

**ANÁLISIS DE DATOS CREDITICIOS APLICACIÓN DE MODELOS
ANALÍTICOS Y EXPLORACIÓN DE DATOS**

VICTOR ALEXANDER MARTIN ROJAS

**UNIVERSIDAD PILOTO DE COLOMBIA
MODELOS ANALÍTICOS PARA LA MINERÍA Y VISUALIZACIÓN DE DATOS
BOGOTÁ D.C SEMESTRE 1 – 2025**

Introducción

Información de calidad, según Wang, R. Y., & Strong, el fundamento de la calidad de en los datos:

"La calidad de los datos es fundamental para la toma de decisiones efectivas, ya que datos inexactos, incompletos o inconsistentes pueden llevar a interpretaciones erróneas y afectar negativamente los resultados organizacionales."[1]

Podemos en esta breve cita observar la trascendencia de la calidad de la información, en un entorno real, el día a día laboral del trabajo con datos tiene múltiples retos al trabajar protegiendo información sensible, con fuentes de datos diversas que no necesariamente están bien estructuradas, normalizadas y congruentes.

Adicionalmente las diversas formas de trabajo, almacenamiento y metodologías de manejo de información generan que no exista un único medio de trabajo con data, es común trabajar con datos de Excel, archivos planos, pdf, bases de datos relacionales y no relacionales, etc, esto sumado a la existencia de diversos tipos de datos como string, numéricos, flotantes, fechas, binarios, etc. Resaltan la importancia de conocer y trabajar en profundidad las herramientas para grandes volúmenes de datos en la que Python junto con librerías como Pandas tienen una gran importancia.

En este trabajo se usa una fuente de datos cuya data original ha sido anonimizada y obfuscada con la finalidad de resguardar información personal y sensible en cumplimiento de las normas PII, aunque esto incrementa el nivel de dificultad al realizar análisis de datos, es una practica común y necesaria ampliamente adoptada en que como ingenieros, Analistas o científicos de datos debemos trabajar sacando el máximo provecho posible a la data manteniendo los estándares de calidad, protección de información personal y cumplimiento de normativas legales y corporativas.

Objetivo:

- Realizar un ejercicio exploratorio y aplicar fundamentos de minería de datos y trabajo de los conceptos de análisis descriptivo, limpieza de data y estadísticas.
 - Cargar data de “credit aproval” y aplicar técnicas de limpieza
 - Realizar análisis descriptivo sobre los datos limpios
 - Analizar el resultado de estos datos y obtener conclusiones respecto a la aprobación de créditos

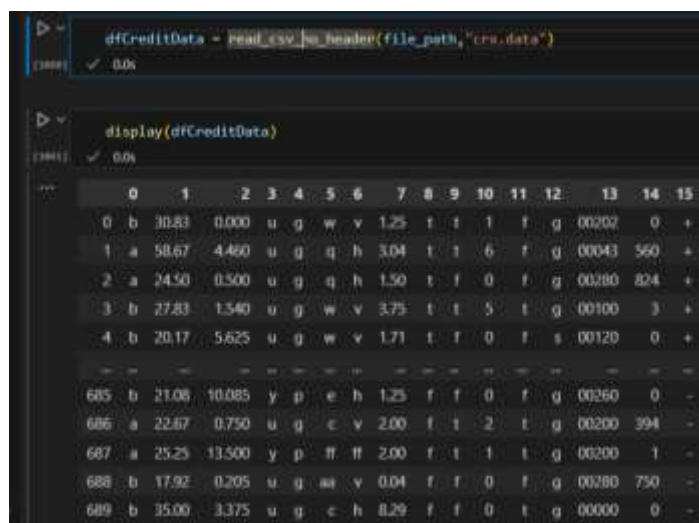
Metodología:

El estudio toma de base la fuente “CREDIT APPROVAL”, analizando la data contenida para observar datos inconsistentes, faltantes, outlayers, tipos de datos, etc.

Sobre estos datos se aplican procesos de limpieza de data para mejorar la calidad del input a las técnicas de análisis descriptivo y búsqueda de información relevante, también se realizarán algunas validaciones y suposiciones dado que la información está incompleta por el efecto de ofuscación en la protección de datos.

