activo riesgoso, sólo si su rendimiento esperado compensa su riesgo (Ross, Westerfield & Jaffe, 2012). Pero ¿cómo saber si ese riesgo no es en vano, y nos encontramos con un incobrable?

Según el BCRA, "es de fundamental importancia que las entidades financieras cuenten con un proceso interno, integrado y global, para evaluar la suficiencia de su capital económico en función de su perfil de riesgo ("Internal Capital Adequacy Assesment Process" - "ICAAP") y con una estrategia para mantener sus niveles de capital a lo largo del tiempo. Si como resultado de este proceso interno se determina que el capital regulatorio es insuficiente, las entidades financieras deberán incrementarlo en base a sus propias estimaciones".

La clasificación de las personas determinando sólo ciertos aspectos, como un evento circunstancial, sin considerar otros aspectos, en cuanto a el riesgo crediticio, hacen que muchas personas se encuentren sin la posibilidad de tener una vivienda propia, un respaldo económico o de cumplir el sueño de concretar un viaje. Por esa razón, se busca encontrar otros patrones o características que puedan ser consideradas a la hora de determinar si es una persona es apta para un crédito, logrando mirar más ampliamente el contexto o la situación más global de la persona, y pudiendo quizás con eso encontrar otros métodos de análisis.

En este estudio se busca conocer: ¿Cuáles son los variables determinantes en el riesgo crediticio y la relación que existe entre ellas?

1.2 Análisis crediticio tradicional y la incorporación de la estadística

Al momento de evaluar una solicitud de crédito, se analizan distintos aspectos, ya sea de la persona como del tipo de crédito que se pretende obtener.

Existe un modelo, llamado “El modelo de las cuatro C”, el cual explica los distintos puntos a analizar, el mismo involucra:

Capacidad: En este aspecto se evaluará la capacidad del individuo para generar ingresos que permitan en un futuro hacer frente a las obligaciones. Se evaluará su flujo de caja, operaciones, inversiones, financiaciones y su capacidad de pago tanto a corto como a largo plazo.

Lo Colateral: Se refiere a los elementos que el individuo o empresa dispone para garantizar o no la obligación que está contrayendo, en este caso el crédito podría ser otorgada con o sin garantía, por ejemplo.

Condiciones: En este caso, se buscará con un acuerdo o contrato determinar las restricciones y limitaciones del mismo; dentro de éstas podrá haber condiciones positivas,

como la obligación de pagar, como también podrá haber condiciones negativas o prohibiciones que determinen por ejemplo al incremento de la deuda.

Carácter: En este concepto, se evaluará el comportamiento pasado y presente del individuo frente a deudas u obligaciones contraídas, estableciendo una tendencia o una probabilidad de actuación similar frente a obligaciones futuras.

La incorporación de la estadística al análisis crediticio

El análisis crediticio en sus principios era realizado, por ejecutivos financieros quienes medían los riesgos basándose en técnicas del negocio, teniendo en cuenta aspectos económico-financieros. Sin embargo, con los años estas técnicas se fueron perfeccionando e incorporando nuevos conceptos y mediciones más precisas con la estadística.

El Credit Scoring, plantea una probabilidad medida con técnicas estadísticas que intentan predecir o estimar el riesgo de que el cliente llegue a la instancia de incumplimiento de la obligación.

Esto no fue así sino hasta el año 1936, cuando el estadístico y biólogo Ronald Aylmer Fisher publica uno de sus trabajos llamado “Análisis Discriminante Lineal”, metodología que estaba relacionada al mundo de la biología y la medicina, pero que luego será utilizada en 1941 por D. Durand dentro del sistema financiero para poder realizar un análisis de préstamos otorgados. Dicha técnica estadística establecía una clasificación en grupos teniendo en cuenta distintas variables que describen al individuo que se intenta clasificar; y en el análisis realizado por Durand, las variables estudiadas fueron: los activos, estabilidad laboral, el sexo y la edad, las mismas determinarían si las operaciones resultarían un buen negocio o un mal negocio para el prestamista.

Luego de diversas investigaciones en el contexto financiero aplicando estas nuevas técnicas, el Credit Scoring toma importancia, y, en consecuencia, en 1968, Edward Altman desarrolla el Z-Score, fórmula que será utilizada para la predicción de quiebras.

EL cálculo para la fórmula del Z-Score es el siguiente:

Altman Z-score = 1,2 * T1 + 1,4 * T2 + 3,3 * T3 + 0,6 * T4 + 1,0 * T5

Donde:

- T1: (Capital Circulante/Activos Totales)
- T2: (Beneficios no distribuidos/Activos Totales)
- T3: (EBITDA/Activos Totales)
- T4: (Capitalización Bursátil/Deuda Total)
- T5: (Ventas Netas/Activos Totales)

Dicha fórmula logra una precisión del 72% con 2 años de antelación respecto de la fecha de la quiebra (dicho porcentaje de precisión aumentará a más años de antelación) y el resultado de falsos negativos se estima en un 6%. De los 22 ratios que originalmente integraban el análisis, se logró quedar con sólo 5 de ellos, los cuales se consideró eran los más relevantes.

Las ventajas de este método por sobre el análisis univariado, es la posibilidad de analizar distintas características en forma simultánea.

En el siguiente gráfico, se muestra un ejemplo del resultado que se obtuvo relevando una muestra de empresas, y determinando si se encuentra expuesta a una futura quiebra. Se demuestra una división en el rango de puntuación, en primer lugar, tenemos a las puntuaciones Z-Score inferiores a 1.81, en donde se encontrarían las empresas con la probabilidad más elevada de caer en quiebra en un futuro próximo, tenemos además, una "zona de ignorancia" que iría desde 1.81 y 2.99 de puntuación; y por ultimo, los mayores a 3 con escasas probabilidades de caer en quiebra.

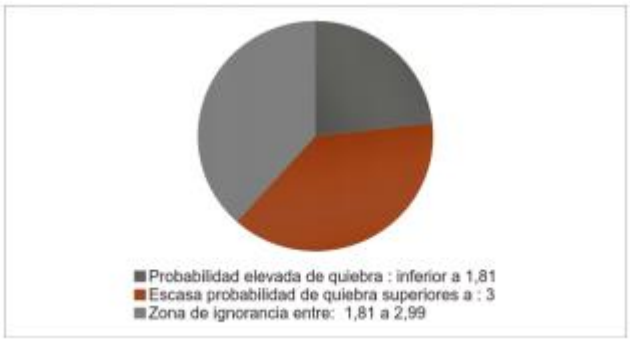


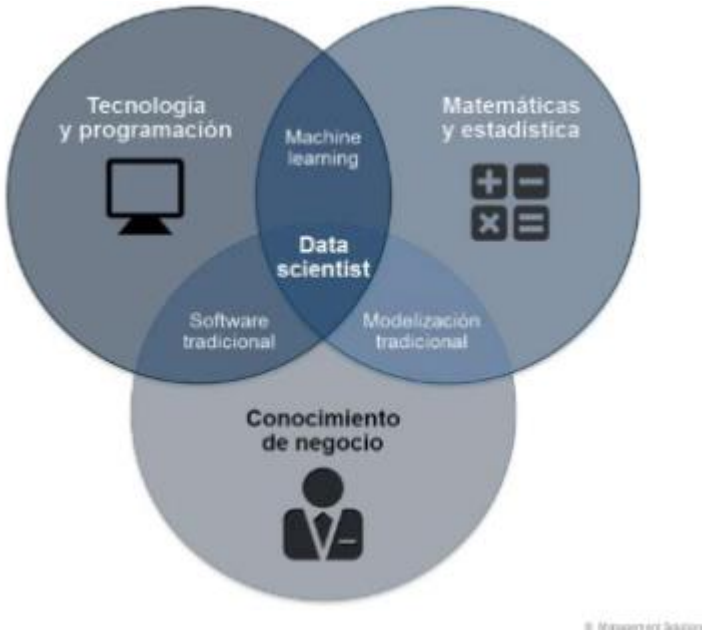
Figura 1: Puntuación Z-Scores Altman Fuente: a partir de la Investigación 1ECO172

1.3 El Scoring bancario y el Big Data

El gran volumen de datos y la transformación digital revolucionan hoy en día al mundo entero, y, por supuesto, el sector financiero no es excluyente en esto, debe adecuarse a los cambios que se presentan y así poder aprovechar toda esa información tanto para la realización de una mejor gestión como para la captación de nuevos clientes, el aumento en las ventas, entre otros objetivos principales.

Si bien la disponibilidad de gran cantidad de información y el rápido acceso a ella nos brinda la posibilidad de realizar un análisis más preciso y llegar a conclusiones en menor tiempo, es necesario tener las herramientas para poder realizar dicho análisis, la capacidad tanto de procesamiento de dicha información como de análisis para detectar o identificar por ejemplo determinados patrones que ayuden en la gestión; sin embargo, al ser un contexto tan cambiante se precisa un actuar rápido y con una metodología que implique el menor error posible.

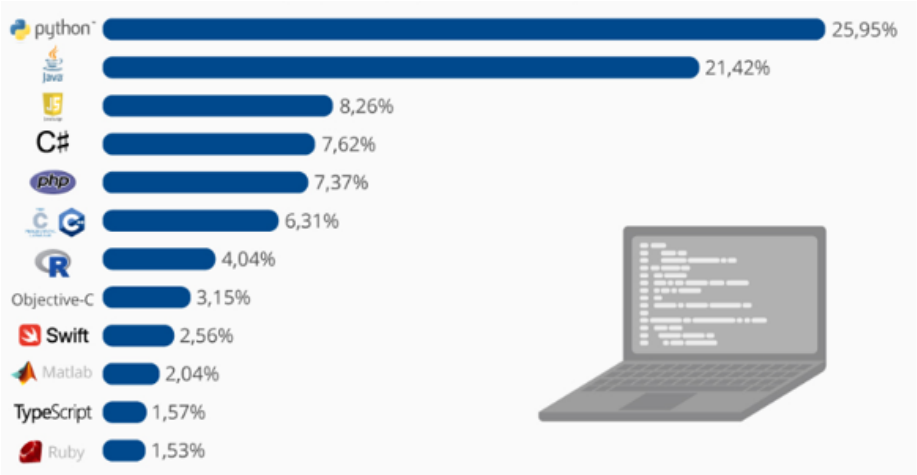
Para dicho fin se empiezan a implementar nuevas técnicas, arquitecturas y también aparece un nuevo rol que ayudará a realizar el análisis adecuado, ya que tendrá conocimientos tanto de estadística, como de programación, tecnología y hasta de negocios, el "*data scientist*".



