

**Estatus tarjetas de crédito a partir de datos personales**

**CoderHouse Data Science**

**12 de Enero 2023**

Marcelo Baeza

Luis Salazar

Lucas Ferrer

Tabla de contenido

[Descripción del caso de negocio](#_heading=h.30j0zll) **3**

[Tabla de versionado](#_heading=h.1fob9te) **3**

[Objetivo](#_heading=h.2et92p0) **3**

[EDA: Exploratory Data Analysis](#_heading=h.3dy6vkm) **4**

[I. Descripción de los datos.](#_heading=h.rpjsmtl6ln49) 4

[II. Limpieza de los datos.](#_heading=h.27uvzi55wvne) 5

[III. Análisis univariado.](#_heading=h.bsy24hcrj4ej) 6

[IV. Análisis bivariado.](#_heading=h.suzu93lpupek) 7

[Relación de estado civil o aprobación de tarjeta de crédito](#_heading=h.1i2ge05mmp68) 9

[Relación de sueldo v/s el número de personas que viven en el hogar](#_heading=h.r0wzmw4dezig) 9

[Tipo de ingreso por género](#_heading=h.x3pyq5odevor) 11

[Distribución de Edades](#_heading=h.3ig21yzfzl7l) 12

[¿Cuáles edades y sexos presentan mayores ingresos?](#_heading=h.wxmz8vfl1wkh) 12

[El tipo de hogar que viven los solicitantes](#_heading=h.osbfje6zm4za) 13

[Análisis Multivariado](#_heading=h.bq47j8qahunu) 13

[¿Cuál es el nivel de correlación entre las variables?](#_heading=h.5hkan8durc0) 13

[Algoritmo Elegido](#_heading=h.1t3h5sf) **14**

[Árbol de decisiones](#_heading=h.oe2bt35qwoi3) 14

[Comparación de Modelos de Machine Learning.](#_heading=h.489srt7tp3gc) 17

[PCA](#_heading=h.h7j877bajo1v) 17

[Trabajando GRID](#_heading=h.8r1z5yyq5wc6) 17

[Operando sobre los 3 modelos](#_heading=h.6bjdwajy59p1) 18

[Bias-Variance Trade Off](#_heading=h.x8tebbtiz6j) 20

[Conclusiones:](#_heading=h.4d34og8) **20**

# Descripción del caso de negocio

Las tarjetas de crédito son un método común de control de riesgos en la industria financiera. Utiliza la información personal y los datos presentados por los solicitantes de tarjetas de crédito para predecir la probabilidad de futuros incumplimientos y préstamos de tarjetas de crédito. Así mismo, el banco puede decidir si emite una tarjeta de crédito al solicitante. En términos generales, los modelos se basan en el análisis de datos históricos y la formulación de modelos de predicción.

Por lo tanto un banco ha solicitado ayuda que les permita identificar patrones en los clientes, para la toma de decisiones sobre a quiénes otorgar una tarjeta de crédito.

# Tabla de versionado

Sólo existe una versión.

# Objetivo

El proyecto está orientado a un banco que pretende evaluar y diferenciar a los clientes, ya sea por capacidad económica, capacidad de devolución, compromiso con las deudas, etc. Si bien este trabajo fue realizado con el dataset de un banco en específico y no se puede difundir, otros bancos podrían usar el mismo modelo para realizar el análisis con sus propios clientes.

Al tener el historial de todos los clientes se puede analizar cuáles son los clientes más confiables al momento de aceptar la solicitud de una tarjeta de crédito o al momento de darles un préstamo. Para esto el banco puede utilizar el historial de deudas de los clientes, así como el tiempo que tardaron en pagarlas. Esto a la larga generará una mayor eficiencia y menor pérdida al banco. Para ellos se realizarán las siguientes tareas:

1. Leer, transformar y preparar datos.
2. Analizar y construir visualizaciones de los datos para identificar patrones en estos.
3. Aplicar algoritmos de machine learning para entrenar modelos predicción e identificar el mejor modelo aplicable.

Así mismo se responderán consultas relativas a la naturaleza de los datos, entre las cuales se encuentran:

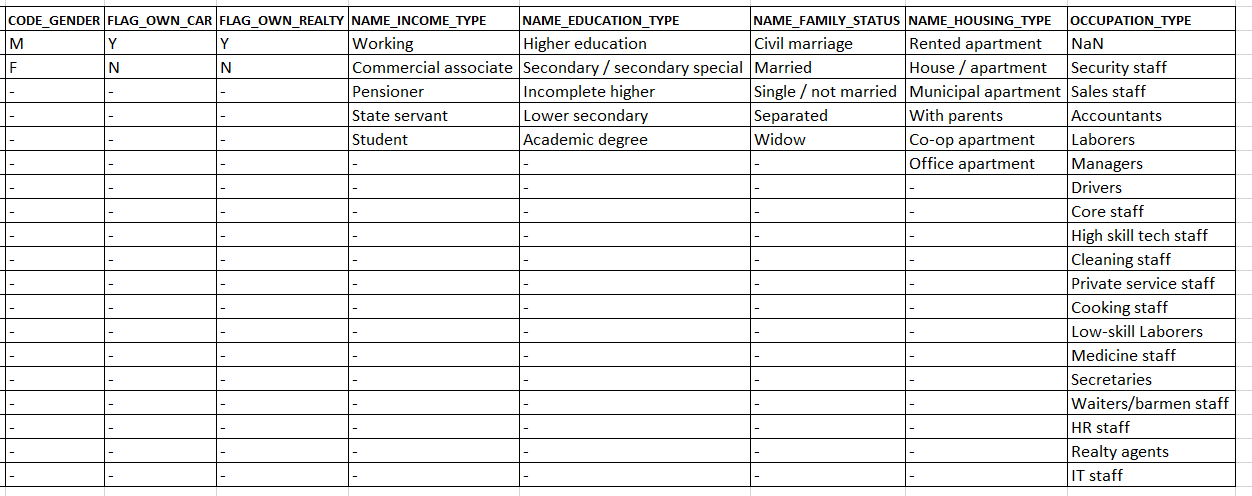
1. ¿Cuál es la correlación entre el estado civil de los clientes y la tasa de aprobación de tarjetas de crédito?
2. ¿Cuál es la distribución entre los niveles académicos de los clientes a los que se les aprobó una tarjeta de crédito?
3. ¿Cuál es la correlación entre la cantidad de miembros en la familia de los clientes y los ingresos anuales?
4. ¿Cuál es la correlación entre las profesiones de los clientes y la cantidad de solicitudes de tarjetas de crédito, teniendo en cuenta el género de los clientes?