1. Carga de data y exploración inicial de estructura

Para trabajar con el set de datos se utiliza Python 3.13, con las librerías Pandas, numpy junto con las librerías para graficación matplotlib, seaborn



```
dfCreditData = pd.read_csv(file_path, "cra.data")

display(dfCreditData)
```

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|-----|-----|-------|--------|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-------|-----|-----|
| 0 | b | 30.83 | 0.000 | u | g | w | v | 1.25 | t | t | 1 | f | g | 00202 | 0 | + |
| 1 | a | 58.67 | 4.460 | u | g | q | h | 3.04 | t | t | 6 | f | g | 00043 | 560 | + |
| 2 | a | 24.50 | 0.500 | u | g | q | h | 1.50 | t | f | 0 | f | g | 00280 | 824 | + |
| 3 | b | 27.83 | 1.540 | u | g | w | v | 3.75 | t | t | 5 | t | g | 00100 | 3 | + |
| 4 | b | 20.17 | 5.625 | u | g | w | v | 1.71 | t | f | 0 | f | s | 00120 | 0 | + |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 685 | b | 21.08 | 10.085 | y | p | e | h | 1.25 | f | f | 0 | f | g | 00260 | 0 | - |
| 686 | a | 22.67 | 0.750 | u | g | c | v | 2.00 | f | t | 2 | t | g | 00200 | 394 | - |
| 687 | a | 25.25 | 13.500 | y | p | ff | ff | 2.00 | f | t | 1 | t | g | 00200 | 1 | - |
| 688 | b | 17.92 | 0.205 | u | g | aa | v | 0.04 | f | f | 0 | f | g | 00280 | 750 | - |
| 689 | b | 35.00 | 3.375 | u | g | c | h | 8.29 | f | f | 0 | t | g | 00000 | 0 | - |

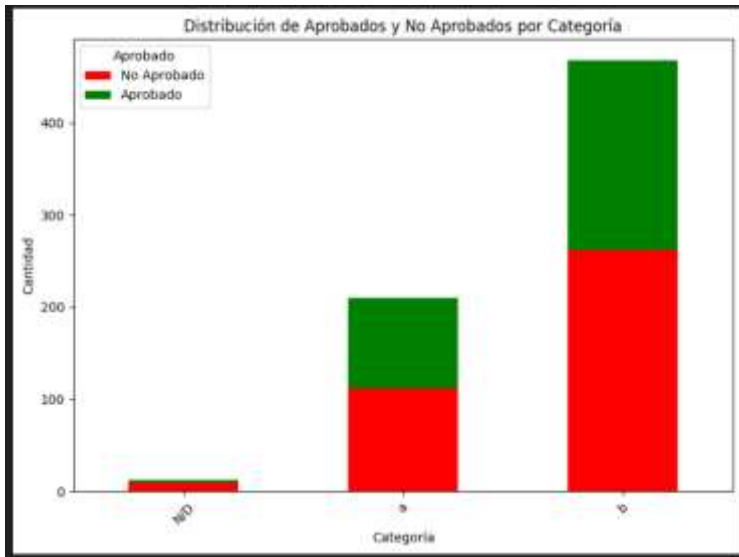
Tipos de datos

En el Dataset se evidencian campos sin nombres aparentes con cadenas de texto, float e integer.

Retos e inconsistencias iniciales

El dataset utiliza el símbolo interrogación para los casos desconocidos, esto genera que campos que originalmente son numéricos sean leídos por pandas como object, este es el caso del campo A2 que contiene la edad del solicitante.

Existe un desconocimiento de los valores de los campos sin embargo de pueden hacer presunciones en base a la información disponible para este ejercicio se tomará el campo 0 como categórico sin mayor impacto dado que no representa una diferencia significativa en un resultado positivo o negativo, 1 como un campo de edad que se encuentra entre 10 y 90 años, 15 como resultado + positivo, favorable, - negativo



Limpieza y preparación

Para realizar la limpieza de data he creado una función para ajustar los campos vacíos dependiendo del caso requerido, en el ejemplo de edades se observa que los datos conservan una distribución normal y los NaN representan una pequeña fracción de los datos, por ello aplica un llenado de valores nulos usando la media que llega como parámetro a la función `limpiezadatos()`:

```
print(dfCreditData[1].isna().sum())
```

