

**Ingeniería Civil Informática**

Facultad de ingeniería

Ciencia de Datos

**Laboratorio Nº1**

**Presentado a:**

Billy Peralta

**Presentado por:**

Gabriel González Cancino

Javier Briones Cañas

Luis Rodriguez Zamora

**Fecha:**

Junio, 2020

# ´Indice

[´Indice 2](#_Toc42370744)

[1. Definición de proyecto 2](#_Toc42370745)

[2. Conformación de los datos 3](#_Toc42370746)

[2.1. Preprocesamiento de Datos 4](#_Toc42370747)

[2.2. Análisis Exploratorio 5](#_Toc42370748)

[2.3. Diseño de Experimentos 9](#_Toc42370749)

[3. Implementación de los algoritmos básicos 9](#_Toc42370750)

[3.1. Revisión de métodos relacionados con “Credit Card Approval” 9](#_Toc42370751)

[3.2. Implementación del algoritmo básico 12](#_Toc42370752)

[3.3. Resultados Obtenidos 13](#_Toc42370753)

[4. Resultados innovación en algoritmos 16](#_Toc42370754)

[4.1. Describir Algoritmos innovadores 16](#_Toc42370755)

[4.2. Implementar Algoritmo Innovador 16](#_Toc42370756)

[4.3. Resultados Obtenidos 16](#_Toc42370757)

[5. Conclusi´on 16](#_Toc42370758)

[Bibliografía 16](#_Toc42370759)

# Definición de proyecto

El laboratorio consiste utilizar distintos conceptos de la Ciencia de Datos para predecir si una persona paga un crédito o no.

Para llevar a cabo esto se pasará por tres etapas, que son, Análisis y preprocesamiento de datos, Implementación de algoritmos básicos y Resultados de innovación de algoritmos.

En la primera etapa de Análisis y preprocesamiento de datos, se analizará los datos mediante la generación de estadísticas descriptivas, luego se debe eliminar datos nulos y generar un nuevo conjunto de datos el procesamiento de estos, reemplazando datos faltantes y/o nulos, con valores medios. Se mostrará cómo se agruparon los datos de cada grupo, eligiendo ciertas variables para la representación que de manera temprana se supuso que representaría de mejor manera el modelo de datos.

En la etapa de Implementación de algoritmos básicos se revisarán otros trabajos que tengan una relación con el proyecto. Se realizará una descripción de los algoritmos básicos elegidos para la tarea. Una vez definido el algoritmo básico, se implementará en la tarea y luego se realizarán los reportes correspondientes.

En última etapa, se revisarán otros trabajos también relacionados al proyecto, pero esta vez relacionados a algoritmos innovadores que de manera distinta resuelvan la problemática. Se observarán los resultados obtenidos por estas técnicas y se implementará un algoritmo innovador. Luego de implementarlo, se realizarán reportes de los resultados obtenidos.

Finalmente, con toda la información obtenida en el proyecto se realizará una presentación al curso, mostrando los procesos realizados, los datos obtenidos durante todo el ciclo del proceso y los resultados finales.

# Conformación de los datos

La investigación para completar el dataset se centró en los casos de los clientes con mora en los pagos en taiwán y compara la exactitud de la predicción de la probabilidad de mora entre seis métodos de extracción de datos.

Desde la perspectiva de la gestión de riesgos, el resultado de le exactitud de la predicción de la probabilidad estimada de incumplimiento será más valioso que el resultado binario de la clasificación: Clientes creíbles o no creíbles. Debido a que se desconoce la probabilidad real de incumplimiento, este estudio presentó el novedoso “Sorting Smoothing Method” para estimar la probabilidad real de incumplimiento. Con la probabilidad real de incumplimiento como variable de respuesta **(Y),** y la probabilidad predictiva de incumplimiento como variable independiente **(X),** el resultado de la regresión lineal simple **(Y = A + BX)** muestra que el modelo de predicción producido por la red neuronal artificial tiene el coeficiente de determinación más alto; su intercepción de regresión **(A)** está cerca de cero, y el coeficiente de regresión **(B)** a uno. Por lo tanto, entre las seis técnicas de minería de datos, la red neural artificial es la única que puede estimar con precisión la probabilidad real de incumplimiento.

Los atributos del dataset son los siguientes:

Esta investigación empleó una variable **binaria**, el pago por defecto (Si = 1, No = 0), como variable de respuesta. También se utilizaron las siguientes **23 variables como** variables explicativas:

**X1**: Monto del crédito concedido (dólar taiwanes): incluye tanto el crédito de consumo individual como el crédito familiar (suplementario).

**X2**: Genero (1 = masculino; 2 = femenino).

**X3**: Educación (1 = escuela de postgrado; 2 = universidad; 3 = escuela secundaria; 4 = otros).

**X4**: Estado civil (1 = casado; 2 = soltero; 3= otros).

**X5**: Edad (años).

**X6 – X11**: Historial de pagos anteriores. Rastreamos los registros de los últimos pagos mensuales (de abril a septiembre de 2005) de la siguiente manera: X6 = el estado de reembolso en septiembre de 2005; X7 = el estado de reembolso en agosto de 2005; … X11 = el estado de reembolso en abril de 2005. La escala de medición para el estado de reembolso es: -1 = pagar debidamente; 1 = retraso en el pago durante un mes; 2 = retraso en el pago durante dos meses; …; 8 = retraso en el pago durante ocho meses; 9 = retraso en el pago durante nueve meses y más.