# EDA: Exploratory Data Analysis

## Descripción de los datos.

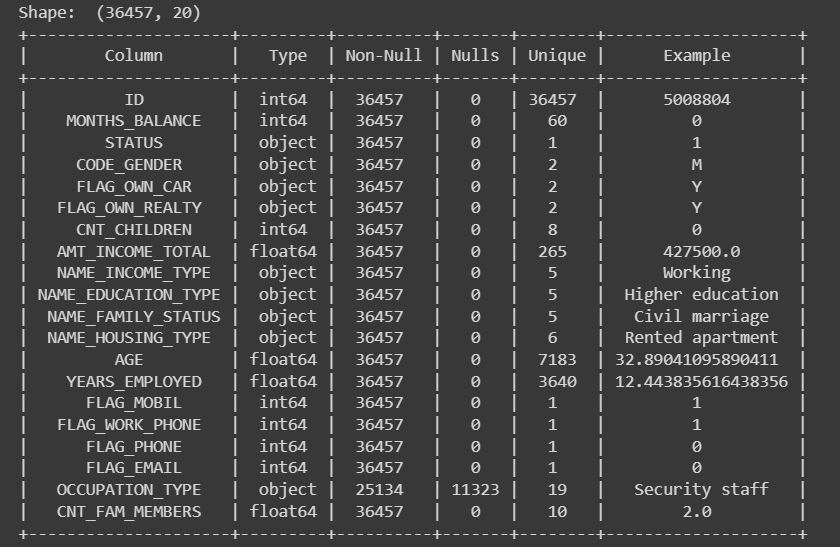
| **Variable** | **Descripción** |
| --- | --- |
| CODE\_GENDER | Indica el género, que será considerado para realizar la evaluación del estatus. |
| FLAG\_OWN\_CAR | ¿Posee carro? que será considerado para realizar la evaluación del estatus. |
| FLAG\_OWN\_REALTY | ¿Posee bienes raíces? que será considerado para realizar la evaluación del estatus. |
| CNT\_CHILDREN | Número de hijos, que será considerado para realizar la evaluación del estatus. |
| AMT\_INCOME\_TOTAL | Ingreso anual, que será considerado para realizar la evaluación del estatus. |
| NAME\_INCOME\_TYPE | Categoría de ingresos, que será considerado para realizar la evaluación del estatus. |
| NAME\_EDUCATION\_TYPE | Nivel educativo, que será considerado para realizar la evaluación del estatus. |
| NAME\_FAMILY\_STATUS | Estado civil, que será considerado para realizar la evaluación del estatus. |
| NAME\_HOUSING\_TYPE | Tipo de vivienda, que será considerado para realizar la evaluación del estatus. |
| OCCUPATION\_TYPE | Ocupación, que será considerado para realizar la evaluación del estatus. |
| CNT\_FAM\_MEMBERS | Tamaño de la Familia, que será considerado para realizar la evaluación del estatus. |
| MONTHS\_BALANCE | Balance mensual |
| STATUS | Estatus de deudas |

Algunas variables categóricas han sido relevadas como niveles, cuyos significados son los siguientes:

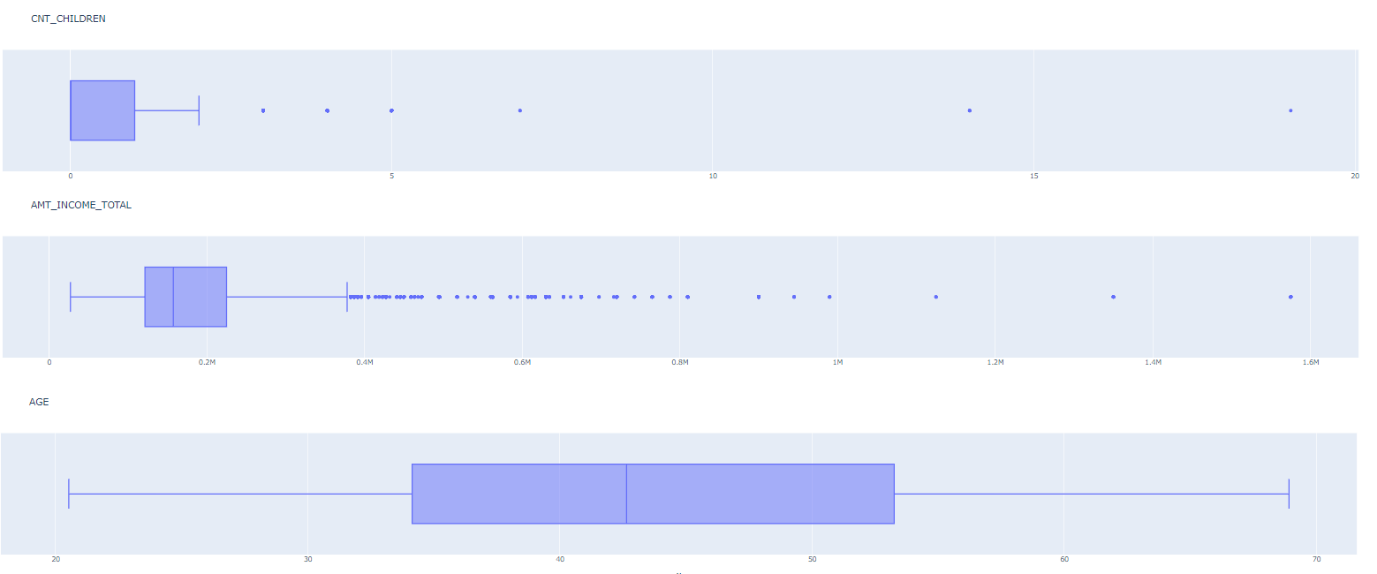


## Limpieza de los datos.

El dataset contiene 20 variables con 36457 registros, entre las que se encuentra la única variable que presenta datos nulos es OCCUPATION\_TYPE .



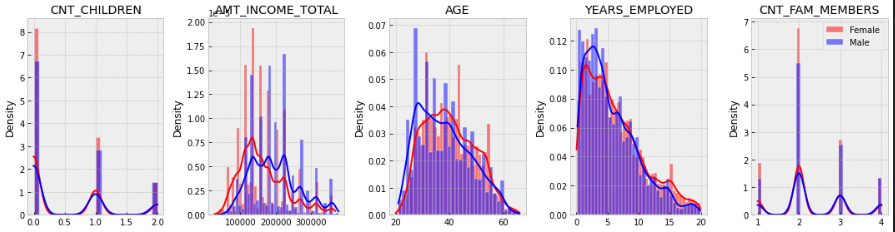
Dichos datos nulos fueron eliminados y se realizó la reducción de la dimensionalidad del dataset basado en análisis de outliers de los variables numéricas.



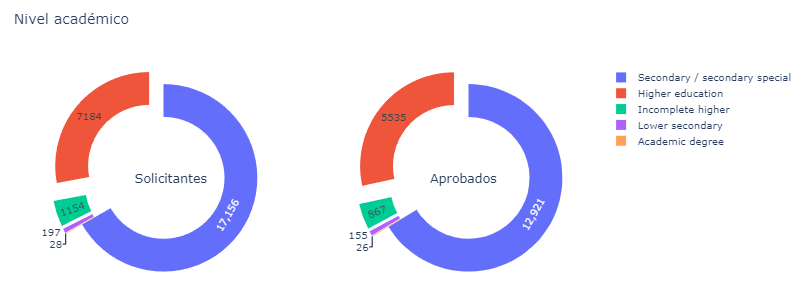
Se identificaron 9942 que fueron eliminados.

## Análisis univariado.

En el análisis exploratorio se estudiaron las distribuciones de las variables numéricas originales y de las variables categóricas. Entre los resultados que se destacan, se observó que no hay diferencias importantes en las distribuciones de las variables ordinales debidas al género de los individuos.

