Universidad de Lima

Facultad de Ingeniería Industrial

Carrera de Ingeniería de Sistemas



Trabajo final ANALÍTICA PREDICTIVA DE DATOS 2019-2

Voltaire Grandez Alfredo

**Código** 20161563

Zúñiga Ramírez Pedro

**Código** 20161620

**Asesor**

José Antonio Taquia Gutiérrez

Lima – Perú  
Noviembre de 2019

análisis y predicción de datos PARA

###### INDICE

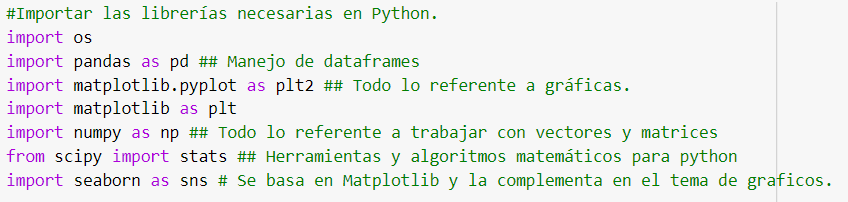
1. Formulación del problema
2. Herramientas analíticas utilizadas
3. Dataset
4. Validación
5. Comparación de técnicas
6. Presentación de resultados

**FORMULACION DEL PROBLEMA**

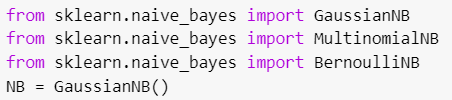
Un banco que brinda crédito de campaña a minoristas cuenta con un histórico de crédito pre aprobado, gracias al recojo de datos de años anteriores, que contiene un campo que presenta a los minoristas que si han aceptado el crédito o no. Se acerca la campaña navideña, por ello el banco ha decidido realizar un análisis de datos para saber cuáles mypes aceptaran el crédito de campaña.

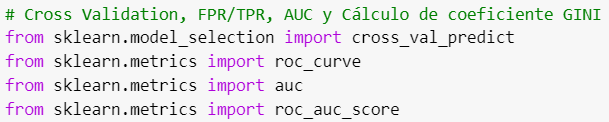
**HERRAMIENTAS ANALÍTICAS UTILIZADAS**

Para el presente análisis se ha utilizado el entorno colaborativo en nube Google Colaboratory. Las principales técnicas utilizadas son la imputación paramétrica, el análisis de cajas (boxplot), histogramas, matriz de confusión, precisión, modelos de clasificación Naive Bayes, K-NN y SVM (Linear, Poly, Rbf y Sigmoid). Las librerías usadas son os, pandas, matplotlib, numpy, scipy, seaborn, sklearn, entre otras.















**DATASET**

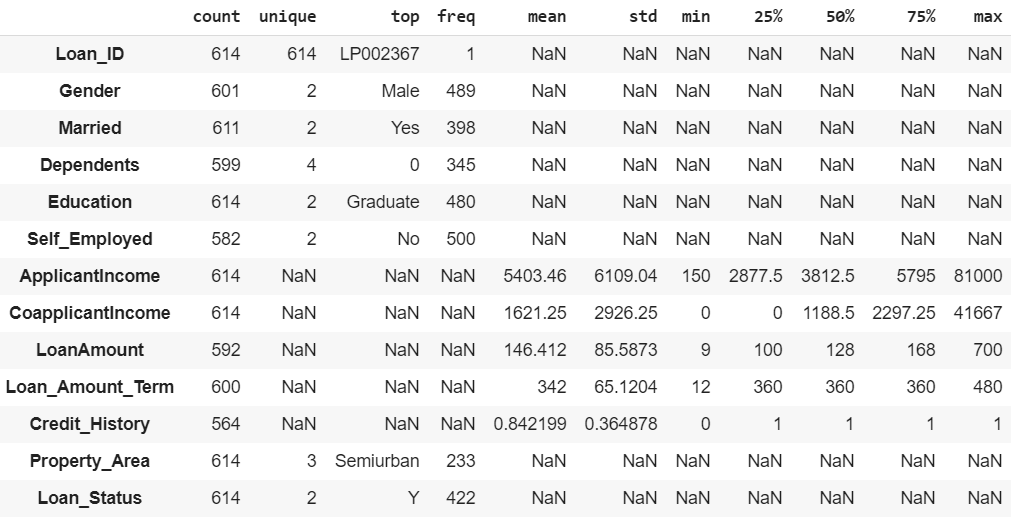
El dataset usado es un archivo csv que contiene 12 campos con 614 registros. Contiene variables cuantitativas y cualitativas. La variable dependiente (y) es el campo “Loan\_Status” el cual presenta 2 tipos de dato, “Y” y “N”. Esta variable dependiente determina a los minoristas que si han aceptado o no un crédito por campaña brindado por el banco.