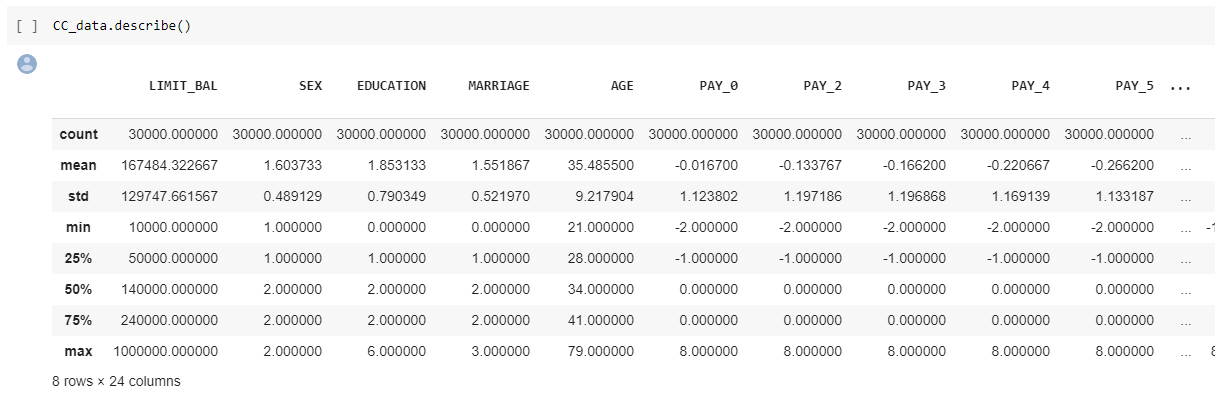
**X12 – X17**: Monto de la factura (dólar taiwanes). X12 = cantidad de la factura en septiembre de 2005; X13 = cantidad de la factura en agosto de 2005; ...; X17 = cantidad de la factura en abril de 2005.

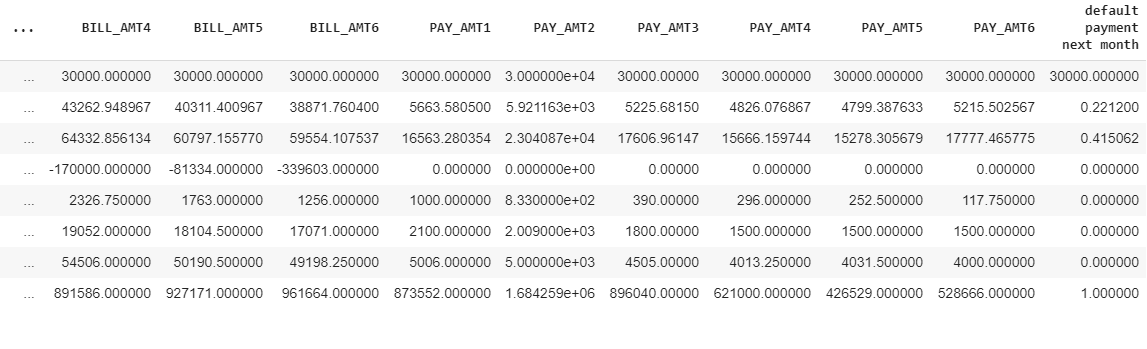
**X18 – X23**: Monto del pago anterior (dólar taiwanes). X18 = cantidad pagada en septiembre de 2005; X1 = cantidad pagada en agosto de 2005; …; X23 = cantidad pagada en abril de 2005.

## Preprocesamiento de Datos

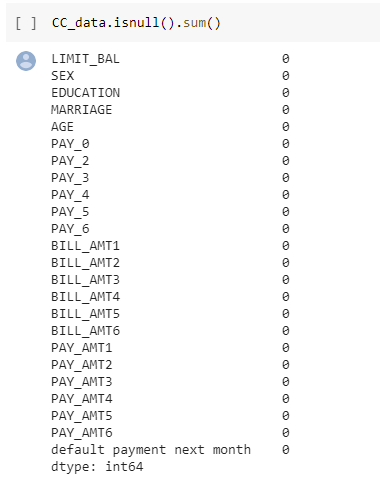
En esta etapa se comprenderán las variables del dataset “default of credit card clients Data Set”, mediante el uso de código en Python, analizando y modificando la información para procesarla de manera correcta.

Se genera estadísticas descriptivas del dataset.





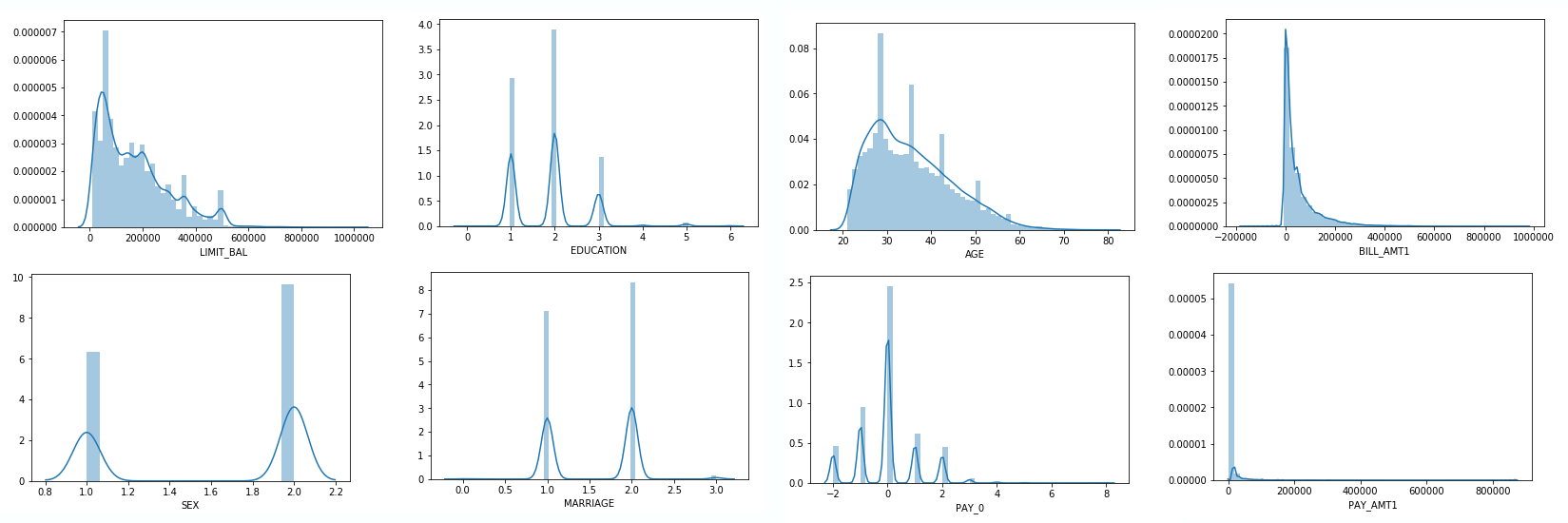
Se comprueba que los datos no presenten valores nulos y de haber, los suma.