[1230] ✓ 0.0s

... 12

```
def limpiezadatos(df,column,metodo,valor=0):
    print(f"limpieza de {df[column].isnull().sum()} registros método {metodo}")
    if metodo == 'media':
        df[column] = df[column].fillna(df[column].mean())
    if metodo == 'mediana':
        df[column] = df[column].fillna(df[column].median())
    elif metodo == 'valor':
        df[column] = df[column].fillna(valor)
    elif metodo == 'verdadero_falso':
        df[column] = df[column].replace({'t': 1, 'f': 0})
    elif metodo == 'string':
        # Múltiples espacios
        df[column] = df[column].str.replace(r'\s+', ' ', regex=True)
        # Caracteres especiales
        df[column] = df[column].str.replace(r'^\w\s', '', regex=True)
    return df
```

[1258] ✓ 0.0s

Llamado función limpieza

```
dfCreditData = limpiezadatos(dfCreditData,8,'verdadero_falso')
dfCreditData = limpiezadatos(dfCreditData,9,'verdadero_falso')
```

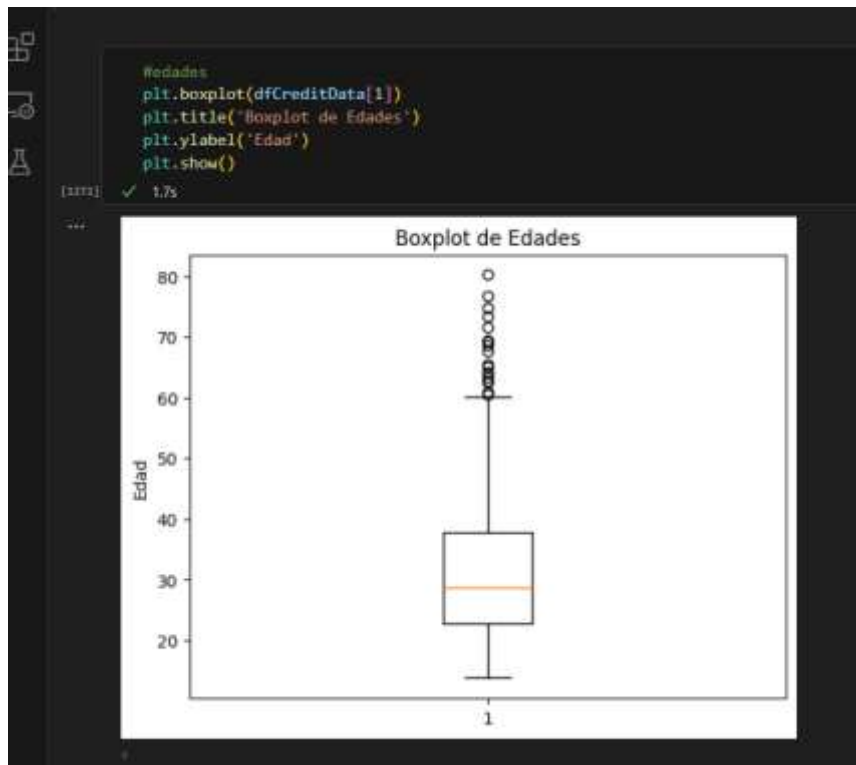
Estructura definida

Para realizar análisis se establecerán correlaciones, para el trabajo con las columnas de valores categóricos que contienen una letra indicio de item_purchased por ejemplo, se usará LabelEncoder para asignar un valor a cada categoría presente en los datos.