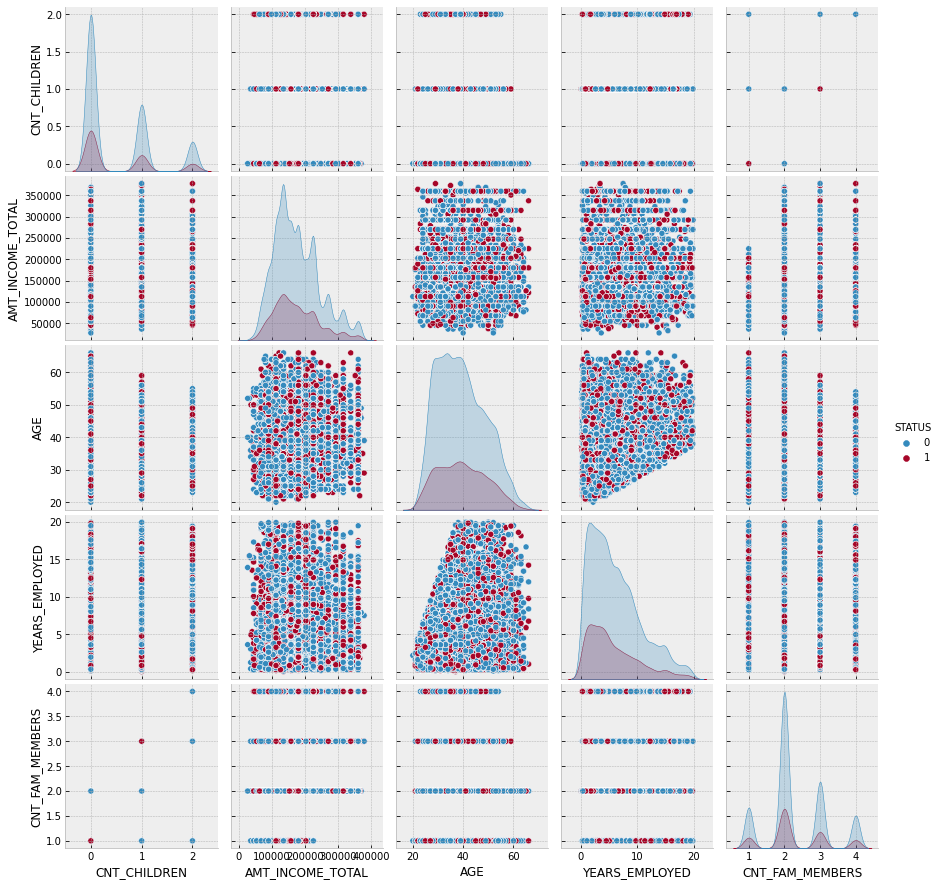


Así mismo, se observa que más de la mitad cuenta con la secundaria completa y más de un cuarto cuenta con educación superior completa. Con lo cual mas del 95% cuenta con nivel educativo alto o medio y no se observa incidencia del nivel educativo en el otorgamiento de TDC mas del 95% cuentan nivel educativo alto o medio y no se observa incidencia del nivel educativo en el otorgamiento de dicha tarjeta.



## Análisis bivariado.

A continuación se presentan las relaciones que existen entre dos variables considerando: cantidad de niños, el ingreso mensual, la edad, los años trabajados y el número de miembros de la familia.

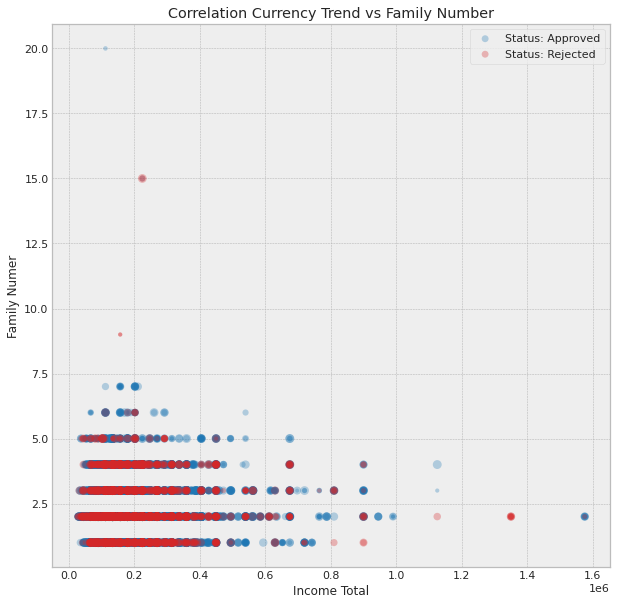


### Relación de estado civil o aprobación de tarjeta de crédito

#### 

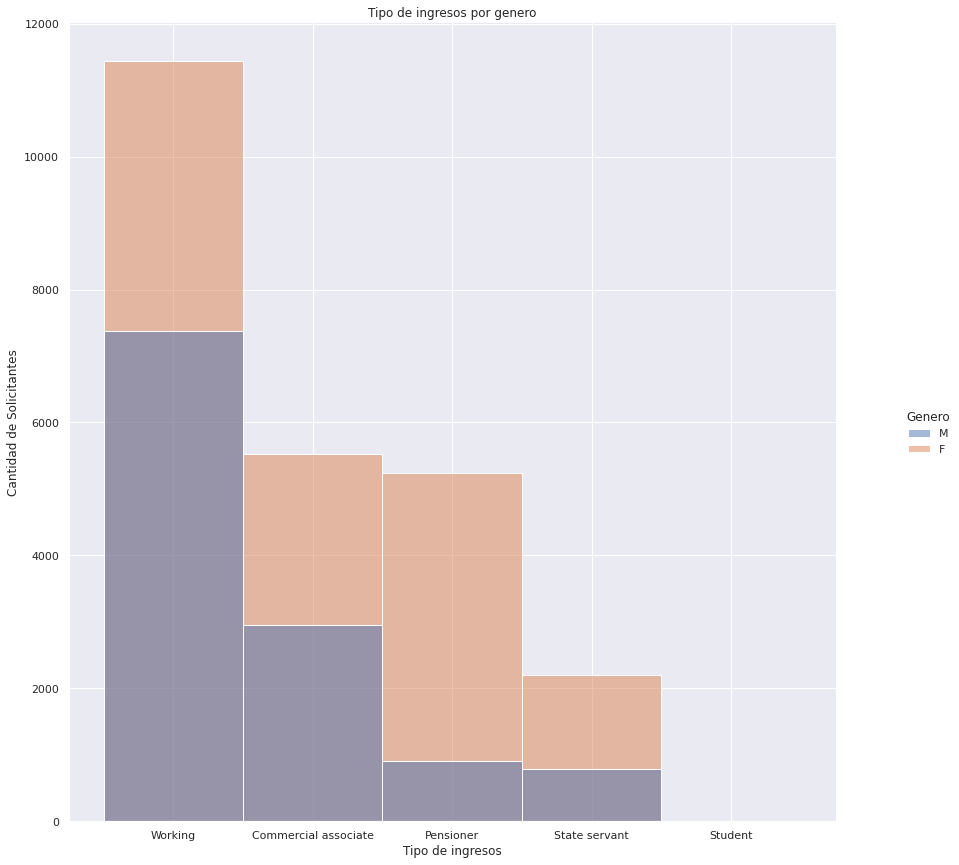
De este gráfico es posible interpretar que existe una gran afluencia de las personas que están casadas y que desean poseer una tarjeta de crédito; pero además no es posible concluir que existe alguna tendencia respecto al estado civil de las personas que indique que se le otorgará con mayor probabilidad su tarjeta de crédito