La Banca Digital 4.0 (Fuente: Management Solutions)

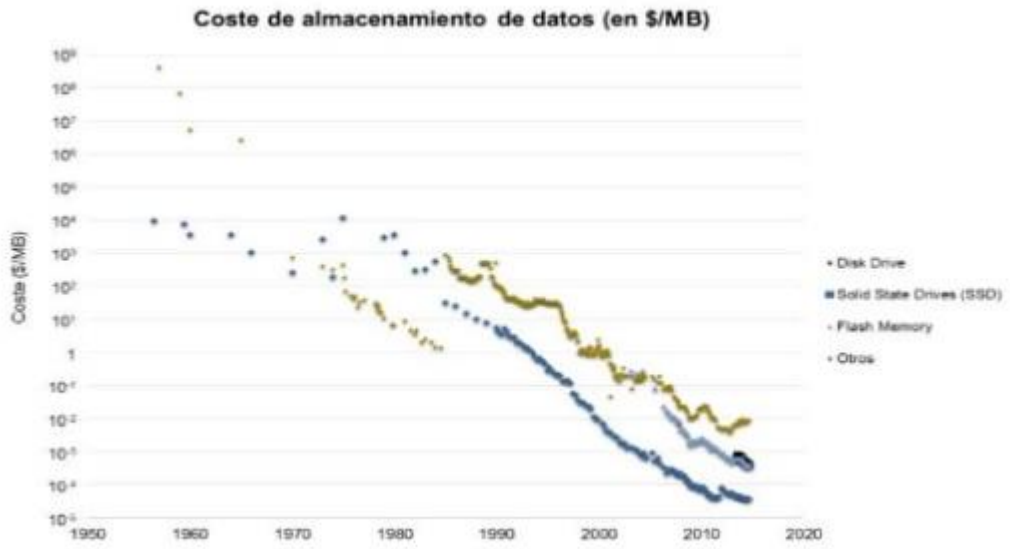
La programación se encuentra muy vinculada al análisis que debe realizarse de los datos, por lo que es importante también identificar los lenguajes de programación más utilizados, ya que determinarán nuestro modo de análisis, ya sea porque ofrece un lenguaje más amigable o porque posee más información o tutoriales a los cuales recurrir a la hora de realizar dicho análisis.

A continuación, se detalla un ranking de los lenguajes de programación más utilizados en 2019:



(Fuente Statista. Datos procedentes del Índice PYPL. Este se basa en las búsquedas de Google de tutoriales de lenguajes de programación. Datos Enero 2019).

Otro de los beneficios de la era digital, es la reducción en los costos por almacenamiento de información, si bien se maneja mayor cantidad de información ésta puede no implicar un espacio físico, el cual determina un costo adicional a la hora de almacenar carpetas, registros o papeles.



La Banca Digital 4.0 (Fuente: Management Solutions)

En conclusión, el sector financiero se encuentra con una reestructuración a la que hacer frente, por medio de la digitalización, llevando también al usuario a cambiar su perfil, y manejarse con los canales digitales, que cada día ganan más terreno por su fácil acceso, por ejemplo, por medio del uso de la telefonía móvil. Esto ayuda a recabar más información en cuanto al perfil del usuario y poder determinar una política de marketing, de publicidad o de captación de clientes que esté basada en el perfil y sea más acertada. Dicha información no debe perderse ni desperdiciarse a la hora de armar un buen plan de negocios.

2. El Análisis de los Datos y la correlación de las variables

Los modelos de evaluación crediticia intentan determinar mediante las relaciones existentes entre las distintas variables que clasifican al solicitante del crédito, una regla que establezca la probabilidad de cumplimiento o incumplimiento de la obligación generada. Para esto se realizará una definición y ponderación de variables cualitativas y cuantitativas, que brindarán información de los clientes para la posterior evaluación crediticia.

Las principales variables son: Edad, Sexo, Calificación Laboral, si es propietario de una casa, si posee ahorros, si tiene cuenta de cheques, el monto del crédito, la duración del mismo y su propósito; dicha base contiene 1000 registros y 9 variables. En este primer análisis, no se hará modificación de los datos, por lo que no tendremos variables categóricas ni dicotómicas que no provengan de la base de datos original y se deja de lado la columna "Risk", que nos ayudará en un análisis posterior a realizar la predicción de si el cliente incumplirá con su pago o no.

La base de datos fue provista por un sitio de internet, donde pueden visualizar distintas "open databases"; en este caso, se utilizó la de un Banco de Alemania, la cual fue provista por el Profesor Hans Hofman, de la Universidad de Hamburgo.

Es necesario aclarar la importancia de la gobernanza de datos, y el marco legal que ampara el uso indebido de información no autorizada. En cuanto a la gobernanza de datos, podemos decir que establece una responsabilidad en cuanto al uso y manipulación de información y que conlleva un conjunto de procedimientos para acceder a ella. Podríamos también mencionar la Ley 25.326 que ampara la protección de datos personales asentados en registros, archivos, etc.

La correlación de las variables

Primero se estableció mediante una Regresión Lineal, las variables más explicativas para la variable a explicar "Risk" con la ayuda de R Studio. En este caso, las variables más representativas para dicho análisis fueron: "Checking Account", "Duration", "Saving Account", y en menor medida, "Sex". (*Apéndice 1*).