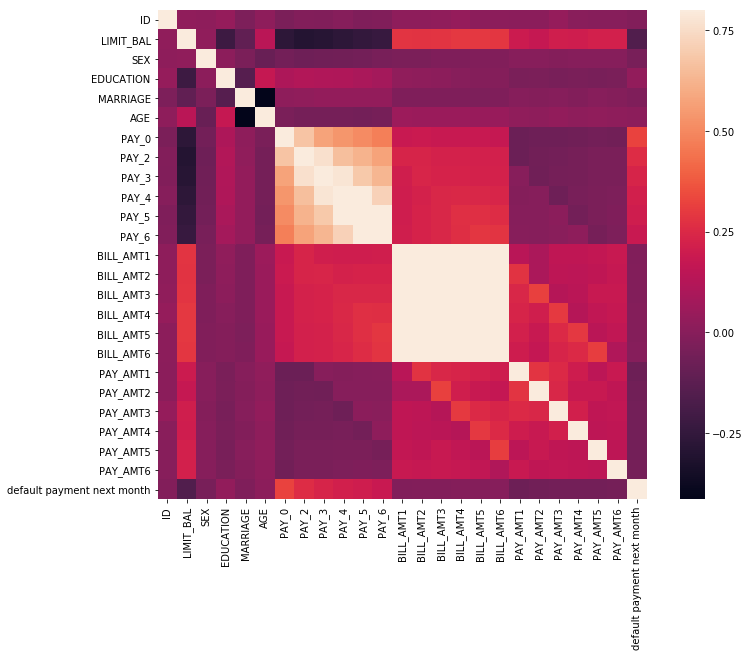
## Análisis Exploratorio

En esta etapa se puede realizar un análisis más avanzado de los datos, utilizando herramientas de visualización de datos que facilitan la comprensión de la relación entre las distintas variables.

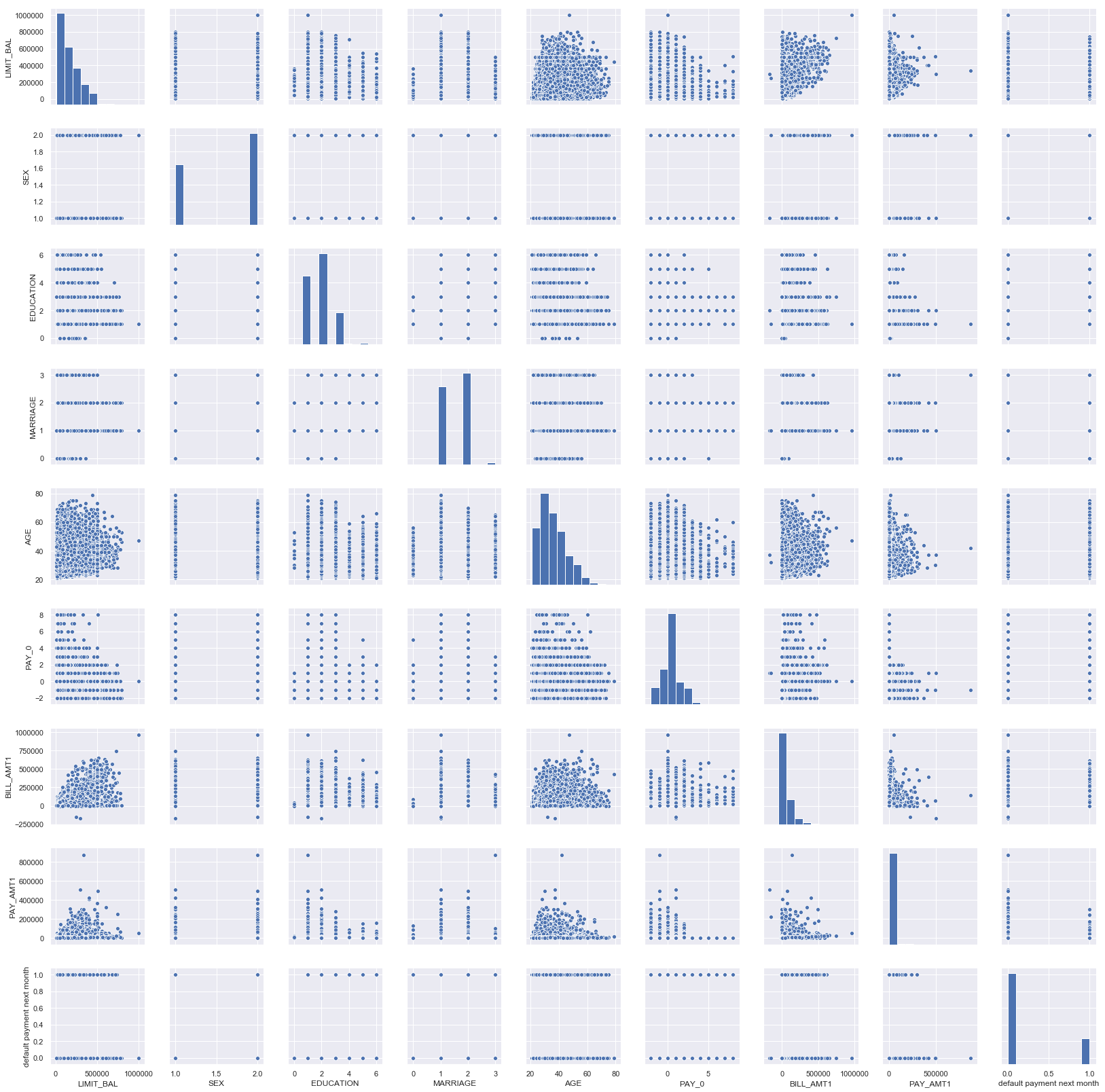
Se mostrará cómo están agrupado los datos en cada grupo. Se eligieron ciertas variables para la representación que tempranamente se supuso que podrían representar de mejor manera este modelo de datos.