### Relación de sueldo v/s el número de personas que viven en el hogar



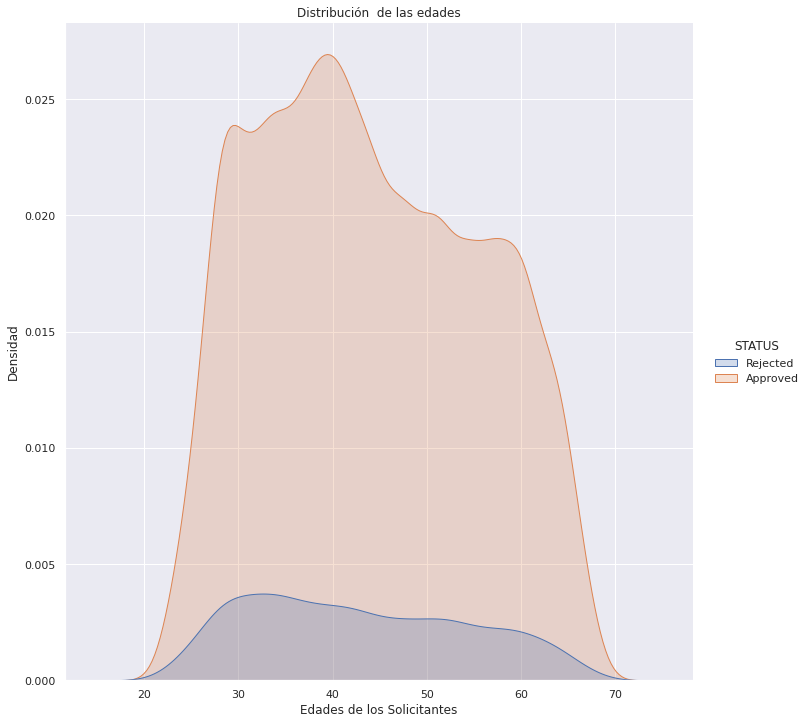
Respecto a la gráfica presentada, es posible concluir que la mayor cantidad de personas que solicitan una tarjeta de crédito tienen un salario anual máximo hasta los 500.000 USD. Además tampoco existe una relación directa entre los miembros familiares vs el salario. Solo se aprecia que dentro del rango especificado predomina el rechazo a la solicitud de tarjeta de crédito.

### Tipo de ingreso por género



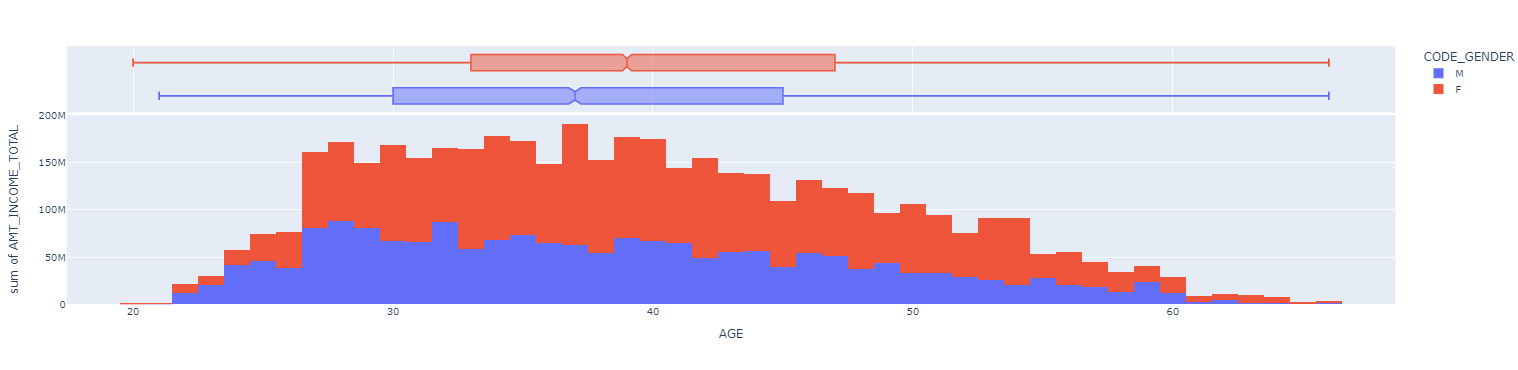
Con respecto al total de los solicitantes, se observa que la mayoría tiene sus ingresos en empleos privados en condición de dependencia, teniendo una mayor incidencia en este grupo los hombres. El segundo grupo corresponde a ingresos provenientes de negocios o inversiones, al igual que el caso anterior los hombres tienen mayor participación. Finalmente, están los pensionados y empleados públicos donde las mujeres tienen mayor incidencia.

### Distribución de Edades



Se observa que la mayoría de solicitantes y aprobados tiene edades entre 35 y 45 años. También es claro que ambas distribuciones tienen un comportamiento similar con lo cual se concluye que la edad no es factor que tuvo impacto en las aprobaciones. Finalmente, otro dato de relevancia es que la mayor parte de solicitantes y aprobados tiene edades por debajo de 50 años.