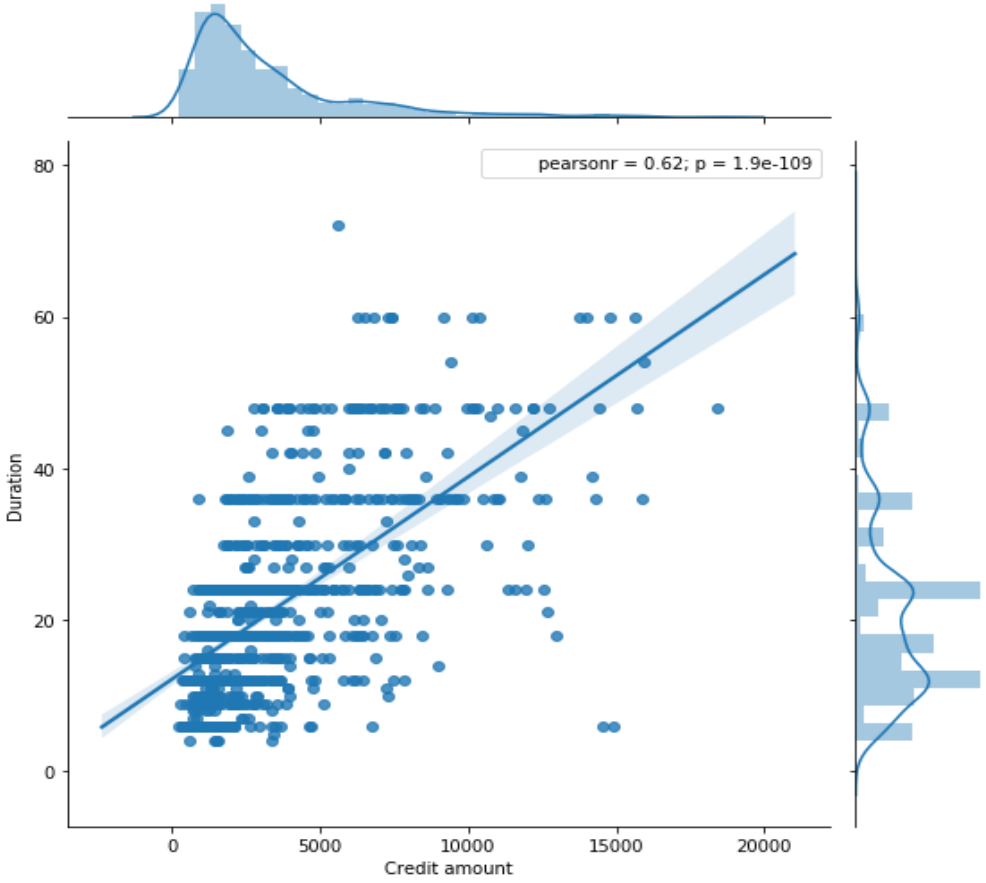
```
Coefficients:
                    Estimate   Std. Error t value   Pr(>|t|)
(Intercept)       0.252024038  0.120526617   2.091    0.03678 *
Age               0.001986757  0.001250807   1.588    0.11252
Checking.account  0.133693531  0.014805953   9.030    < 2e-16 ***
Credit.amount    -0.000005086  0.000006248  -0.814    0.41582
Duration         -0.006700870  0.001435848  -4.667 0.00000348 ***
Housing          -0.010420778  0.027260660  -0.382    0.70235
Job               0.003815223  0.021400277   0.178    0.85854
Purpose           0.010682500  0.006847963   1.560    0.11909
Saving.accounts   0.042575814  0.011080883   3.842    0.00013 ***
Sex               0.075295643  0.029840279   2.523    0.01178 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Luego, a través del coeficiente de correlación lineal de Pearson se pudo determinar, la relación entre las variables Monto del Crédito y Duración del Crédito, lo cual tiene mucho sentido, ya que a mayor monto mayor es la duración del crédito. Por lo que se determina una relación directa entre las variables, cuando una aumenta la otra actúa de igual modo, este coeficiente establece el siguiente rango:

$R = 0$ → No hay relación.

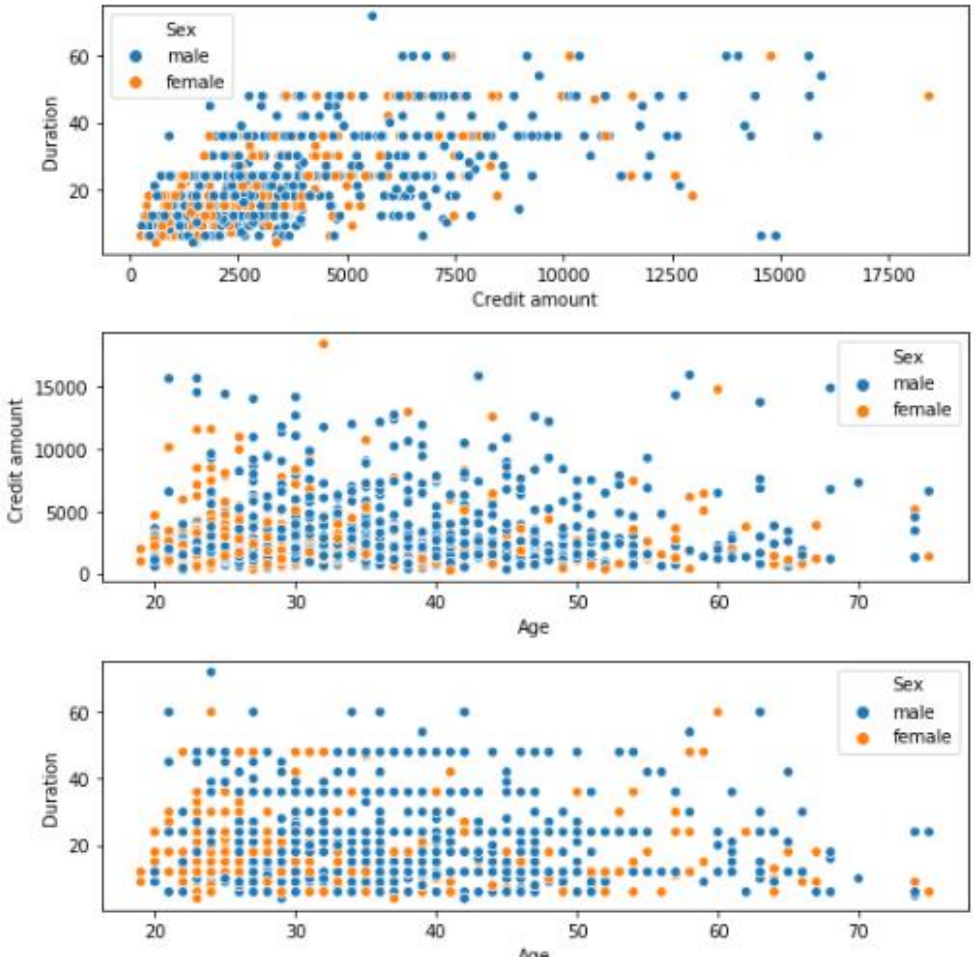
$0 < R < 1$ → Correlación positiva.

$R = -1$ → Correlación inversa.



Como se muestra en el gráfico, en nuestro caso el coeficiente es de 0.62, por lo que obtenemos una relación positiva; este análisis fue realizado con Python (*Apéndice 2*).

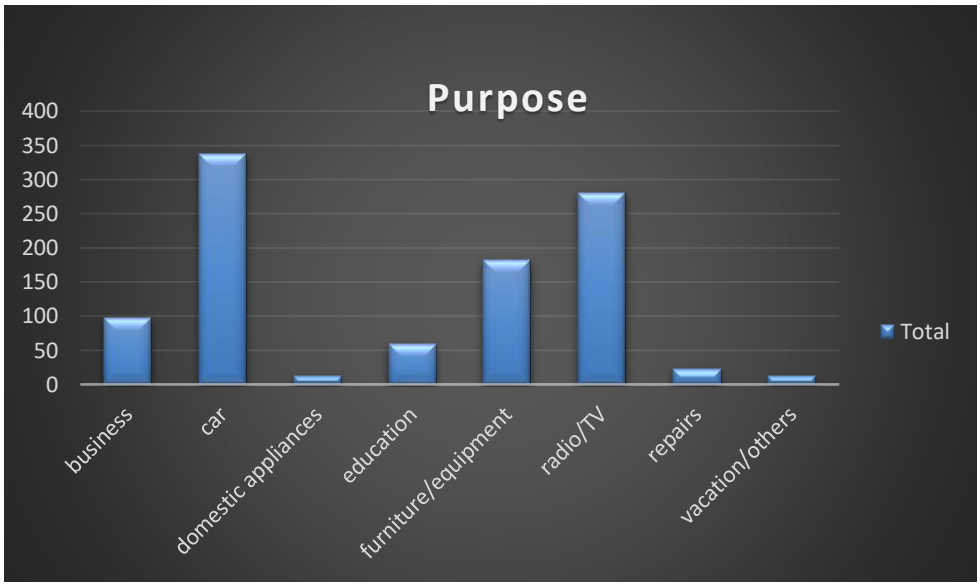
A través de la correlación lineal de la base de datos, se obtiene, por ejemplo, la relación que hay entre el sexo, y la duración y el crédito a otorgar, y se obtuvo el siguiente gráfico de dispersión:



Del mismo, se puede observar que, entre la población de la muestra a analizar, las mujeres tienden a ser más jóvenes que los hombres a la hora de solicitar un préstamo en dicha entidad bancaria.

Uno de los análisis que nos servirá a posteriori, y lo complementaremos con cada uno de los análisis es ponderar el propósito del crédito, en este caso, esto podría servirnos para en un futuro estableciendo el perfil del cliente y un adecuado análisis de los datos provistos de las muestras, realizar una campaña que permita una buena captación de clientes, por ejemplo, en este caso, determinar la edad y el propósito para poder direccionar esa campaña.

Propósito	Cantidad de Casos
business	97
car	**337**
domestic appliances	12
education	59
furniture/equipment	**181**
radio/TV	**280**
repairs	22
vacation/others	12
Total	**1000**



En este caso, determinando los grupos con clustering y analizando los últimos pedidos de crédito, se puedo visualizar el perfil del cliente a captar y, por ejemplo, que la campaña será destinada a que el cliente adquiera autos.

2.1. Las distintas técnicas de aprendizaje automático

El aprendizaje automático es aquel que realiza el sistema, ingresando datos históricos que ayudarán posteriormente a determinar patrones de comportamiento los cuales al brindarle casos particulares podrán predecir o inferir en el comportamiento de dicha problemática en análisis. Se busca que con este método la computadora "aprenda" mediante distintos algoritmos y técnicas para luego generalizar esos comportamientos y aplicarlos a una nueva base de datos y así poder inferir sobre ella.

Existen distintos modelos, los cuales brindarán distintas conclusiones; tenemos los *modelos geométricos* que trabaja en múltiples dimensiones, los *modelos probabilísticos* que clasifican las variables y su distribución y, por último, los *modelos lógicos*, como lo son los árboles de decisión.

En el presente trabajo, las técnicas a utilizar son Random Forest, Decision Tree y Regresión Logística, las primeras hacen referencia a los Árboles de Decisión y son utilizadas para realizar la predicción en cuanto a si el individuo será un potencial moroso o no; además, se realizará un agrupamiento con el método Clustering, el cual nos brindará información sobre el perfil del cliente, y dicha información podrá ser utilizada a futuro.

2.2. Análisis de Componentes principales con R

El Análisis de Componentes principales o PCA, es una técnica del Análisis Multivariado, que busca describir el conjunto de datos utilizando nuevos componentes no correlacionados entre sí, los cuales son confeccionados con las variables originales que sí están correlacionadas entre sí y se muestran dentro de un nuevo componente que las identifica y reemplaza.

Esta técnica pretende no tener información redundante a la hora de realizar un análisis de los datos y a la vez perder la menor cantidad de información en el proceso, busca la mejor proyección de los datos en términos de mínimos cuadrados.

Para realizar este análisis se necesitara de los autovalores y autovectores, ya que con éstos se formará la Transformación Lineal, y se encontrará así un nuevo sistema de coordenadas donde la varianza con más tamaño representará al primer eje o Componente Principal, y la segunda más grande, al segundo eje y así sucesivamente.

Para la realización de esta técnica se utilizó R Studio, y como resultado se obtuvieron nueve componentes, de estos nueve sólo nos quedamos con tres componentes, los cuales están identificados de la siguiente manera:

- Importancia de los componentes:

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp4    Comp.5
Standard deviation     1.3962839 1.1911010 1.0741990 1.0022205 0.9232950
Proportion of Variance 0.2166232 0.1576357 0.1282115 0.1116051 0.0947193
Cumulative Proportion  0.2166232 0.3742589 0.5024704 0.6140755 0.7087948


                           Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     0.91932227 0.87627160 0.80543207 0.59926659
Proportion of Variance 0.09390594 0.08531688 0.07208009 0.03990227
Cumulative Proportion  0.80270076 0.88801764 0.96009773 1.00000000
```

- Autovalores:

```
Component variances:

Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
1.9496088 1.4187216 1.1539034 1.0044459 0.8524737 0.8451534 0.7678519 0.6487208

   Comp.9
0.3591205
```

- Autovectores

```
Component loadings:
                      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5
Age                0.159283718  0.56692852  0.22343528  0.08632314  0.03526222
Checking.account -0.062097916  0.36325479 -0.58097418 -0.18825620  0.26198468
Credit.amount     0.581044069 -0.22521869 -0.14807747  0.02688783 -0.14342206
Duration          0.552413070 -0.25844386 -0.13998908 -0.03529712 -0.33094490
Housing          -0.323526093 -0.43510160 -0.28633610  0.20404683  0.06419643
Job               0.358821642 -0.08683025 -0.13242314 -0.34353763  0.72191385
Purpose          -0.187069957 -0.03430737  0.03754752 -0.88193854 -0.36495857
Saving.accounts  -0.005984088  0.30004129 -0.65469338  0.13214233 -0.35286888
Sex               0.244022651  0.37558842  0.20362784  0.01444596 -0.13051511

                      Comp.6      Comp.7      Comp.8       Comp.9
Age                0.450541777  0.22621481  0.57361155  0.106822624
Checking.account -0.004918676 -0.63303595  0.14545254  0.025831988
Credit.amount     0.093147035 -0.09663824  0.22056369 -0.708364942
Duration          0.077255830 -0.16000597  0.05795909  0.680256998
Housing          -0.235424869  0.17604071  0.69563124  0.089189640
Job              -0.100843151  0.42781675 -0.06290345  0.099199709
Purpose           0.013063833  0.11473504  0.18332321 -0.066824264
Saving.accounts  -0.018573079  0.53464741 -0.22810132 -0.033390501
Sex              -0.846305646 -0.01104560  0.15591176  0.003921554
```

Con estos datos entonces, podremos armar nuevos componentes, que así reduzcan la cantidad de variables a analizar, y los identificaríamos con nombres de acuerdo a las variables originales a las que están representando. Por ejemplo:

El componente 1: Las variables más correlacionadas con este factor son la duración y el monto del crédito. Tiene que ver más con el crédito en sí.

El componente 2: Se correlaciona con las variables Edad, Sexo, Saldo de la cuenta corriente y de ahorros. Describe a la persona y sus características, ya sea cualitativas como financieras.

El componente 3: Correlaciona las cajas de ahorro y cuenta corriente del cliente con una relación inversa, podría hablarse aquí sólo del estado financiero.

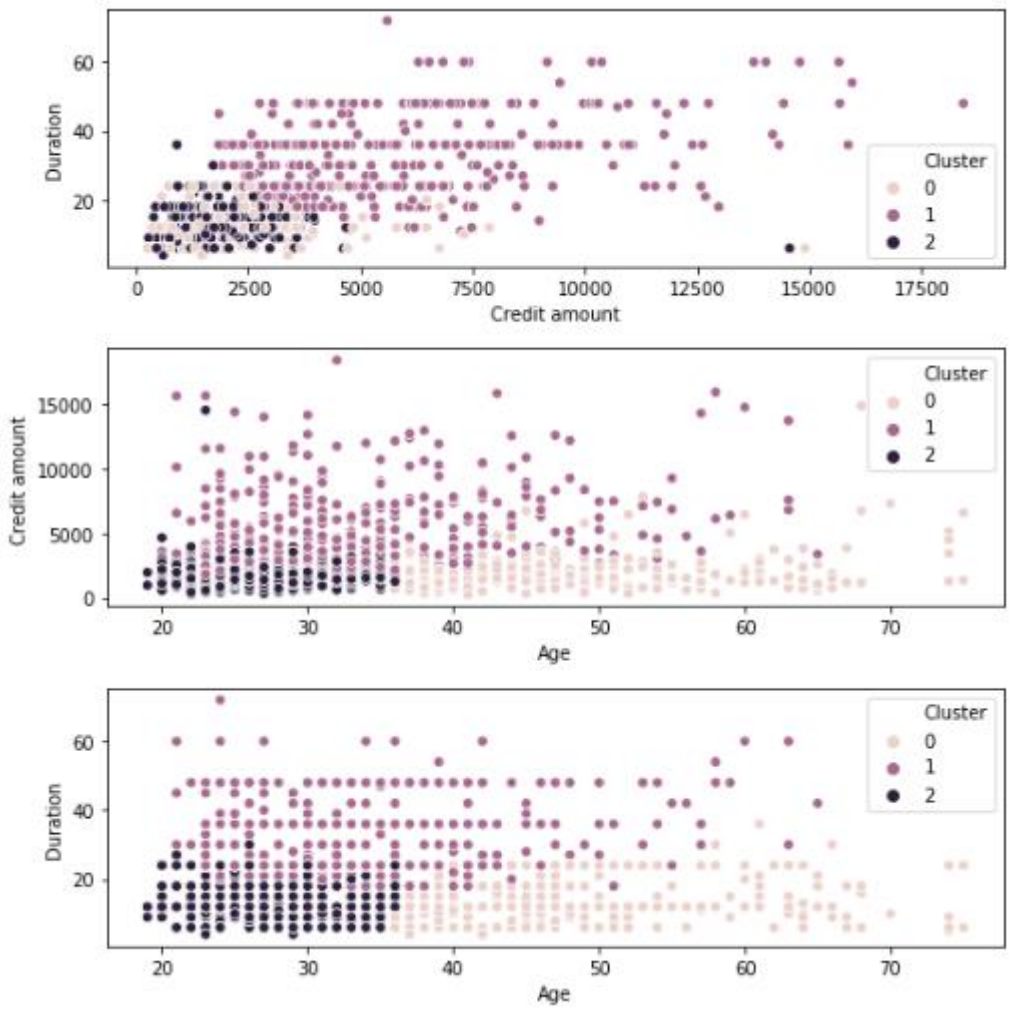
El componente 4: Correlaciona el trabajo con el propósito del crédito. Está ligado al fin al que se le dará al crédito en caso que el cliente lo obtenga.