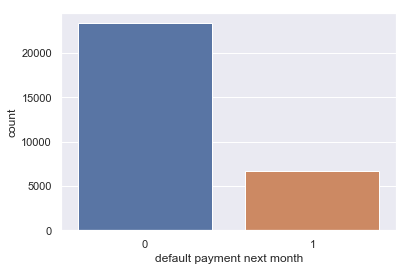


Matriz de correlación de las variables donde los datos que son mas cercanos a 1 son directamente proporcional y lo mas cercanos a -1 son inversamente proporcional

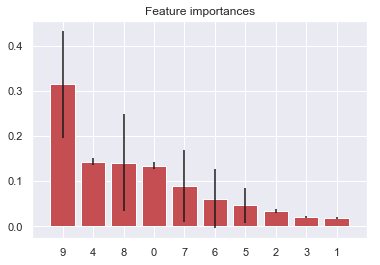


Grafica de dispersión de las variables, aquí se muestra las gráficas que creímos eran mas representativas para mostrar los datos.



Se muestra el recuento de los que pagaron (1) y los que no pagaron (0).

De acuerdo con el random forest classifier, PAY\_5 es la característica más importante, seguida de AGE y PAY\_4.



## Diseño de Experimentos

Un experimento seria contrastar dos tipos de clasificadores y calcular su accuracy y de esa forma establecer qué técnica es más adecuada para el modelo en cuanto al rendimiento de precisión en los resultados logrados.

Un segundo experimento sería realizar clustering, es el experimento perfecto para representar la conducta de asociación de las variables con respecto al resultado final de la clase.

# Implementación de los algoritmos básicos

## Revisión de métodos relacionados con “Credit Card Approval”

Proceedings of the 2017 Industrial and Systems Engineering Research Conference K. Coperich, E. Cudney, H. Nembhard, eds

Real Time Credit Card Default Classification Using Adaptive Boosting-Based Online Learning Algorithm

**Hongya Lu, Haifeng Wang and Sang Won Yoon Department of Systems Science and Industrial Engineering State University of New York at Binghamton, Binghamton, NY 13902**

Este estudio propone un algoritmo de aprendizaje en línea basado en el impulso adaptativo (AdaBoost) para reconocer el crédito potencial clientes predeterminados de tarjeta para escenarios en tiempo real. La clasificación de los solicitantes de tarjetas de crédito es clave y desafiante problema de competitividad para las compañías de tarjetas de crédito para reducir su riesgo de mercado. En este documento, la información del solicitante está representado por un vector de características que incluye nivel de educación, puntajes de crédito, registros históricos de pagos, etc. el proceso de aprendizaje, el conjunto de datos de capacitación se divide en dos partes, muestras originales y recién agregadas, basadas en

Tiempo de solicitud del cliente. El algoritmo AdaBoost se aplica para mejorar la capacidad básica de aprendizaje del árbol de decisión utilizando muestras originales. Para estabilizar los cambios dinámicos y estocásticos del proceso, se adopta una técnica de reponderación cuando

Se reciben nuevos datos. Luego, el mecanismo de aprendizaje en línea se utiliza para actualizar los parámetros del modelo para lograr el algoritmo autoadaptación en línea. El método de aprendizaje en línea basado en AdaBoost se compara con otras clasificaciones de vanguardia.

Métodos que incluyen Extreme Learning Machine (ELM), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayesian (NB) y K-Nearest Neighborhood (K-NN) utilizando una evaluación de validación cruzada de 5 veces.

Los resultados experimentales muestran que el algoritmo AdaBoost en línea no solo puede reducir el tiempo de entrenamiento, sino también obtener estabilidad y

precisión de clasificación competitiva en el conjunto de datos de clientes predeterminados de tarjetas de crédito.

Lu, Hongya & Wang, Haifeng & Yoon, Sang Won. .(2017). Real Time Credit Card Default Classification Using Adaptive Boosting-Based Online Learning Algorithm.

<https://www.researchgate.net/profile/Haifeng_Wang38/publication/319689046_Real_Time_Credit_Card_Default_Classification_Using_Adaptive_Boosting-Based_Online_Learning_Algorithm/links/59b991e1458515bb9c48a3f8/Real-Time-Credit-Card-Default-Classification-Using-Adaptive-Boosting-Based-Online-Learning-Algorithm.pdf>

**Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients**

NH Niloy, MAI Navid, Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients, *American Journal of Data Mining and Knowledge Discovery*. Vol. 3, No. 1, 2018, pp. 1-12. doi: 10.11648/j.ajdmkd.20180301.11

Los árboles de decisión usan una herramienta de soporte de decisiones que utiliza un árbol como modelo de gráfico y toma decisiones. El clasificador naïve bayesiano es un clasificador binario para obtener sí / no de los datos y es un método muy primitivo para encontrar una clasificación verdadera o falsa de un conjunto de datos. Ambos algoritmos se pueden usar como modelo predictivo en el aprendizaje automático y la minería de datos. Aquí, se realiza un análisis comparativo entre estos dos algoritmos de aprendizaje automático. Los datos que tenemos se utilizan para clasificar si el cliente es el titular de la tarjeta de crédito predeterminado o no. En la perspectiva de la gestión de riesgos, el resultado puede usarse para obtener con precisión el resultado de clasificar clientes creíbles o no creíbles.