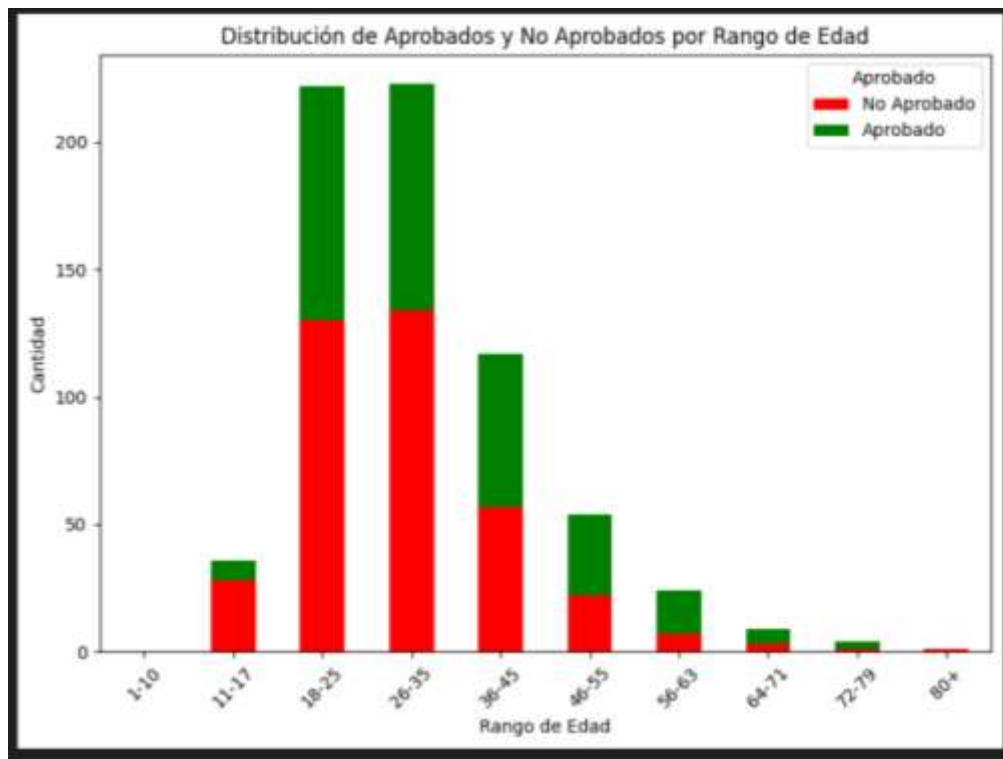
```
dfCreditData['4_cod'] = label_encoder.fit_transform(dfCreditData[3])  
dfCreditData['5_cod'] = label_encoder.fit_transform(dfCreditData[4])
```

2. Técnicas de análisis

Con el set de datos depurado, se construye un análisis de los aplicantes, comenzando con las edades, mediante un grafico de caja y bigotes podemos observar la media en 31 años, con valores extremos en edades de 80+, normal considerado la longevidad de una persona



En relación con éxito o fracaso en el crédito la edad no es un factor determinante, de hecho se ve una proporción entre diferentes rangos de edades, para gestionar este punto se crearon categorías, rangos para facilitar análisis y visualización:



Búsqueda de data relevante

En búsqueda de los factores relevantes en el crédito se aplicó una relación de correlación con esto se construyó el dataframe correlación, en el se usaron los campos codificados hacia valores numéricos:

```
# Calcular la correlación entre las columnas 8, 9 y 15
correlacion = dfCreditData[['4_cod', '5_cod', '6_cod', '7_cod', 8, 9, 11, 15]].corr()

# Mostrar la matriz de correlación
```

Se presume en base a los valores que contiene la columna A4 son referentes al estado Civil: el campo sería: casado? Donde:

U posiblemente es unmarried

Y Yes

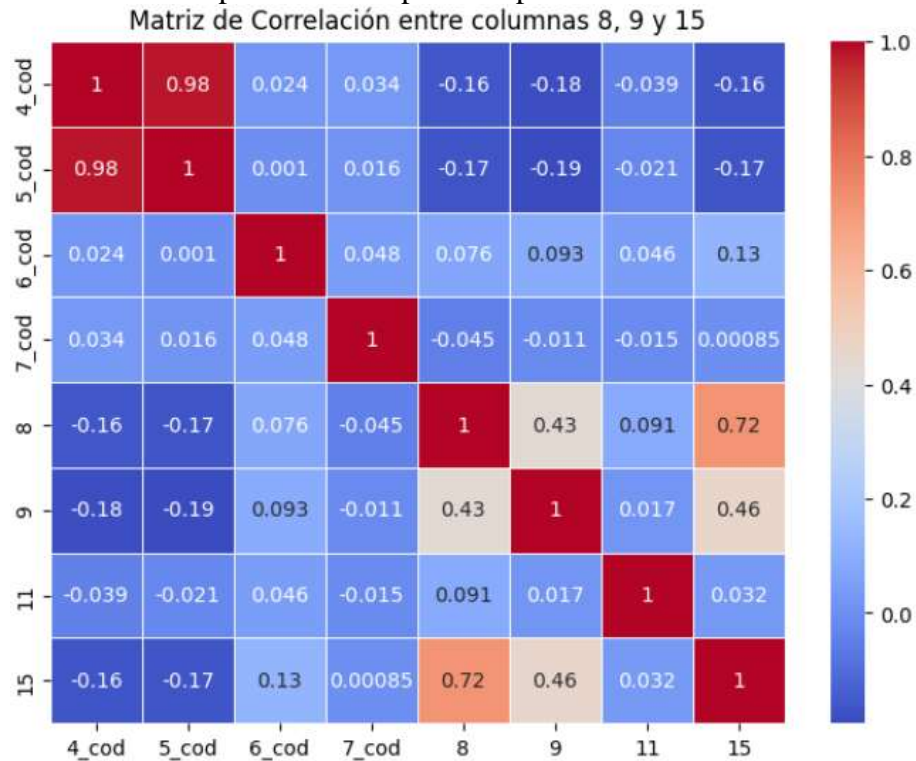
L living together

Para este caso los campos desconocidos se asignarán con U por ser el estado por defecto que tienen las personas antes de los demás

```
dfCreditData = limpiezadatos(dfCreditData, 11, verdadero_falso)
dfCreditData = limpiezadatos(dfCreditData, 3, 'valor', 'u')
```

Relación de data en el proceso de aprobación de crédito

Mediante un mapa de calor se pueden apreciar las relaciones de datos:



Podemos encontrar una fuerte relación entre los campos 4(presumiblemente estado civil) y 5(no determinado) sin embargo se evidencia que estos dos campos no tienen mayor impacto en el resultado (15), por ello aún con una fuerte relación entre ellas no se consideran significativas.

Sin embargo, se encuentran una relación alta entre el campo 9(8 en el orden) y el resultado 16 (15 en el orden) y una relación razonable positiva entre el campo 10 y el resultado.

Esta relación positiva denota un cambio en el que los casos en estado f tienen una alta relación en casos negativo(- denegado el crédito) con los identificados con u en el campo A4, el porcentaje de rechazo s aumenta considerablemente lo que afecta la fuerza de la correlación entre el campo 10(9 en el orden) presumiblemente el genero

y el resultado mostrando el sesgo hacia las mujeres no casadas.



Hallazgos

Pese a no tenerla información completa o las descripciones detalladas de los campos del set de datos es posible realizar análisis de los datos buscar relaciones, depurar data inconsistente y tomar decisiones para mejorar la calidad e integridad que permiten obtener una descripción de los factores considerados relevantes en el proceso de un crédito como los campos 9 y 10.

Conclusiones

Una funcionalidad modular y reutilizable para limpieza de datos garantiza uniformidad, el set de datos debe ser depurado limpiando valores nulos, outliers, mejorando calidad de los resultados.

Existen valores en el set de datos que no tienen impacto significativo en el resultado de aprobación o rechazo, para ello una matriz de correlación permite describir estos comportamientos y sugerir los atributos a tener en cuenta.

El reemplazo de datos es de gran ayuda, ajustar todos los valores V y f a un booleano 0 y 1 o ajustar los '?' por valores nulos, mejora la calidad con un valor nulo el tipo de dato se puede convertir a float, in, para valores numéricos, esto ajusta el datatype al valor que realmente corresponde.

Los campos 9 y 10 tiene un impacto en el resultado, que se complementa con el campo 4 (presumiblemente el estado civil), el bien adquirido con el préstamo (presumiblemente A6), no tiene mayor impacto en la aprobación o denegación del crédito.

Bibliografía

Lichman, M. (1999). *UCI Machine Learning Repository: Credit Approval Dataset*. Universidad de California, Irvine. Recuperado de <https://archive.ics.uci.edu/dataset/27/credit+approval>

"La calidad de los datos es fundamental para la toma de decisiones efectivas, ya que datos inexactos, incompletos o inconsistentes pueden llevar a interpretaciones erróneas y afectar negativamente los resultados organizacionales."

— Wang, R. Y., & Strong, D. M. (1996). *Beyond accuracy: What data quality means to data consumers*. *Journal of Management Information Systems*, 12(4), 5-33.

Anexos

- Codificación en Python
- Dataset de ingeta