Entonces podríamos así elegir estos cuatro componentes que albergan una clara identificación de los factores o componentes principales, abarcando dentro de los mismos todas las variables necesarias o explicativas del modelo en cuestión.

2.3. La aplicación de Clustering con K-Means

Como se mencionó en el apartado anterior, el método de Clustering busca agrupar a los individuos en distintos grupos donde se compartan características similares dentro de los grupos y, a su vez, sean diferentes entre los grupos.

En este caso, las variables que se utilizaron para dicho análisis fueron: “Age”, “Duration” y “Sex”, y se observan entonces una división en tres grupos, los cuales se muestran en el siguiente gráfico de dispersión:



Si agrupamos la información, se puede establecer la siguiente tabla:

	Age	*Credit amount*	*Duration*
Cluster			
0	48.6	1970.5	13.9
1	34.0	5665.4	32.2
2	27.7	1737.5	14.3

Del análisis se concluye, que se puede estimar para el primer grupo una edad promedio de 49 años, monto bajo y una duración baja también. Por el contrario, el tercer

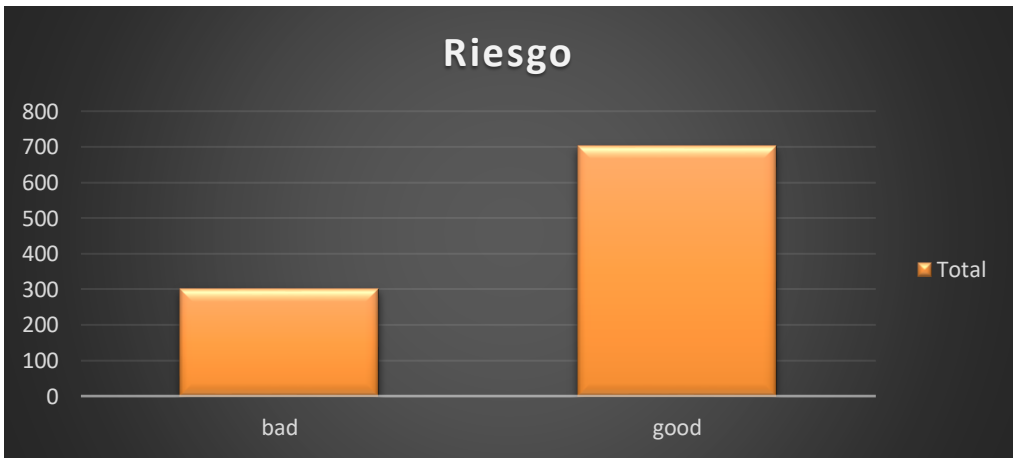
grupo con una edad promedio de 34 años es el del monto más elevado a la hora de pedir un crédito, y, como hemos analizado antes, a mayor crédito se establece mayor duración del crédito. El tercer y último grupo, establece una edad promedio de 28 años, un monto bajo y a la vez, una duración baja del crédito también.

3. Métodos predictivos

Los métodos predictivos a utilizar en este apartado son: Random Forest, Regresión Logística y Decision Tree, ambos son técnicas de Árboles de decisión y se medirá su performance para poder predecir en este caso el incumplimiento de los clientes, considerando las distintas variables analizadas a lo largo del trabajo.

3.1. La problemática en análisis

El siguiente gráfico muestra como están distribuidos los datos, de acuerdo a los distintos resultados en el riesgo, como se ve a continuación existen más personas con cumplimiento en sus deudas que las que resultaron ser morosas:



Lo que se busca a través de estas técnicas de aprendizaje automático, es encontrar una metodología que nos permita predecir con distintas variables el comportamiento del cliente, y así poder reducir la morosidad, sin necesidad de llegar a incurrir en costos innecesarios inclusive a la hora de realizar la gestión del recupero de la deuda.

Se analizarán estos distintos métodos seleccionados y se establecerá cuál tiene mejor performance en cuanto a predicción de los datos, estableciendo una división en la base de datos que servirá de entrenamiento y otra que servirá para la prueba final.

3.2. Estimación del modelo

El primer análisis se realizó con Random Forest, o también conocido como "Bosques Aleatorios", el primer paso fue modificar la base de datos estableciendo variables solo numéricas, ya sea categóricas como dicotómicas.

Luego se realiza una división entre los datos que van a ser tomados como datos históricos y que ayudarán a realizar la predicción, es decir, se realiza el entrenamiento de los datos, dividiendo así la base de datos en 80% y 20% para el test final.

Se determina luego distintas profundidades en Random Forest, en estos casos se tomó como profundidad 100, 1000 y 2000. Los resultados fueron los siguientes en cuanto a la exactitud del modelo:

```
100 exactitud: 0.76
1000 exactitud: 0.75
2000 exactitud: 0.755

Matriz de Confusión:

[[ 28  30]
 [ 18 124]]

Curva ROC (Característica Operativa del Receptor- la razón de
verdaderos positivos)

Curva ROC - AUC del modelo:
0.6779990286546866
```

Para comprobar se realizó con distintos árboles, que marcaban un salto de 500, y se obtuvo la misma conclusión:

```
1000 5 exactitud: 0.7629523809523809 0.02114392961672895
1500 5 exactitud: 0.7639523809523809 0.022063400707815924
2000 5 exactitud: 0.7623571428571428 0.02340833427176489
2500 5 exactitud: 0.7637857142857143 0.021730205108534052
3000 5 exactitud: 0.7634285714285713 0.022031388657272652
3500 5 exactitud: 0.7633809523809523 0.0225787360081979
4000 5 exactitud: 0.7635714285714286 0.021998891953131056
4500 5 exactitud: 0.7632142857142858 0.022426119570765454
1000 10 exactitud: 0.7578571428571428 0.018236136908413594
1500 10 exactitud: 0.7587380952380952 0.020666831247973087
2000 10 exactitud: 0.7571190476190475 0.01779614980398302
2500 10 exactitud: 0.7553095238095239 0.018233121291159685
3000 10 exactitud: 0.7585000000000001 0.019857285599578285
3500 10 exactitud: 0.7576666666666667 0.01761109859295534
4000 10 exactitud: 0.7585238095238094 0.01866393323038097
4500 10 exactitud: 0.7577619047619049 0.01893783840241561
1000 15 exactitud: 0.7532857142857143 0.019512569354438982
1500 15 exactitud: 0.7511428571428572 0.01948204010419119
2000 15 exactitud: 0.7511904761904762 0.017938702632971922
2500 15 exactitud: 0.7516904761904761 0.01904187413364404
3000 15 exactitud: 0.7538333333333334 0.0191329043067403
3500 15 exactitud: 0.7533333333333333 0.020943585816379968
4000 15 exactitud: 0.7525119047619048 0.019332116428978568
4500 15 exactitud: 0.7522142857142857 0.020815652361142918
1000 20 exactitud: 0.7489047619047619 0.01930883387881404
1500 20 exactitud: 0.751107142857143 0.017884962456940413
2000 20 exactitud: 0.7498452380952381 0.01871580782614087
2500 20 exactitud: 0.7505119047619047 0.018873069272489048
3000 20 exactitud: 0.7523452380952381 0.01940945750958351
3500 20 exactitud: 0.75125 0.0192820396264035
4000 20 exactitud: 0.7518690476190477 0.01746504445393079
4500 20 exactitud: 0.7515833333333333 0.018262758408061
1000 25 exactitud: 0.7506904761904762 0.019585942075481748
1500 25 exactitud: 0.7508095238095238 0.019117453826708385
2000 25 exactitud: 0.7500357142857144 0.01934960024079203
2500 25 exactitud: 0.752 0.01959281504431011
3000 25 exactitud: 0.7517023809523808 0.01887400791043531
3500 25 exactitud: 0.7523809523809524 0.01827053251212574
4000 25 exactitud: 0.7511071428571429 0.01848938685581153
4500 25 exactitud: 0.7517142857142857 0.018310259134407682
```

Por lo tanto, a medida que aumento la profundidad en el árbol de decisión perdemos precisión, y la exactitud máxima lograda es en este caso de 76%.

Por otro lado, se analizaron los datos por medio de Regresión Logística, esta técnica es utilizada para predecir el resultado de una variable en este caso categórica (“Risk”) en

función de otras variables. Realizando el entrenamiento del modelo, se obtuvo el siguiente resultado:

```
Exactitud del modelo:

0.715
```

Matriz de Confusión:

```
[[ 23  35]
 [ 22 120]]

Curva ROC - AUC del modelo:
0.6208110733365712
```

Finalmente, se realiza el análisis con Decision Tree, una rama de la técnica de Arboles de decisión, con la cual podemos predecir un comportamiento o patrón. Del análisis de los datos se obtuvo el siguiente resultado:

```
Exactitud del modelo:
0.665
```

Matriz de Confusión:

```
[[ 27  32]
 [ 35 106]]


Curva ROC - AUC del modelo:
0.604700084144729
```

3.3 Comparación de resultados

A raíz de los resultados obtenidos, se determina que el modelo con mejor performance a la hora de predecir si el cliente no podrá cumplir con la deuda es Random Forest, obteniendo una exactitud en dicha predicción del 76%.

La Curva ROC también demuestra una superioridad en el modelo por sobre los demás analizados, ya que ésta mide los verdaderos positivos, o bien los aciertos que se tuvo al momento de realizar la predicción, y ese ratio me da casi 68%, mientras que los demás solamente llegan a 62%.