NH Niloy &  MAI Navid ( 2018). Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients

<http://www.sciencepublishinggroup.com/journal/paperinfo?journalid=603&paperId=10027054>

**The comparisons of data mining techniques for the predictive**

**accuracy of probability of default of credit card clients**

Esta investigación tuvo como objetivo el caso de los pagos predeterminados de los clientes en Taiwán y compara la precisión predictiva de la probabilidad de incumplimiento entre seis métodos de minería de datos. Desde la perspectiva de la gestión de riesgos, el resultado de la precisión predictiva de la estimación

La probabilidad de incumplimiento será más valiosa que el resultado binario de la clasificación: clientes creíbles o no creíbles. Porque el verdadero problema

se desconoce la capacidad de incumplimiento, este estudio presentó la novela "Método de suavizado de clasificación" para estimar la probabilidad real de incumplimiento.

Con la probabilidad real de incumplimiento como la variable de respuesta (Y), y la probabilidad predictiva de incumplimiento como la variable independiente(X), el resultado de regresión lineal simple (Y = A + BX) muestra que el modelo de pronóstico producido por la red neuronal artificial tiene el alto coeficiente de determinación de esta; su intersección de regresión (A) es cercana a cero y el coeficiente de regresión (B) a uno. Por lo tanto, entre los seis

técnicas de minería de datos, la red neuronal artificial es la única que puede estimar con precisión la probabilidad real de incumplimiento.

 2007 Elsevier Ltd. Todos los derechos reservados.

I-Cheng Yeh & Che-hui Lien( 2007). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients

[**https://www.sciencedirect.com/science/article/pii/S0957417407006719**](https://www.sciencedirect.com/science/article/pii/S0957417407006719)

First International Conference on Artificial Intelligence and Cognitive Computing

Mejora de la precisión de predicción de incumplimiento de crédito utilizando técnicas de conjunto

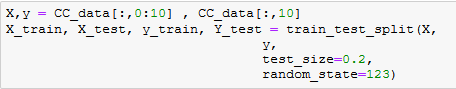
De E. Sivasankar y B. Emil Richard Singh

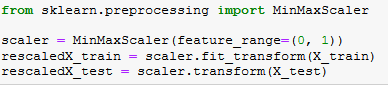
 La calificación crediticia de una institución o individuo proporciona una imagen financiera sugerente y la fortaleza del individuo o la institución. Le da al prestamista la capacidad de visualizar la potencialidad en la medida en que la institución o el individuo puedan aprovechar el crédito. La predicción predeterminada de la suma de todos los atributos, como el historial de pagos, es un instrumento común utilizado para generar la calificación crediticia. El objetivo de esta investigación es comparar la precisión predictiva del conjunto de clasificadores básicos utilizando técnicas de embolsado, refuerzo y bosque aleatorio en la predicción del incumplimiento de los clientes de tarjetas de crédito y sugiriendo la técnica con la mayor precisión. El pago predeterminado de los clientes en el conjunto de datos de Taiwán se utiliza para construir el modelo. Los algoritmos de clasificación ML, como el vecino K más cercano, Naive Bayesian, el árbol de decisión y las máquinas de vectores de soporte, se aplican para crear el modelo base en el conjunto de datos. El ensacado, el aumento y el bosque aleatorio se aplican en el conjunto de datos para generar un modelo para la predicción. Se tabula la precisión de cada uno de los modelos para varios grados. El método de filtro de características de ganancia de información se utiliza para identificar características con máxima entropía. Las características con alta entropía sugeridas por la ganancia de información junto con las técnicas de conjunto se utilizan para construir el nuevo modelo. La precisión del nuevo modelo se tabula. Se encuentra que la técnica de conjunto de refuerzo tiene la mejor precisión de predicción.

Emil Richard Singh B., Sivasankar E. (2019) Enhancing Prediction Accuracy of Default of Credit Using Ensemble Techniques. In: Bapi R., Rao K., Prasad M. (eds) First

## Implementación del algoritmo básico

Primero se utilizará el algoritmo de entrenamiento, después de tener listo el preprocesamientos de los datos se crean nuevas variables para introducir características y etiquetas. donde se dividen los datos en conjunto de entrenamiento y conjunto de prueba. estos conjuntos son normalizados entre los rangos 0 y 1 para luego realizar el entrenamiento y mapeo de los datos.

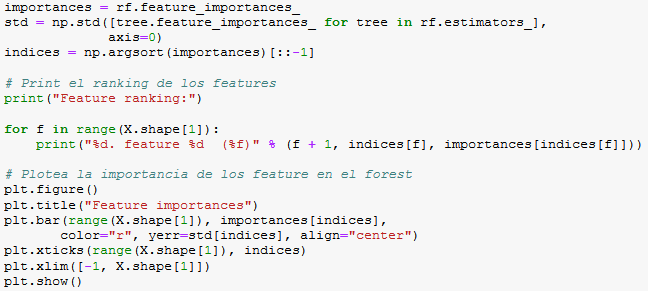


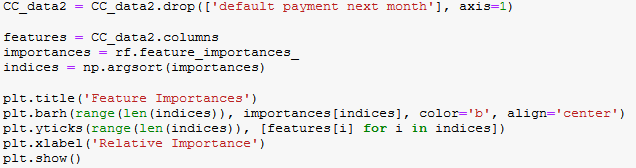


Para saber que variable es la más importante se utilizara *RandomForestClassifier* el cual es un método versátil de aprendizaje automático capaz de realizar tanto tareas de regresión como de clasificación. También lleva a cabo métodos de reducción dimensional, trata valores perdidos, valores atípicos y otros pasos esenciales de exploración de datos. Es un tipo de método de aprendizaje por conjuntos, donde un grupo de modelos débiles se combinan para formar un modelo poderoso.

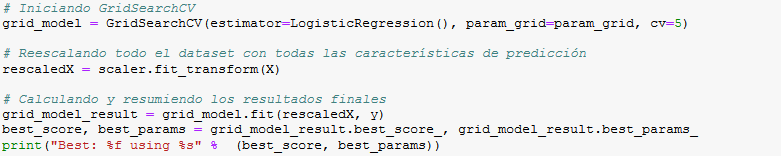
En Random Forest se ejecutan varios algoritmos de árbol de decisiones en lugar de uno solo. Para clasificar un nuevo objeto basado en atributos, cada árbol de decisión da una clasificación y finalmente la decisión con mayor “votos” es la predicción del algoritmo.