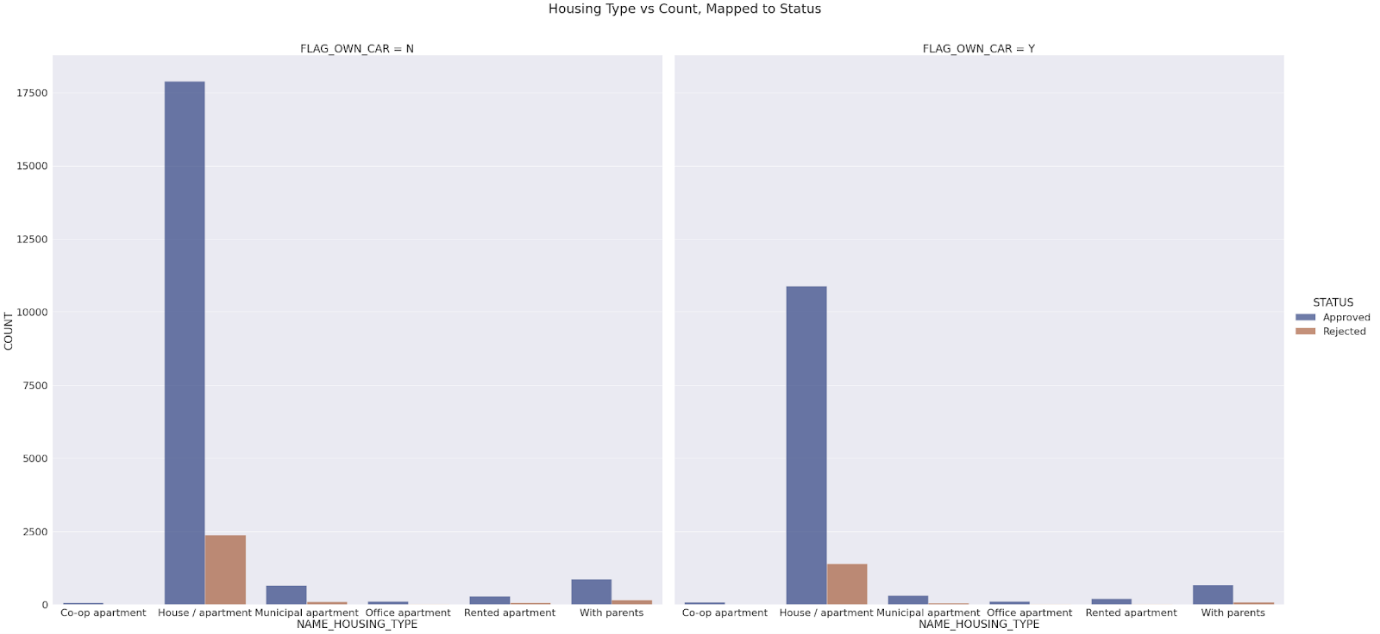
### ¿Cuáles edades y sexos presentan mayores ingresos?



En general el nivel de ingreso de las mujeres es mayor que el de los hombres y en ambos sexos dicho nivel disminuye a medida que avanza la edad. Lo cual se relaciona con el gráfico anterior, donde los más jóvenes cuentan con una mayor tasa de aprobación.

### El tipo de hogar que viven los solicitantes

Esta información es para conocer el número de participantes que se les ha asignado una tarjeta de crédito, relacionada al tipo de “hogar” donde viven.



Para el análisis de la gráfica en donde se presentan dos escenarios: las personas que poseen o no un vehículo, y a la vez comparando su situación respecto al tipo de casa en que vive. La predominancia en ambas gráficas es para personas que poseen una casa o departamento. Además concluir que no existe una tendencia respecto a qué personas se le asigna una tarjeta o quienes no.

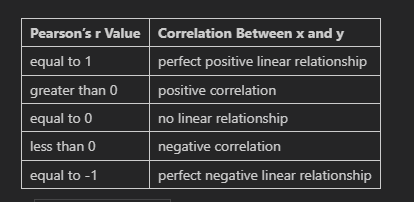
## Análisis Multivariado

Para esta ocasión se desea revisar la multi relación que podrían presentar las variables respondiendo la pregunta:

### ¿Cuál es el nivel de correlación entre las variables?

| MONTHS\_BALANCE | 0.108973 |
| --- | --- |
| AMT\_INCOME\_TOTALxFamilymember | 0.005676 |
| CODE\_GENDER | -0.004718 |
| FLAG\_OWN\_CAR | 0.010172 |
| FLAG\_OWN\_REALTY | -0.025758 |
| CNT\_CHILDREN | -0.008741 |
| AMT\_INCOME\_TOTAL | -0.003468 |
| NAME\_INCOME\_TYPE | -0.009521 |
| NAME\_EDUCATION\_TYPE | 0.011216 |
| NAME\_FAMILY\_STATUS | 0.003670 |
| NAME\_HOUSING\_TYPE | 0.004042 |
| AGE | 0.012067 |
| YEARS\_EMPLOYED | -0.035856 |
| FLAG\_MOBIL | NaN |
| FLAG\_WORK\_PHONE | -0.016932 |
| FLAG\_PHONE | -0.006743 |
| FLAG\_EMAIL | 0.017049 |
| OCCUPATION\_TYPE | 0.001093 |
| CNT\_FAM\_MEMBERS | -0.006830 |

Donde las características de relación se distribuyen:



Permitiendo así:

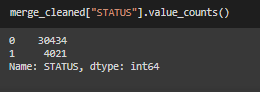
1. No es posible visualizar una relación directa considerable entre las variables, ya que él correlativo más alto es el sueldo con el 0.1
2. Continuando así el nivel educacional que de igual forma sigue siendo bajo.