Se estableció además para comprobar la profundidad del árbol, es decir, hasta cuantos arboles debería yo aceptar en mi análisis y así obtener la mejor performance, y lo que se obtuvo que con 100 arboles yo obtenía la mejor performance, y una vez que iba aumentando en profundidad iba perdiendo precisión, y así el porcentaje de exactitud en la predicción era cada vez más bajo.

Conclusión

A raíz de la base de datos analizada, pudimos sacar distintas conclusiones:

- La primera se estableció con la correlación que se determinó entre las variables, existiendo una correlación fuerte entre la edad, el monto del crédito y la duración del mismo.
- La segunda conclusión se basa en el análisis de clustering y los grupos bien definidos que se establecen a raíz de este análisis, ya que divide a la muestra en tres grandes grupos:

  -El primer grupo está conformado por personas con una edad promedio de 49 años, los cuales obtienen montos bajo de préstamos y la duración de estos préstamos es corta.

  - El segundo grupo establece un cliente de una edad promedio de 34, el cual busca préstamos de montos elevados y con larga duración de los mismos.

  - Y el tercer y último grupo, es aquel que está formado por personas con una edad promedio de 28 y toma créditos de bajo monto y financiados a corto plazo.

  Esta observación puede determinar una gran información del perfil del cliente que se acerca al banco para realizar la petición de un producto, en este caso un crédito, pero que puede ser identificado y analizado para ofrecerle otros productos con una captación de clientes o campaña que sea exclusivamente dirigida en este caso a personas de una edad promedio de 34, las cuales, si analizamos la base de datos, en su mayoría buscan créditos para poder financiar la compra de autos.

  - Por último, el modelo predictivo estableció que la mejor predicción que puede realizarse en base a los métodos que fueron analizados, en este caso tres (Random Forest, Decision Tree y Regresión Logística), se realizará con Random Forest, ya que se obtuvo una exactitud por encima de los demás y

una curva ROC con valores aceptables, con lo cual nos serviría para el trabajo o análisis que pretendemos hacer sobre el incumplimiento de los clientes que se acercan a la entidad a obtener un crédito.

Si bien la base de datos, daba lugar a mucho más análisis y podríamos hasta utilizarlo, como se mencionó antes para comercializar otros productos, se han podido aplicar distintas técnicas de aprendizaje automático y de clasificación, y a la vez se ha podido utilizar y perfeccionar el uso de las herramientas como Python y R, que son lenguajes muy utilizados en el mercado para dichas predicciones o análisis; por lo que en el futuro todos estos conocimientos podrán ser utilizados, ya sea en la rama bancaria como en cualquier ámbito que se necesite de un análisis estadístico de los datos.

Referencias bibliográficas y bibliografía

1. BCRA (2019), Lineamientos para la Gestión de Riesgos en Entidades Financieras.
2. Vargas Sanchez, A. & Castelú, S. (2014), Medición del Riesgo Crediticio mediante aplicación de Métodos Basados en Calificaciones Internas.
3. Puertas Medina, R. & Martí Selva, M.L. (2013), Análisis del Credit Scoring.
4. Ross, Westerfield & Jaffe, (2012), Finanzas Corporativas.
5. Joos, P., Vanhoof, K., Ooghe, H. & Sierens, N. (1998), Credit classification: a comparison of logit models and decision trees. Faculteit Economie en Bedrijfskunde.
6. Mays, E. (2001), Handbook of Credit Scoring, The Glenlake Publishing Company Ltda.
7. Ong, M. (1999), Internal Credit Risk Models, Risk Books.
8. Kiviat, B (2017) The art of deciding with data: evidence from how employers translate credit reports into hiring decisions.
9. Pascal, R. (1998), Decisiones financieras, 3ª ed., Edit. Macchi.
10. Altman Edward (2002) Revisiting Credit Scoring Models in a Basel 2 Environment New York University (NYU) - Salomon Center; New York University (NYU) - Department of Finance May 2002
11. Foster Provost, T. F., 2013. Data Science for Business. O'Reilly.
12. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837–845.
13. Kashyap, P. (2018). Machine Learning for Decision Makers: Cognitive Computing Fundamentals for Better Decision Making.
14. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Anexos y Apéndices

1- REGRESION LINEAL CON R

```
Rcmdr>  RegModel.2 <-
Rcmdr+    lm(Risk~Age+Checking.account+Credit.amount+Duration+Housing+Jo
b+Purpose+Saving.accounts+Sex,
Rcmdr+     data=german_credit_data2_)

Rcmdr>  summary(RegModel.2)

Call:
lm(formula = Risk ~ Age + Checking.account + Credit.amount +
    Duration + Housing + Job + Purpose + Saving.accounts + Sex,
    data = german_credit_data2_)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1289 -0.4256  0.1504  0.3054  0.7920

Coefficients:
                    Estimate   Std. Error t value   Pr(>|t|)
(Intercept)       0.252024038  0.120526617   2.091    0.03678 *
Age               0.001986757  0.001250807   1.588    0.11252
Checking.account  0.133693531  0.014805953   9.030    < 2e-16 ***
Credit.amount    -0.000005086  0.000006248  -0.814    0.41582
Duration         -0.006700870  0.001435848  -4.667 0.00000348 ***
Housing          -0.010420778  0.027260660  -0.382    0.70235
Job               0.003815223  0.021400277   0.178    0.85854
Purpose           0.010682500  0.006847963   1.560    0.11909
Saving.accounts   0.042575814  0.011080883   3.842    0.00013 ***
Sex               0.075295643  0.029840279   2.523    0.01178 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4217 on 990 degrees of freedom
Multiple R-squared:  0.1618,   Adjusted R-squared:  0.1542
F-statistic: 21.23 on 9 and 990 DF,  p-value: < 2.2e-16
```

2- COEFICIENTE DE PEARSON CON PYTHON

```
import scipy.stats as stats

r1 = sns.jointplot(x="Credit amount",y="Duration", data=data, kind="reg", height=8)

r1.annotate(stats.pearsonr)

plt.show()
```

3- ANÁLISIS DE COMPONENTES PRINCIPALES CON R

```
Rcmdr>  local({
Rcmdr+    .PC <-
Rcmdr+    princomp(~Age+Checking.account+Credit.amount+Duration+Housing+
Job+Purpose+Saving.accounts+Sex,
Rcmdr+     cor=TRUE, data=german_credit_data2_)
Rcmdr+    cat("\nComponent loadings:\n")
Rcmdr+    print(unclass(loadings(.PC)))
Rcmdr+    cat("\nComponent variances:\n")
Rcmdr+    print(.PC$sd^2)
Rcmdr+    cat("\n")
Rcmdr+    print(summary(.PC))
Rcmdr+  })

Component loadings:
                  Comp.1      Comp.2      Comp.3      Comp.4      Comp.5

Age               0.159283718  0.56692852  0.22343528  0.08632314  0.035
26222
Checking.account -0.062097916  0.36325479 -0.58097418 -0.18825620  0.261
98468
Credit.amount     0.581044069 -0.22521869 -0.14807747  0.02688783 -0.143
42206
Duration          0.552413070 -0.25844386 -0.13998908 -0.03529712 -0.330
94490
Housing          -0.323526093 -0.43510160 -0.28633610  0.20404683  0.064
19643
Job               0.358821642 -0.08683025 -0.13242314 -0.34353763  0.721
91385
Purpose          -0.187069957 -0.03430737  0.03754752 -0.88193854 -0.364
95857
Saving.accounts  -0.005984088  0.30004129 -0.65469338  0.13214233 -0.352
86888
Sex               0.244022651  0.37558842  0.20362784  0.01444596 -0.130
51511
                      Comp.6      Comp.7      Comp.8       Comp.9
Age               0.450541777  0.22621481  0.57361155  0.106822624
Checking.account -0.004918676 -0.63303595  0.14545254  0.025831988
Credit.amount     0.093147035 -0.09663824  0.22056369 -0.708364942
Duration          0.077255830 -0.16000597  0.05795909  0.680256998
Housing          -0.235424869  0.17604071  0.69563124  0.089189640
Job              -0.100843151  0.42781675 -0.06290345  0.099199709
Purpose           0.013063833  0.11473504  0.18332321 -0.066824264
Saving.accounts  -0.018573079  0.53464741 -0.22810132 -0.033390501
Sex              -0.846305646 -0.01104560  0.15591176  0.003921554

Component variances:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
Comp.8
1.9496088 1.4187216 1.1539034 1.0044459 0.8524737 0.8451534 0.7678519 0.
6487208
   Comp.9
0.3591205

Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
```

```
Standard deviation     1.3962839 1.1911010 1.0741990 1.0022205 0.9232950
Proportion of Variance 0.2166232 0.1576357 0.1282115 0.1116051 0.0947193
Cumulative Proportion  0.2166232 0.3742589 0.5024704 0.6140755 0.7087948

                           Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     0.91932227 0.87627160 0.80543207 0.59926659
Proportion of Variance 0.09390594 0.08531688 0.07208009 0.03990227
Cumulative Proportion  0.80270076 0.88801764 0.96009773 1.00000000
```

4- CLUSTERING CON K-MEANS EN PYTHON

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans, AffinityPropagation
import warnings
warnings.filterwarnings("ignore")

data = pd.read_csv("german_credit_data_.csv")

data.head()
```

Out:

	ID	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	0	67	male	2	own	NaN	little	1169	6	radio/TV	good
1	1	22	female	2	own	little	moderate	5951	48	radio/TV	bad
2	2	49	male	1	own	little	NaN	2096	12	education	good
3	3	45	male	2	free	little	little	7882	42	furniture/equipment	good
4	4	53	male	2	free	little	little	4870	24	car	bad

```
data.drop(data.columns[0], inplace=True, axis=1)
```

```
print("Database has {} observations (customers) and {} columns (attributes).".format(data.shape[0],data.shape[1]))
print("Missing values in each column:\n{}".format(data.isnull().sum()))
print("Columns data types:\n{}".format(data.dtypes))
```

```
Out: Database has 1000 observations (customers) and 10 columns (attributes).
Missing values in each column:
Age                   0
Sex                   0
Job                   0
Housing               0
Saving accounts     183
Checking account    394
Credit amount         0
Duration              0
Purpose               0
Risk                  0
dtype: int64
Columns data types:
Age                  int64
Sex                 object
Job                  int64
Housing             object
Saving accounts     object
Checking account    object
Credit amount        int64
Duration             int64
Purpose             object
Risk                object
dtype: object
```

```
n_unique = data.nunique()
print("Number of unique values:\n{}".format(n_unique))
```

```
Out:
Number of unique values:
Age                  53
Sex                   2
Job                   4
Housing               3
Saving accounts       4
```