Para finalizar, se predecirá con regresión logística para modelar la relación entre una respuesta discreta y un conjunto de variables independientes, o explicativas





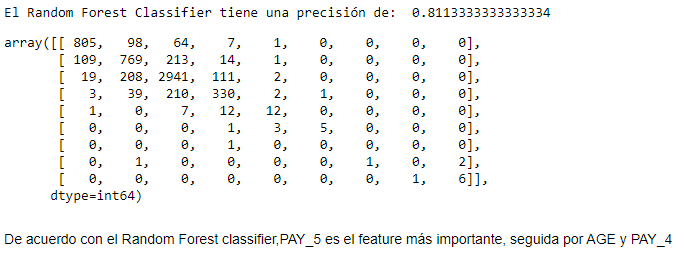
La identificación del mejor modelo de regresión logística se realiza mediante la comparación de modelos (para este caso *X\_tes*t e *Y\_train*) utilizando el cociente de verosimilitud, que indica a partir de los datos de la muestra cuanto más probable es un modelo frente al otro. La diferencia de los cocientes de verosimilitud entre dos modelos.Dicho esto se podra realizar una busqueda en rejilla donde se automatiza realizando una búsqueda, no aleatoria  en puntos de dicho espacio repartidos de forma espaciada en una rejilla. Una vez entrenado, podemos acceder al mejor modelo, a la mejor puntuación y a los mejores parámetros encontrados a través de los atributos *best\_estimator\_, best\_score\_ y best\_params\_* respectivamente:

****

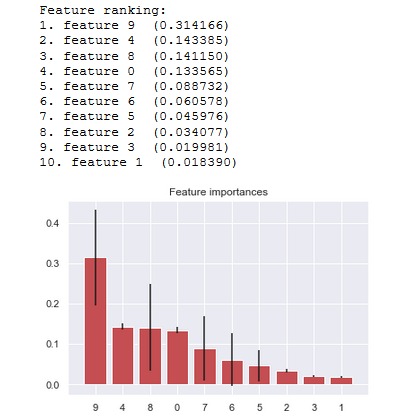
## Resultados Obtenidos

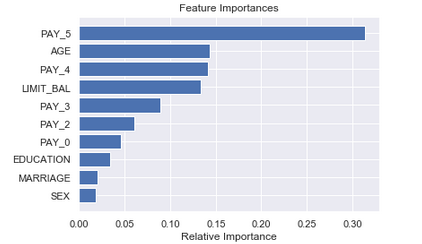
El análisis de nuestro laboratorio mostro varios resultados poco esperados

Para empezar el algoritmo *RandomForestClassifier* nos demostro que la precisión del algoritmo , la matriz de confusión y las variables más importantes. El cual fue:

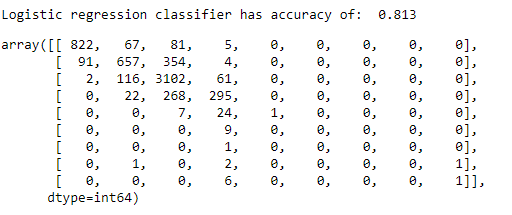


Cabe destacar que también el algoritmo nos genero una lista de los features mas importante al menos importante ,mostrando como resultado lo siguiente:





Luego con la regresión logística nos mostró que:



Luego para finalizar en se mostró la mejor búsqueda con rejilla el cual se muestra que el resultado final es:



# Resultados innovación en algoritmos

## Describir Algoritmos innovadores

## Crossvalidation es a modo general, una técnica que consiste en tomar el conjunto de datos de entrada y a partir de ellos, crear dos conjuntos separados, un conjunto de entrenamiento y prueba y el otro conjunto de validación, usualmente se suele usar el 80% de los datos de entrada para entrenar el modelo y el otro 20% de los datos es usado para hacer el testeo, acto seguido realizar el análisis al conjunto de entrenamiento y validar el análisis en el otro conjunto. El proceso de ajuste optimiza los parámetros del modelo para que se ajusten a los datos de entrenamiento lo mejor posible.

## Esta técnica sirve para obtener los mejores resultados en cuanto a la medición de la calidad de la predicción del modelo cuando no se tiene un conjunto explícito de datos de prueba, ayuda a medir el comportamiento del modelo creado, encontrar un mejor modelo rápidamente y mitigar el sobreajuste u overfitting mediante K-Fold, que es un método o enfoque de Crossvalidation más sofisticado que generalmente da como resultado un modelo menos sesgado en comparación con otros métodos.

## Un dato para tener en cuenta es que Crossvalidation solo produce resultados significativos si el conjunto de validación y prueba se han extraído de la misma población.

## No es recomendable usar Crossvalidation con datos que están sujetos a evolucionar con el tiempo, ya que esto puede provocar la existencia de diferencias sistemáticas entre los conjuntos de entrenamiento y validación.