# Algoritmo Elegido

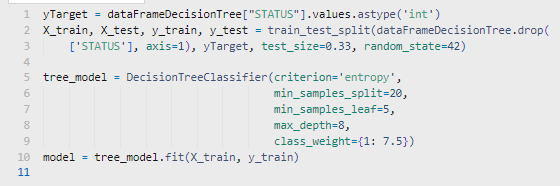
A continuación, se expondrán algoritmos con los que se trabajaron para saber si es posible es posible clasificar a los candidatos

### Árbol de decisiones

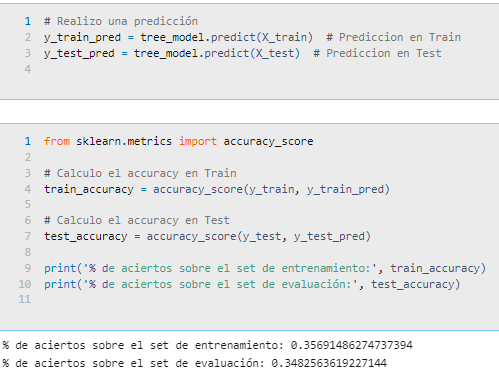
En base a esta información es posible verificar que existe una diferencia de 7.5 entre la cantidad de muestra entre el estatus 0 y el 1. Esto es necesario considerar porque se generará un peso de esta diferencia para la clase 1 en el árbol de decisión.



Se buscó la mejor profundidad y en base al algoritmo

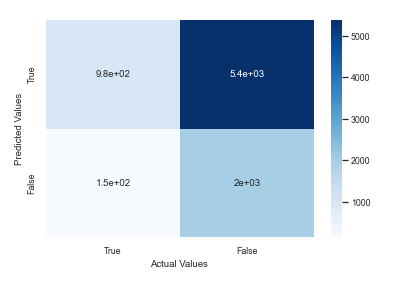


Dando como resultado:



Permitiendo tener el siguiente resultado: [link](https://drive.google.com/file/d/1HqS-kQH_FlqvSB1cy3SgcSfa719EBDdL/view?usp=share_link)

Matriz de confusión



Métricas:

* **Precision**: 0.2686040195545899.
* **Recall**: 0.931261770244821.
* **Specify**: 0.1536769327467002.
* **Accuracy**: 0.3482563619227144.
* **F1\_Score**: 0.41694772344013487.

En base a los valores numéricos se concluye:

1. Posee una precisión bastante baja, por lo que es posible comprender que posee una gran dispersión de los datos. Y además es necesario comprobar que considerar casos positivos, el status 0 o 1.
2. El recall de la muestra indica que no es tan eficaz para identificar correctamente a las personas que se le otorgó el crédito.
3. Respecto al accuracy, ya contempla aumentar, y nos permite saber que está cercano a una estimación "positiva".
4. Al ser medianamente alta la especificidad nos permite entender que el algoritmo es factible para identificar los casos negativos.
5. Para el caso del f1-score, es posible decir que el modelo no logra clasificar correctamente.

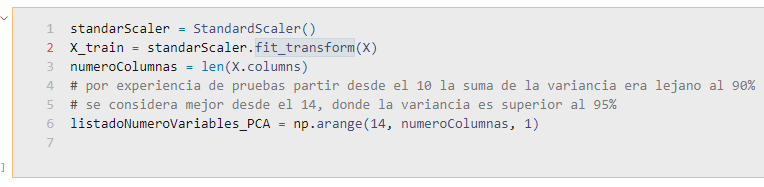
**Para la medida adicional que se obtiene a través del reporte de clasificación de skelearn:**

1. Es posible verificar que el modelo permite verificar la etiqueta 0, que se le otorgará el crédito eficazmente.
2. El modelo tiene un déficit al momento de clasificar la etiqueta 1, que no se le otorgará el crédito.
3. Esto se puede deber al soporte que posee, osea la cantidad de elementos que se usaron para construir el modelo.

### Comparación de Modelos de Machine Learning.

Para la comparación los algoritmos de machine learning es necesario considerar el PCA que se trabajó:

#### PCA



En donde por experiencia el mejor valor de solución fue cuando la dimensión del dataset desde 14.

Por experiencia y prueba los modelos que tuvieron que ser borrados por que tuvieron mal performance fueron:

* LogisticRegression
* SVC
* GaussianNB
* SGDClassifier



Se realizó la comparación de los siguientes modelos por medio de validación cruzada con la ayuda de las librerías Sklearn, StratifiedKFold. Teniendo como [**resultado**](https://drive.google.com/file/d/1I-Q5IGn-g64ScFWDkufvX_Khbk2WtAJN/view?usp=share_link)

Permitiendo obtener los mejores 3 modelos en base a la medida roc\_auc, permitiendo diferenciar entre dos clases (aprobado o rechazado).

Los resultados obtenidos son siguientes, evaluando el área bajo la curva.

#### Trabajando GRID

Además se aplica una lógica de “Grid” con la idea de operar sobre los mejores algoritmos PARAMETRIZANDOLOS y buscar su mejor rendimiento. Para tal caso se siguió la siguiente lógica:

1. Generando grid genéricos por indíce de modelos.
2. StratifieldKold, ya que permite trabajar con iguales etiquetas, que es necesario porque no existe igualdad entre estas dos
3. Se trabaja con la mejor selección de modelo: RandomizedSearchCV.
4. Se trabaja con los 3 mejores modelos de la selección anterior.
5. Se registra el score de las ejecuciones (kfold) con el propósito de trabajar con las 3 mejores combinaciones.