```
Checking account      3
Credit amount       921
Duration             33
Purpose               8
Risk                  2
dtype: int64
```

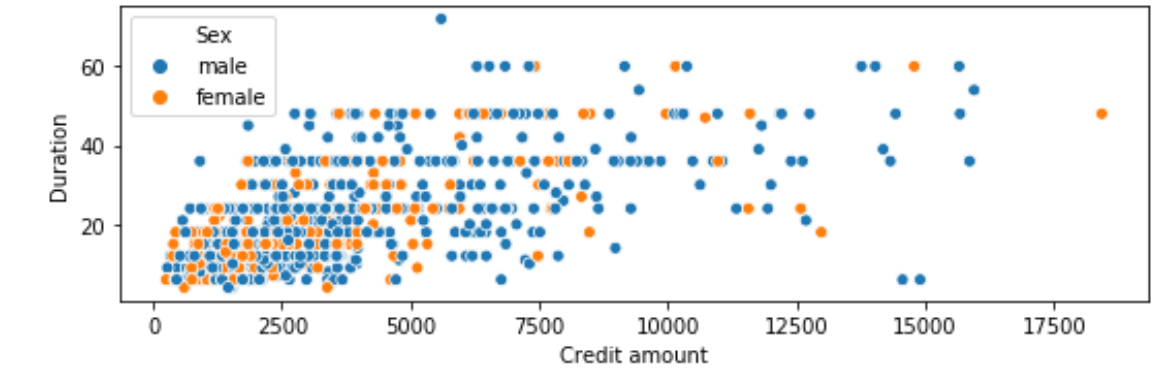
```
for col in data.select_dtypes(include=[object]):
  print(col,":", data[col].unique())
```

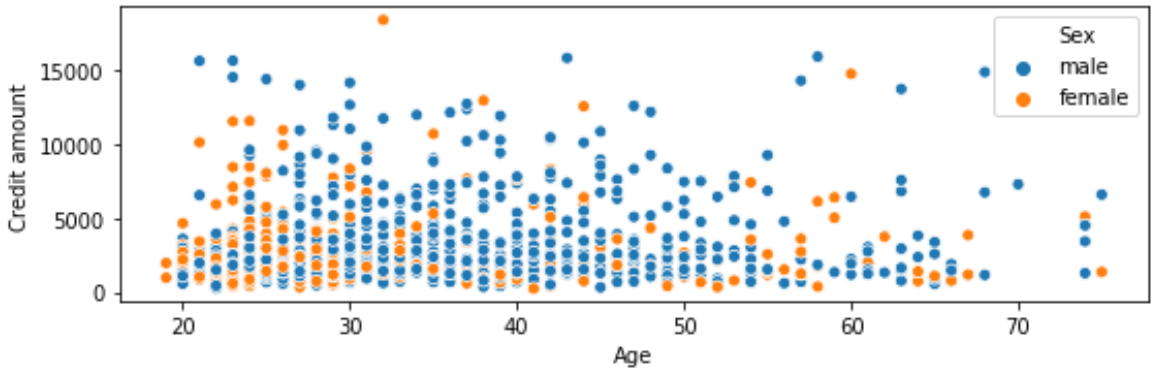
```
Unique values in each categorical column:
Sex : ['male' 'female']
Housing : ['own' 'free' 'rent']
Saving accounts : [nan 'little' 'quite rich' 'rich' 'moderate']
Checking account : ['little' 'moderate' nan 'rich']
Purpose : ['radio/TV' 'education' 'furniture/equipment' 'car' 'business'
 'domestic appliances' 'repairs' 'vacation/others']
Risk : ['good' 'bad']
```

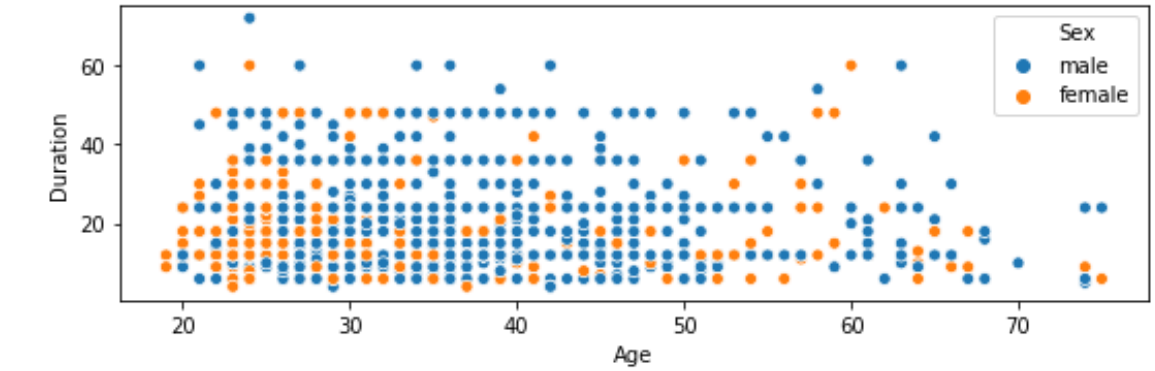
```
def scatters(data, h=None, pal=None):
  fig, (ax1, ax2, ax3) = plt.subplots(3,1, figsize=(8,8))
  sns.scatterplot(x="Credit amount",y="Duration", hue=h, palette=pal, data=data, ax=ax1)
  sns.scatterplot(x="Age",y="Credit amount", hue=h, palette=pal, data=data, ax=ax2)
  sns.scatterplot(x="Age",y="Duration", hue=h, palette=pal, data=data, ax=ax3)
  plt.tight_layout()
```

In [12]:

```
scatters(data, h="Sex")
```



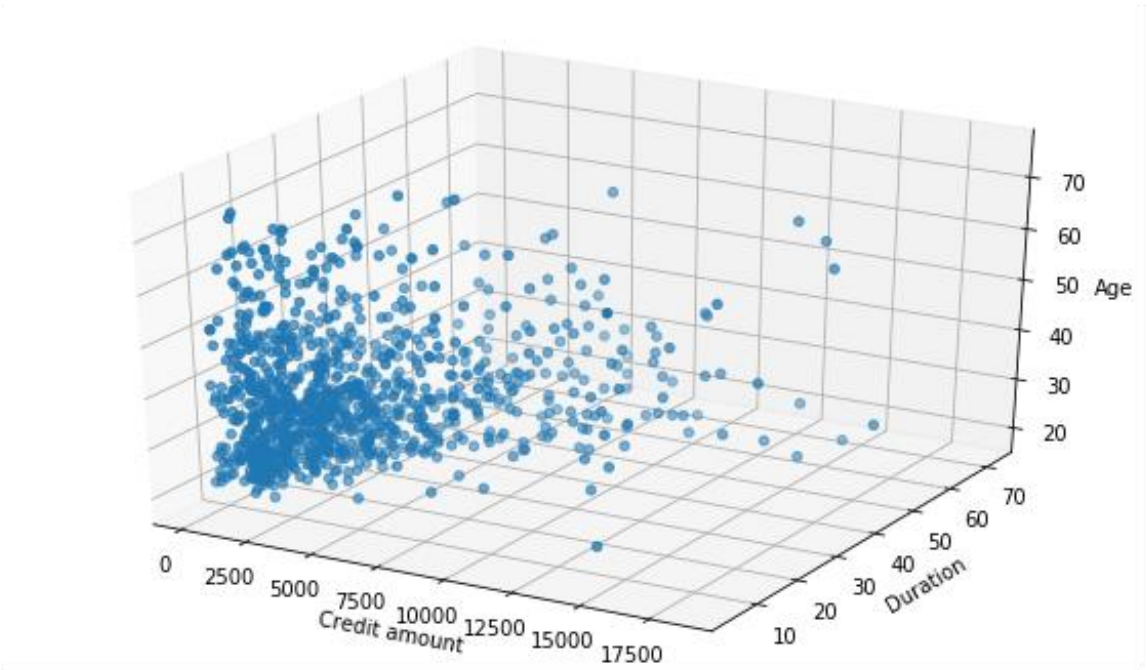




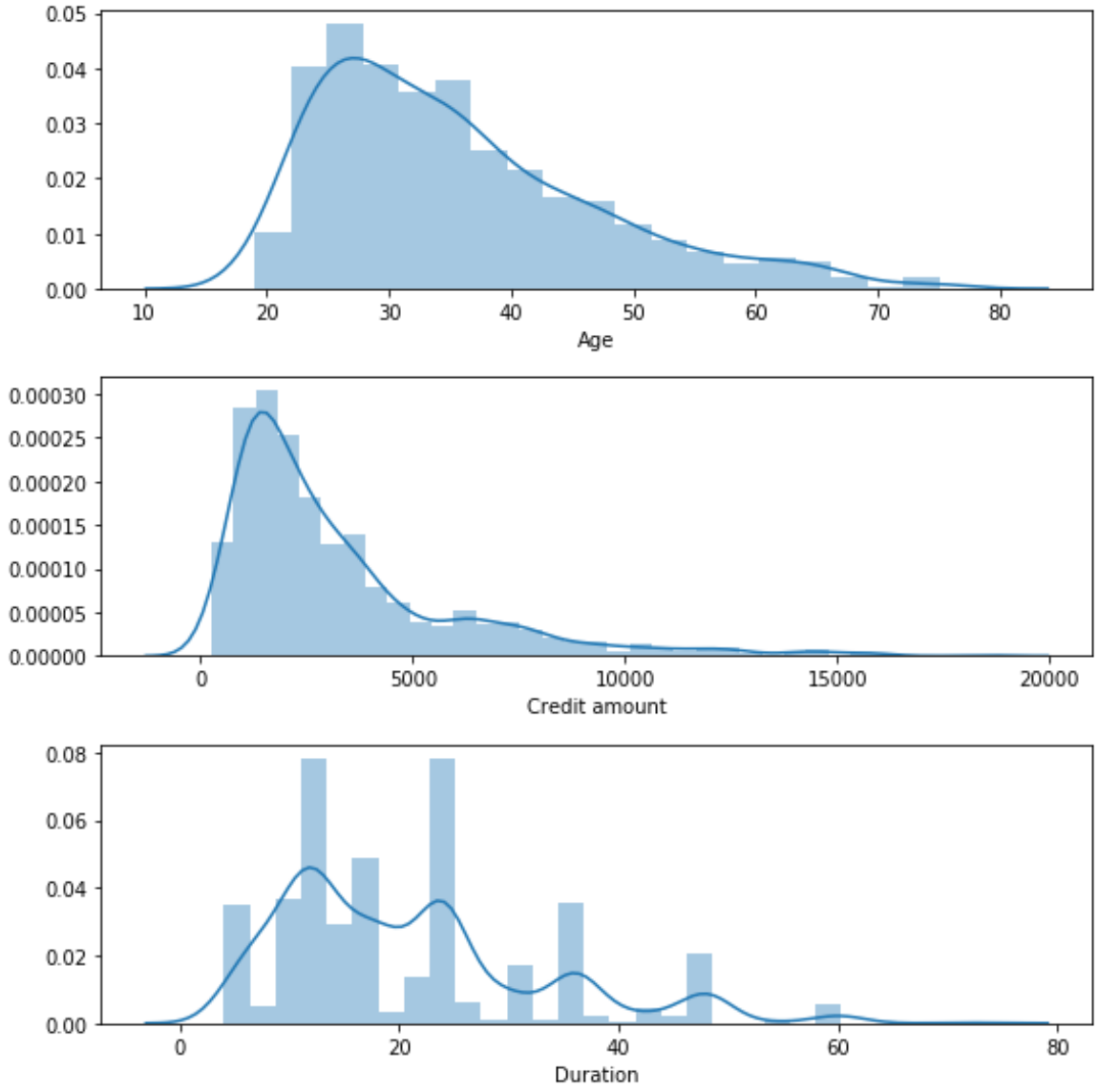