## Comparación con y sin crossvalidation

## Sin crossvalidation

Este experimentos se realizo sin crossvalidation solamente con la función [sklearn.model\_selection.train\_test\_split](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) que nos permite dividir el dataset en dos bloque, un bloque está destinado al entrenamiento y el otro bloque para la validación del modelo



## 

Se realizaron prueba de 5 en 5 hasta 60 como se muestra se ve en la gráfica con un estimador de valor 500 para demostrar que en cierto punto ya no mejora el modelo según el número del estimador

Esta es la gráfica solo con el estimador normal que va en aumento de 5 en 5 hasta 60

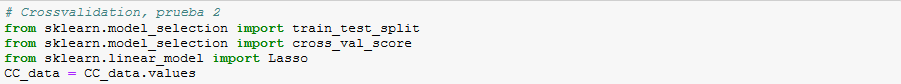
## *Sin olvidar que el resultado sin crossvalidation fue*

## 

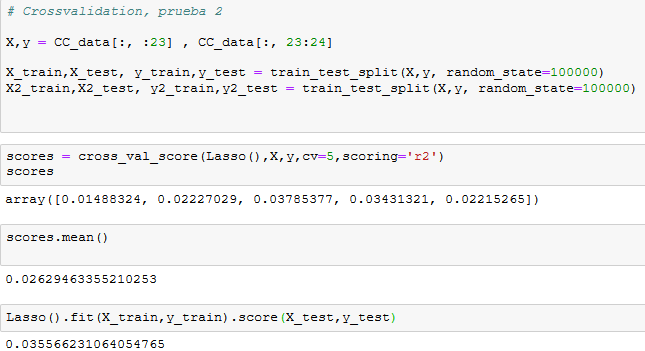
**Crossvalidation**

Para ocupar crossvalidation tuvimos que elegir el tipo de validación que queríamos realizar nosotros optamos por la validación que revisa la puntuación del estimador

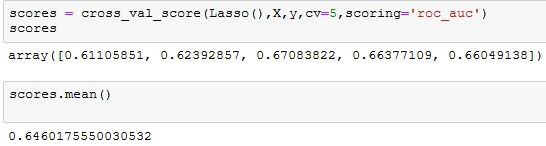
Aquí se están cargando las librerías a ocupar

****

Aquí se guardo la dataset para la variable x e y y luego se separan para entrenamiento y testing

****

Aquí se ve la puntuación de estimador y al final se muestra el resultado promedio



En esta grafica se ve una mejora en comparación del anterior modelo y se muestra que en cierto punto ya no aumenta la predicción en el modelo

Se ve que la gráfica tiene una tendencia logarítmica donde en un punto ya no podrá seguir aumentando



**Análisis de comparativa:**

El modelo que presenta un mejor resultado es el modelo que se realizó con crossvalidation por que elimina cierto ruido que presentan los datos esto se puede ver en los resultado del modelo sin crossvalidarion que fueron 0,8070 en comparación con el modelo de crossvalidation que fue 0,8094 este aumente se puede pensar que es mínimo pero al pensar en la cantidad de datos que procesa es importante por lo que es recomendable ocupar crossvalidation en los modelos de predicción para tener más precisión en los datos

## Implementar Algoritmo Innovador

**K-means**

K-means es un algoritmo de aprendizaje automático no supervisado, que agrupa puntos de datos similares y descubre patrones subyacentes.

El trabajo de K-means comienza seleccionando un primer grupo de centroides al azar, que se utilizan como puntos de inicio para cada grupo, acto seguido realiza cálculos iterativos para optimizar las posiciones de los centroides. Estas iteraciones se detienen cuando los centroides se han estabilizado, ósea no hay cambios en sus valores indicando que el agrupamiento ha sido exitoso, o se ha alcanzado el numero definido de iteraciones.

Un dato para considerar es que su rendimiento generalmente no es tan competitivo como el de las otras técnicas de agrupación sofisticadas, ya que ligeras variaciones en los datos podrían conducir a una gran variación.

Se puede decir que el uso de K-means está sujeto a la necesidad de descubrir nuevas relaciones entre los parámetros de un data set, utilizando como medida de similitud la distancia euclidiana u otras funciones como Manhattan, Lavenshtein, Mahalanobis, etc.

**Implementación de K-means**

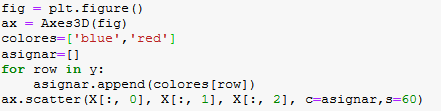
Para aplicar K-means fue necesario visualizar ciertas variables para realizar el procesamiento de esta, se tuvo que elegir las variables mas representativas ya que el software Jupyter presentaba problemas de procesamiento, por esta razón se escogio a las variables

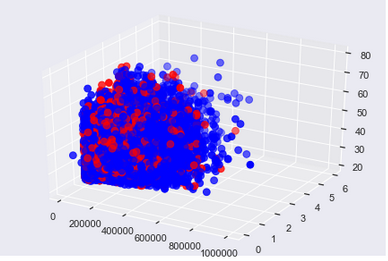
*“"LIMIT\_BAL","EDUCATION","AGE","PAY\_0","BILL\_AMT1","PAY\_AMT1*"”.

Con esto se ligo a la variable “payment next month” ya que esta era la que informaba la finalidad de la predicción que era saber si se pagaría la tarjeta de crédito.

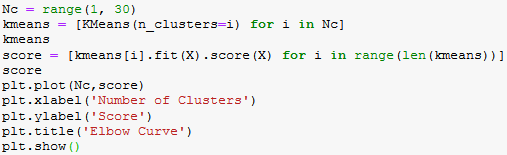


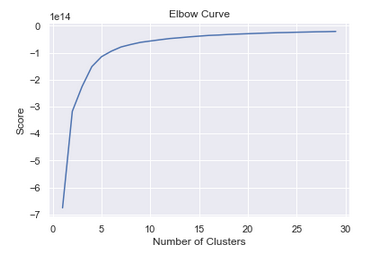
Luego se presento una modelo 3D con los datos, para tener una mejor comprensión y visualización de estos, donde fueron agrupados en 2 grupos



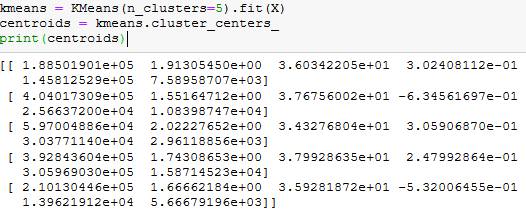


Con esto se comprobó que los datos, se encuentran demasiado acoplados entre sí, por esta razón no se puede realizar un análisis previo, por esta razón se buscó un mejor K para el K-means que se adecuara a los datos, como se ve en esta curva:

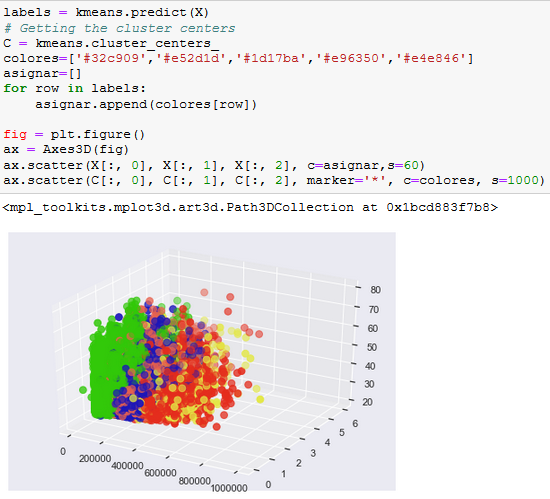




Se decidió como equipo elegir el k =5 debido a la inclinación pronunciada de la curva. Para luego ejecutar el algoritmo donde se obtuvieron las etiquetas y los centroids.



Luego se mostro la misma grafica anteriormente visualizada , a diferencia que ahora está representada con k = 5, es decir 5 agrupaciones

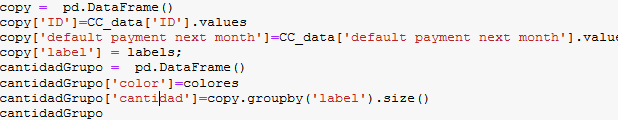


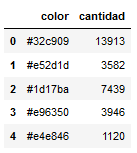
.

Aquí podemos ver que el Algoritmo de K-Means con K=5 ha agrupado a los 30.000 datos que pueden o no pueden pagar la tarjeta de crédito.

Después se pueden mostrar los datos que están asociados a cada cluster

.

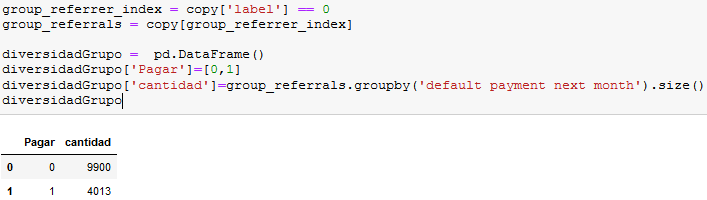


****

|  |  |  |
| --- | --- | --- |
|  | Claro | Oscuro |
| Verde | #32c909 | #0d7f05 |
| Rojo | #e52d1d | #c32518 |
| Azul | #1d17ba | #130f7a |
| Rosado | #e96350 | #e54933 |
| Amarillo | #e4e846 | #c8cc19 |
| negro | #4f4f4f | #000000 |
| Morado | #c720ce | #621066 |
| Celeste | #25c3dd | #1b98ac |
| naranja | #e08538 | #c2691e |
| café | #914a17 | #773d13 |
| blanco | #ffffff | #dbdbdb |
| Dorado | #b8a91d | #a3961a |
| esmeralda | #78de2a | #5cb11b |
| turqueza | #31df90 | #1ec278 |
| guinda | #cc0026 | #7b0017 |

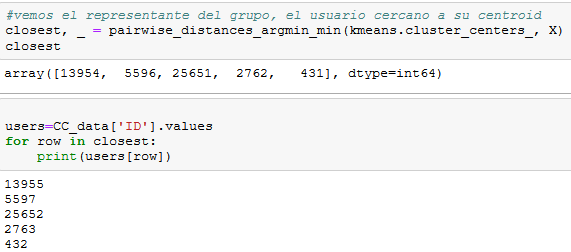
Tabla de colores utilizados en el algoritmo

Donde en esta tabla se muesta que al cluster de color verde claro (32c909) están asociados 13.913 Personas, rojo claro (e52d1d) con 3.582 personas,azul claro(1d17ba) con 7.439, Rosado claro (e96350) con 3.946 y Amarillo claro (e4e846) con 1.120.Esto en simples palabras significa que el cluster verde claro es el que presenta mayor densidad de datos



Con esto podemos concluir que para el cluster con mayor numero de datos asociados hay mayor cantidad de que personas que no pagan su tarjeta que son 9.900 personas expecificamente.

Tambien se puede Buscar los usuarios que están más cerca a los centroids de cada grupo que podríamos decir que tienen los rasgos característicos que representan a cada cluster:



Aquí se puede apreciar que las personas con ID:13.955,5.597,25.652, 2.763 y 432 son los representantes de cada cluster.

**Comparativa de cantidad de Clusterrs**

A medida que el K crece la cantidad de muestras de datos se ira haciendo más pequeñas por ende estas terminaran siendo no significativas y agrupaciones no representativas sin relación alguna.

# 

# Conclusi´on

# ----------

En este trabajo se estudió el caso de predecir si una persona paga un crédito o no, mediante el análisis, procesamiento e implementación de algoritmos básicos y algoritmos innovadores como K-Means y Crossvalidation o validación cruzada.

se vio que los resultados pueden variar dependiendo del modelo de predicción, al realizar el experimento,.esto demuestra que es mejor utilizar el crossvalidation ya que hay una mejora en el ordenamiento de los datos, como se pudo apreciar en la leve mejora al comparar ambos modelo (con y sin crossvalidation), a pesar de ser el 0,2% de diferencia ,teniendo en cuenta la cantidad de datos que se procesan se considera que es una mejora leve pero significativa. además, gracias a la comparativa del random forest se comprobo que los modelos llegan a un punto en que no pueden seguir mejorando. Como lo demuestra la grafica donde la tendencia de la gráfica era una logarítmica esto demuestra que se acerca al mejor valor, pero nunca va a llegar al 100%

Otra comparación realizada fue con k-mean pudimos ver que grupo son los que pagan mas contra los morosos donde cada cluster nos arrojaba que la mayoría dentro de los grupos es morosa

# Bibliografía

[1] *Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.*

*aprendemachinelearning.(s.f.).Obtenidodehttps://www.aprendemachinelearning.com/k-means-en-python-paso-a-paso/*