Teniendo como resultado:

* [Fold 1/10]
  + Score: 0.6798037424137752. Tiempo ejecución 253.02872586250305 segundos
* [Fold 2/10]
  + Score: 0.6717604663696969. Tiempo ejecución 235.69638681411743 segundos
* [Fold 3/10]
  + Score: 0.6796626033472244. Tiempo ejecución 267.5591015815735 segundos
* [Fold 4/10]
  + Score: 0.6793060414948856. Tiempo ejecución 251.11597061157227 segundos
* [Fold 5/10]
  + Score: 0.6867359221700058. Tiempo ejecución 251.85966444015503 segundos
* [Fold 6/10]
  + Score: 0.6656364910544975. Tiempo ejecución 255.90586876869202 segundos
* [Fold 7/10]
  + Score: 0.6863809877153928. Tiempo ejecución 244.39102745056152 segundos
* [Fold 8/10]
  + Score: 0.6586012861736334. Tiempo ejecución 261.63546538352966 segundos
* [Fold 9/10]
  + Score: 0.6862375298870476. Tiempo ejecución 245.60595989227295 segundos
* [Fold 10/10]
  + Score: 0.6761290722160287. Tiempo ejecución 260.96279764175415 segundos
* Avg scores - 0.6770254142842187

#### Operando sobre los 3 modelos

Al ya tener modelos con una performance que permita asignar a un candidato con una mayor probabilidad al 50%. Se trabajó considerando el rendimiento de tres algoritmos:

* RandomForestClassifier
  + Parametría:
  + 'random\_state': 246
  + ‘'n\_jobs': 4
  + 'n\_estimators': 150,
  + 'min\_samples\_split': 5
  + 'max\_depth': 14
  + ‘criterion': 'entropy'
  + ‘'class\_weight': 'balanced\_subsample',
  + 'bootstrap': True
* RandomForestClassifier
  + Parametría:
  + 'random\_state': 246
  + 'n\_jobs': 4
  + 'n\_estimators': 200
  + 'min\_samples\_split': 10,
  + 'max\_depth': 16
  + 'criterion': 'log\_loss'
  + 'class\_weight': 'balanced\_subsample'
  + 'bootstrap': True
* RandomForestClassifier
  + Parametría:
  + ‘'random\_state': 61
  + 'n\_jobs': 4
  + 'n\_estimators': 200
  + 'min\_samples\_split': 10
  + 'max\_depth': 14
  + 'criterion': 'log\_loss'
  + 'class\_weight': 'balanced'
  + 'bootstrap': True

Teniendo como resultado

* Modelo 1
  + Mean: 0.4177842061914342
  + std: 0.15728666599892036
  + Variance: 0.02473909530105593
* Modelo 2
  + Mean: 0.39911369195997526
  + std: 0.18000259380793657
  + Variance: 0.032400933777585
* Modelo 3
  + Mean: 0.418842534017898
  + std: 0.15490990647930442
  + Variance: 0.023997079125426844
* Apreciando que en comparativa de los podamos existen 3 que resalta a la hora de efectuar una mejor categorización.
* En los 3 de los modelos existe una baja desviación estándar, permitiendo entender que no existe una gran dispersión de los datos, por lo que los modelos clasificaron de una manera uniforme, sin tanta diferencia.
* A pesar de poseer una desviación estándar el promedio de categorización es bajo. Por lo que la probabilidad de etiquetar correctamente será baja
* Respecto a la probabilidad promediada de donde se presenta anteriormente, la probabilidad mínima es muy baja para considerar en un futuro, y el 75% supera el 13% lo cual sigue siendo bajo. Pero se mantiene la desviación estable.

#### Bias-Variance Trade Off

Además se decidió trabajar con Bias-Variance con la idea de mejorar el rendimiento de los modelos. Teniendo así un poco bias y poca varianza, poseyendo un modelo robusto. Y como resultado:

* Modelo 1
* Tiempo: 338.7708840370178 segundos.
* MSE: 0.309
* Bias: 0.223
* Variance: 0.086
* Modelo 2
* Tiempo 448.3412914276123 segundos.
* MSE: 0.298
* Bias: 0.218
* Variance: 0.080
* Modelo 3
* Tiempo 353.31135749816895 segundos.
* MSE: 0.310
* Bias: 0.223
* Variance: 0.086

# Conclusiones:

Con respecto a los datos disponibles se observa que no hay diferencias importantes en las distribuciones de las variables ordinales debidas al género de los individuos. Así mismo, mas del 95% cuenta con nivel educativo alto o medio y no se observa incidencia del nivel educativo, el estado civil o edad en el otorgamiento de TDC.

Se observa que la mayoría de solicitantes y aprobados tiene edades entre 35 y 45 años. Otro dato de relevancia es que la mayor parte de solicitantes y aprobados tiene edades por debajo de 50 años.

Por otro lado, el análisis multivariado no arrojó fuerte correlación (cerca a 1) de alguna variable con respecto al status de aprobación (variable independiente) por si sola.

Inicialmente se aplicó el algoritmo de árbol decisiones. Sin embargo, las métricas indican una gran dispersión de los datos por lo cual que no es tan eficaz para identificar correctamente a las personas que se le otorgó el crédito, aunque al ser medianamente alta la especificidad nos permite entender que el algoritmo es factible para identificar los casos negativos.

Luego de aplicar validación cruzada, se realizó la comparación de diferentes modelos por medio de validación cruzada con la ayuda de las librerías Sklearn, StratifiedKFold. Permitiendo obtener los mejores 3 modelos en base a la medida roc\_auc, permitiendo diferenciar entre dos clases (aprobado o rechazado). Todos estos modelos asociados al algoritmo RandomForestClassifier con diferentes parámetros. Sin embargo, la probabilidad de etiquetar correctamente al usar dichos modelos será baja.