```
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(figsize=(10,6))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(data["Credit amount"], data["Duration"], data["Age"])
ax.set_xlabel("Credit amount")
ax.set_ylabel("Duration")
ax.set_zlabel("Age")
```

Out[18]:

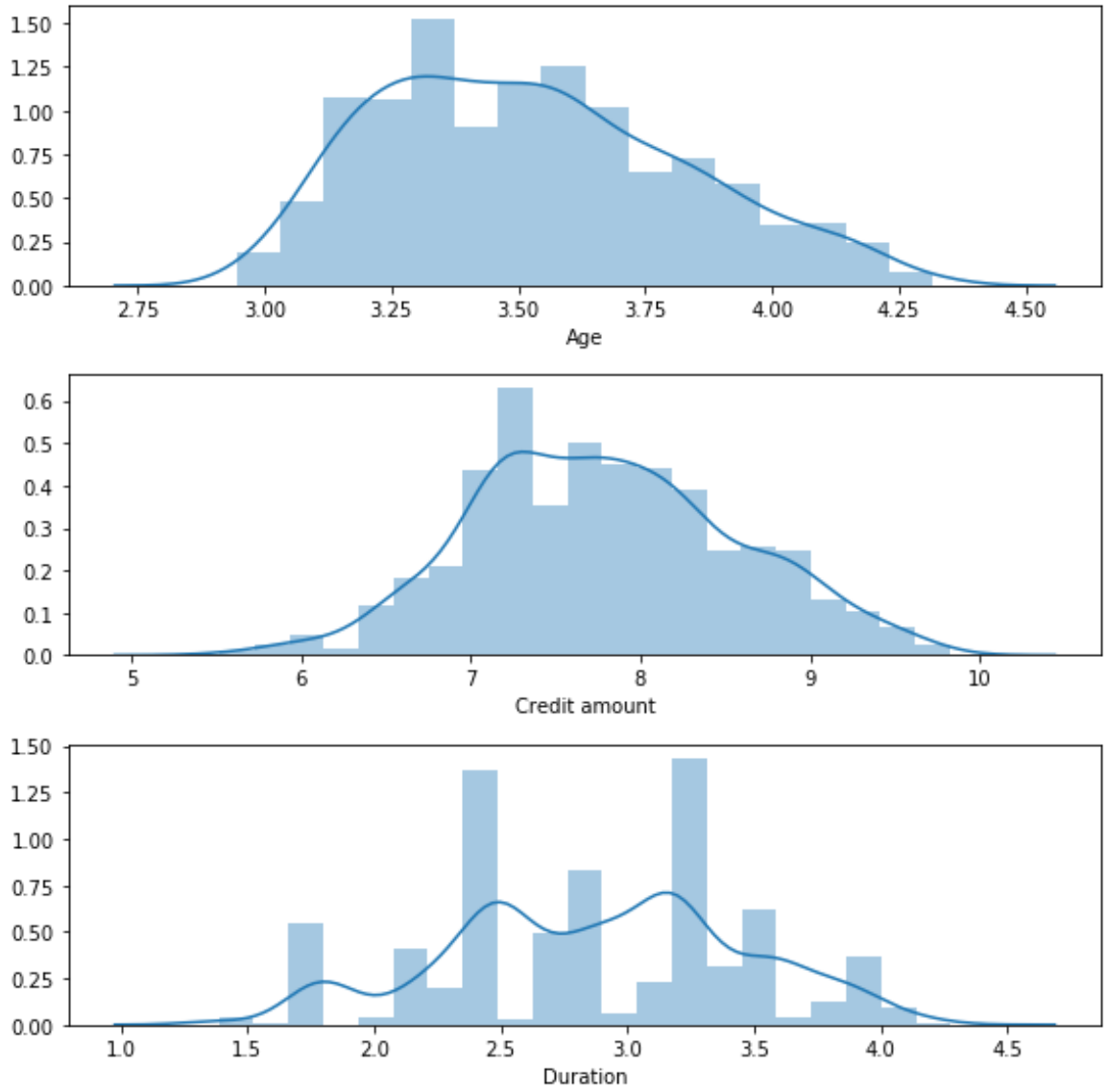
```
Text(0.5, 0, 'Age')
```




```
#Selecciono las columnas a analizar
selected_cols = ["Age","Credit amount", "Duration"]
cluster_data = data.loc[:,selected_cols]


def distributions(df):
    fig, (ax1, ax2, ax3) = plt.subplots(3,1, figsize=(8,8))
    sns.distplot(df["Age"], ax=ax1)
    sns.distplot(df["Credit amount"], ax=ax2)
    sns.distplot(df["Duration"], ax=ax3)
    plt.tight_layout()


distributions(cluster_data)
```




```
cluster_log = np.log(cluster_data)
distributions(cluster_log)
```



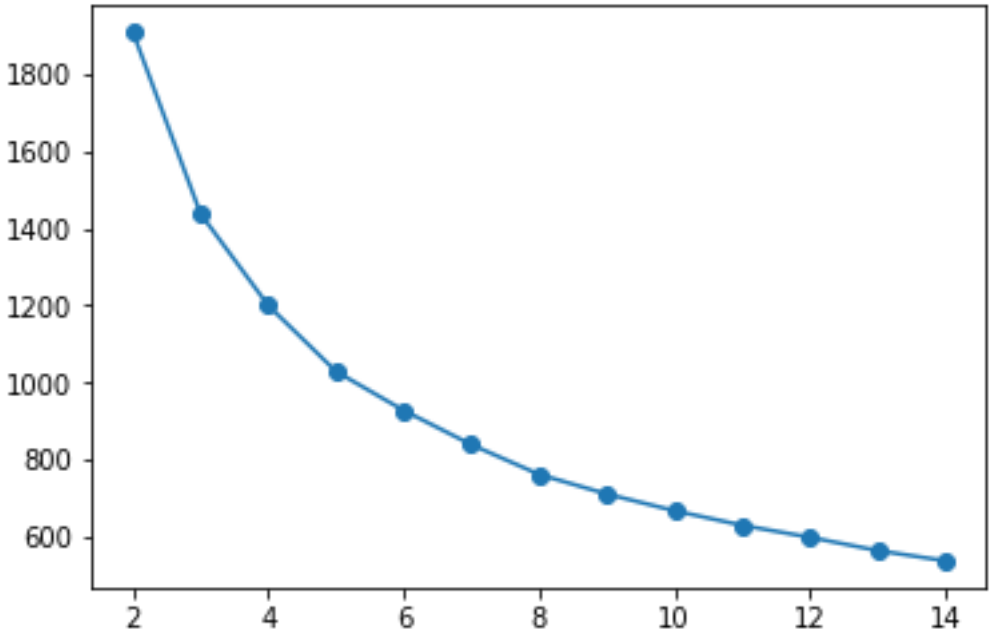

```
scaler = StandardScaler()
cluster_scaled = scaler.fit_transform(cluster_log)
```

```
clusters_range = [2,3,4,5,6,7,8,9,10,11,12,13,14]
inertias =[]

for c in clusters_range:
    kmeans = KMeans(n_clusters=c, random_state=0).fit(cluster_scaled)
    inertias.append(kmeans.inertia_)
```

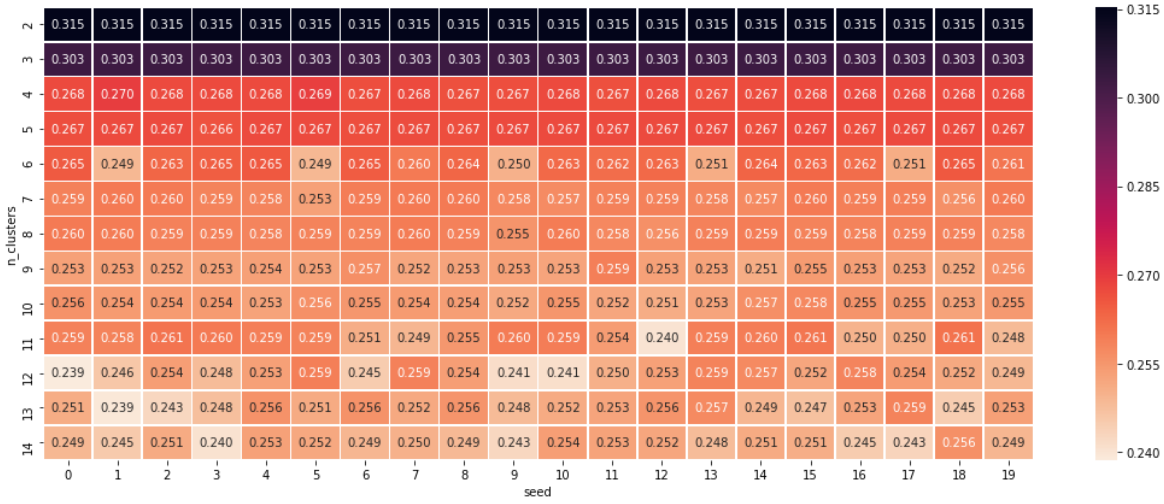
```
plt.figure()
plt.plot(clusters_range,inertias, marker='o')
```

Out[24]:

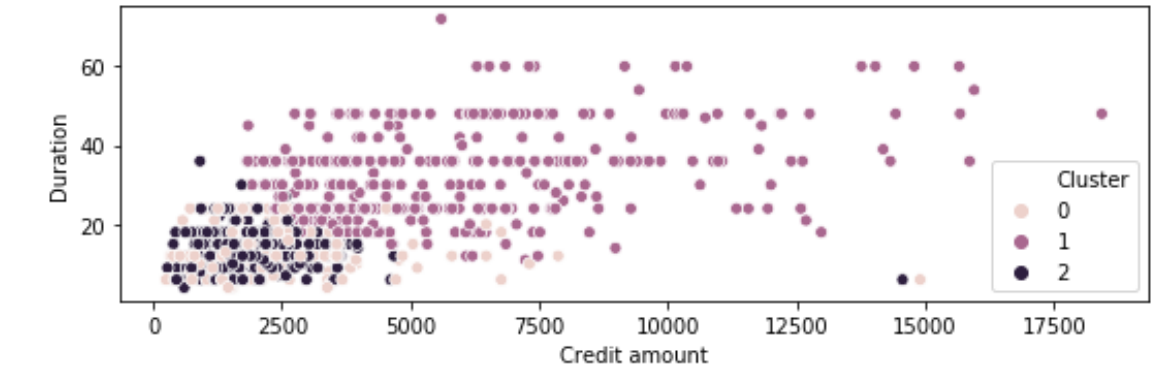
```
[<matplotlib.lines.Line2D at 0x2057bf89438>]
```

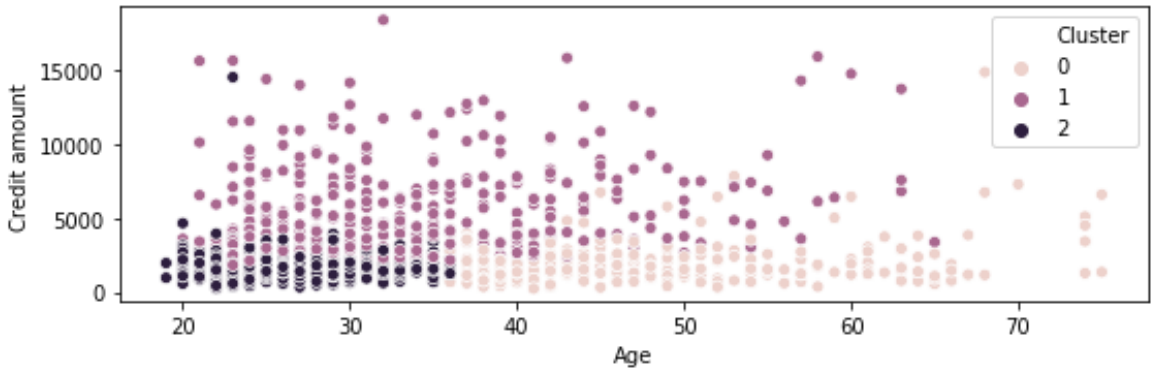


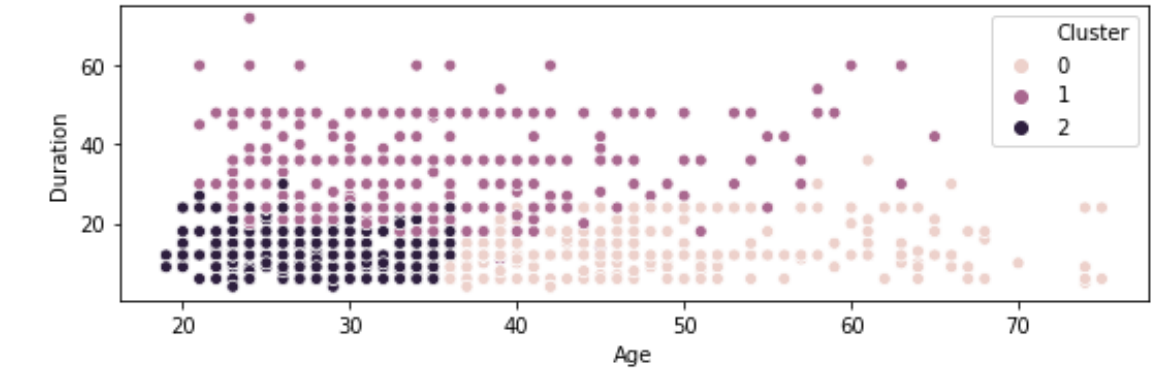