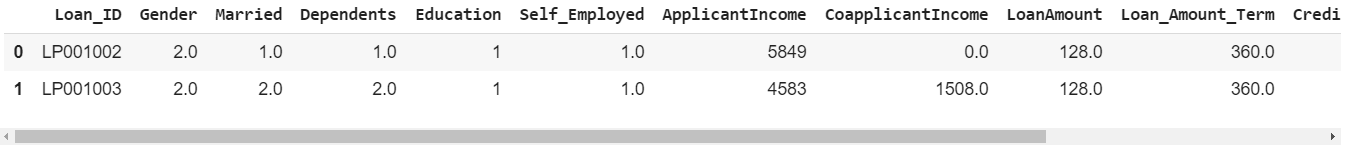
Para hacer un correcto análisis del dataset debemos alterar el formato de los valores. Para ello, verificamos los valores nulos que presenta y su suma con una función básica llamada unm\_missing. Luego, realizamos la imputación paramétrica para llenar los valores nulos con la moda, en el caso de las varibles cualitativas y, la mediana, para las variables cuantitativas. Finalmente, verificamos que todos los registros tengan valores numéricos y no nulos.

**VALIDACIÓN**

Con el dataset modificado, podemos empezar a generar resultados mediante técnicas de análisis. El primer análisis es el Box Plot sobre el campo “LoanAmount” que muestra un sesgo muy pronunciado, al igual que el análisis de densidad que sustenta lo anterior. Luego, mostramos un histograma del campo “Dependents” para detallar un gráfico de diferentes valores. Finalmente, le quitamos la variable Loan\_ID o identificador único para poder modelar la información

Ahora comenzaremos a crear la data de train y test mediante un split y, asi, obtener X\_train, X\_test, y\_train e y\_test.





**COMPARACIÓN DE TÉCNICAS**

Para el análisis de datos, se han utilizado modelos de clasificación para determinar cuál es el mejor. A continuación, se muestra técnicas como Naive Bayes, K-NN, SVM, que a su vez presenta 4 tipos de kernel; linear, poly, rbf y sigmoid.

El primer modelo de clasificación realizado será el bayesiano Naives Bayes. Importamos las 3 librerias requeridas, GaussianNB, MultinomialNB y BernoulliNB. Además, importamos las librerías necesarias para realizar el cross-validation como cross\_val\_predict, roc\_curve, entre otras. Elegimos el método de predicción probabilística y hacemos un plot para mostrar la curva ROC de valores verdaderos positivos y falsos positivos con su respectivo score AUC. Ahora, obtendremos la matriz de confusión de acuerdo al training y predict de test. Finalmente, presentamos un reporte de clasificación que muestra la precisión, f1-score, support, etc.

El segundo modelo de clasificación realizado será el K-NN (KNeighbors Classifier). Importamos las librerias requeridas. Además, importamos las librerías necesarias para realizar el cross-validation como cross\_val\_predict, roc\_curve, entre otras. Elegimos el método de predicción probabilística y hacemos un plot para mostrar la curva ROC de valores verdaderos positivos y falsos positivos con su respectivo score AUC. Debido al bajo resultado de la prueba, decidimos hacer una reducción de dimensiones (n\_components = 3) utilizando PCA. Despues, verificamos el plot y comparamos el anterior AUC con el actual con reducción de dimensionalidades. Ahora, obtendremos la matriz de confusión de acuerdo al training y predict de test. Finalmente, presentamos un reporte de clasificación que muestra la precisión, f1-score, support, etc.

El tercer modelo de clasificación realizado será el SVM. Primero, como sabemos que los kernel de SVM son sensibles a las escalas de las características, tendremos que estandarizar nuestras variables predictoras tanto para train y test. Hecho esto, importamos la libreria requerida SVC y asignamos los 4 tipos de kernel a usar; linear, poly, rbf y sigmoid. Además, importamos las librerías necesarias para realizar el cross-validation como cross\_val\_predict, roc\_curve, entre otras. Elegimos el método de función de decisión para cad uno y hacemos un plot para mostrar la curva ROC de valores verdaderos positivos y falsos positivos con su respectivo score AUC. Las curvas muestran el área obtenida por cada kernel y, como están todas presentes en el gráfico, es fácil saber cuál es mejor. Ahora, obtendremos la matriz de confusión de acuerdo al training y predict de test de cada kernel. Finalmente, presentamos un reporte de clasificación que muestra la precisión, f1-score, support, etc.

**PRESENTACIÓN DE RESULTADOS**