```
from sklearn.metrics import silhouette_samples, silhouette_score

clusters_range = range(2,15)
random_range = range(0,20)
results =[]
for c in clusters_range:
  for r in random_range:
    clusterer = KMeans(n_clusters=c, random_state=r)
    cluster_labels = clusterer.fit_predict(cluster_scaled)
    silhouette_avg = silhouette_score(cluster_scaled, cluster_labels)
    #print("For n_clusters =", c," and seed =", r,  "\nThe average silhouette_score is :", silhouette_avg)
    results.append([c,r,silhouette_avg])

result = pd.DataFrame(results, columns=["n_clusters","seed","silhouette_score"])
pivot_km = pd.pivot_table(result, index="n_clusters", columns="seed",values="silhouette_score")

plt.figure(figsize=(15,6))
sns.heatmap(pivot_km, annot=True, linewidths=.5, fmt='.3f', cmap=sns.cm.rocket_r)
plt.tight_layout()
```




```
kmeans_sel = KMeans(n_clusters=3, random_state=1).fit(cluster_scaled)
labels = pd.DataFrame(kmeans_sel.labels_)
clustered_data = cluster_data.assign(Cluster=labels)


scatters(clustered_data, 'Cluster')
```








```
grouped_km = clustered_data.groupby(['Cluster']).mean().round(1)
grouped_km
```

Cluster	Age	Credit amount	Duration
0	48.6	1970.5	13.9
1	34.0	5665.4	32.2
2	27.7	1737.5	14.3

5- METODOS PREDICTIVOS CON PYTHON

RANDOM FOREST

```
import sklearn as sk


from sklearn import model_selection

X_train, X_test, y_train, y_test = sk.model_selection.train_test_split(X, y, test_size=0.2, random_state=0)


from sklearn.ensemble import RandomForestClassifier

for i in [100, 1000, 2000]:
  clf = sk.ensemble.RandomForestClassifier(n_estimators=i)

  _ = clf.fit(X_train, y_train)
  y_pred = clf.predict(X_test)
  score = accuracy_score(y_pred, y_test)
  print(i, "exactitud:", score)
```

Out:

```
100 exactitud: 0.76
1000 exactitud: 0.755
2000 exactitud: 0.76
```

```
for max_depth in range(5, 30, 5):
  for n_estimators in range(1000, 5000, 500):
    clf = sk.ensemble.RandomForestClassifier(n_estimators=n_estimators, max_depth=max_depth)

    scores = sk.model_selection.cross_val_score(clf, X, y, scoring="roc_auc", cv=5, n_jobs=-1)
    print(n_estimators, max_depth, "exactitud:", scores.mean(), scores.std())
```

Out:

```
1000 5 exactitud: 0.7629523809523809 0.02114392961672895
1500 5 exactitud: 0.7639523809523809 0.022063400707815924
2000 5 exactitud: 0.7623571428571428 0.02340833427176489
2500 5 exactitud: 0.7637857142857143 0.021730205108534052
3000 5 exactitud: 0.7634285714285713 0.022031388657272652
```

```
3500 5 exactitud: 0.7633809523809523 0.0225787360081979
4000 5 exactitud: 0.7635714285714286 0.021998891953131056
4500 5 exactitud: 0.7632142857142858 0.022426119570765454
1000 10 exactitud: 0.7578571428571428 0.018236136908413594
1500 10 exactitud: 0.7587380952380952 0.020666831247973087
2000 10 exactitud: 0.7571190476190475 0.01779614980398302
2500 10 exactitud: 0.7553095238095239 0.018233121291159685
3000 10 exactitud: 0.7585000000000001 0.019857285599578285
3500 10 exactitud: 0.7576666666666667 0.01761109859295534
4000 10 exactitud: 0.7585238095238094 0.01866393323038097
4500 10 exactitud: 0.7577619047619049 0.01893783840241561
1000 15 exactitud: 0.7532857142857143 0.019512569354438982
1500 15 exactitud: 0.7511428571428572 0.01948204010419119
2000 15 exactitud: 0.7511904761904762 0.017938702632971922
2500 15 exactitud: 0.7516904761904761 0.01904187413364404
3000 15 exactitud: 0.7538333333333334 0.0191329043067403
3500 15 exactitud: 0.7533333333333333 0.020943585816379968
4000 15 exactitud: 0.7525119047619048 0.019332116428978568
4500 15 exactitud: 0.7522142857142857 0.020815652361142918
1000 20 exactitud: 0.7489047619047619 0.01930883387881404
1500 20 exactitud: 0.751107142857143 0.017884962456940413
2000 20 exactitud: 0.7498452380952381 0.01871580782614087
2500 20 exactitud: 0.7505119047619047 0.018873069272489048
3000 20 exactitud: 0.7523452380952381 0.01940945750958351
3500 20 exactitud: 0.75125 0.0192820396264035
4000 20 exactitud: 0.7518690476190477 0.01746504445393079
4500 20 exactitud: 0.7515833333333333 0.018262758408061
1000 25 exactitud: 0.7506904761904762 0.019585942075481748
1500 25 exactitud: 0.7508095238095238 0.019117453826708385
2000 25 exactitud: 0.7500357142857144 0.01934960024079203
2500 25 exactitud: 0.752 0.01959281504431011
3000 25 exactitud: 0.7517023809523808 0.01887400791043531
3500 25 exactitud: 0.7523809523809524 0.01827053251212574
4000 25 exactitud: 0.7511071428571429 0.01848938685581153
4500 25 exactitud: 0.7517142857142857 0.018310259134407682
```

print(confusion_matrix(y_test, y_pred))
print("\n")

Out:

```
[[ 28  30]
 [ 18 124]]
```

```
from sklearn.metrics import roc_auc_score
roc_auc = roc_auc_score(y_test, y_pred)
print('Curva ROC - AUC del modelo:')
print(roc_auc)
plt.show(roc_auc)
```

Out:

```
Curva ROC - AUC del modelo: 0.6779990286546866
```

REGRESION LOGISTICA

```
from sklearn import datasets

from sklearn.preprocessing import StandardScaler

escalar = StandardScaler()

X_train = escalar.fit_transform(X_train)
X_test = escalar.transform(X_test)


from sklearn.linear_model import LogisticRegression
algoritmo = LogisticRegression()


algoritmo.fit(X_train, y_train)
                                                                    Out[65]:
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=Tr
ue,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)


y_pred = algoritmo.predict(X_test)


from sklearn.metrics import precision_score


precision = precision_score(y_test, y_pred)
print('Precisión del modelo:')
```

```
print(precision)
Precisión del modelo:
0.7741935483870968
```

```
from sklearn.metrics import accuracy_score
exactitud = accuracy_score(y_test, y_pred)
print('Exactitud del modelo:')
print(exactitud)
```

```
Exactitud del modelo:
0.715
```

```
print(confusion_matrix(y_test, y_pred))
print("\n")
```

```
[[ 23  35]
 [ 22 120]]
```

```
from sklearn.metrics import roc_auc_score
roc_auc = roc_auc_score(y_test, y_pred)
print('Curva ROC - AUC del modelo:')
print(roc_auc)
plt.show(roc_auc)
```

```
Curva ROC - AUC del modelo:
0.6208110733365712
```

DECISION TREE

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)

feature_cols = ['Age', 'Sex', 'Job', 'Housing','Saving accounts','Checking account','Checking account',
        'Credit amount','Duration','Purpose']
X = data[feature_cols]
y = data.Risk
```

Creación del Árbol de Decisión

```
clf = DecisionTreeClassifier()
```

Entrenamiento

```
clf = clf.fit(X_train,y_train)
```

Predicción

```
y_pred = clf.predict(X_test)
```

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.685
```

```
from sklearn.metrics import accuracy_score
exactitud = accuracy_score(y_test, y_pred)
print('Exactitud del modelo:')
print(exactitud)
```

```
Exactitud del modelo:
0.665
```

```
print(confusion_matrix(y_test, y_pred))
print("\n")
```

```
[[ 27  32]
 [ 35 106]]
```

```
from sklearn.metrics import roc_auc_score
roc_auc = roc_auc_score(y_test, y_pred)
print('Curva ROC - AUC del modelo:')
print(roc_auc)
plt.show()
```

```
Curva ROC - AUC del modelo:
0.604700084144729
```