Instituto Tecnológico
y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

PREDICCIÓN DE COMPORTAMIENTO CREDITICIO

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
MAESTRO EN CIENCIA DE DATOS

Presenta: **ALDO EMMANUEL VILLARREAL PALOMINO**

Director: **DR. JAIME EMMANUEL ALCALÁ TEMORES**

Tlaquepaque, Jalisco, septiembre de 2023.

AGRADECIMIENTOS

Gracias a mis compañeros de estudio, especialmente a Elisa, que siendo completos extraños dedicó tiempo a ayudar a un compañero perdido. A Yared, por haberme ayudado académica y personalmente. A Ángel, Luis, Jesús, Alejandra y Alex, con quienes pasé tantas horas haciendo trabajos y que terminamos siendo amigos. Gracias a Paola Montoya por sus asesorías extracurriculares y su clase que fue la más cercana al mundo real. A Fernando por ser el profesor más preciso, dedicado y con un curso tan bien pensado y desarrollado con el que tuve el gusto de tomar clase. Gracias a mi alma mater ITESO por permitirme continuar mis estudios y su excelente calidad humana. Gracias a Jaime por recibir mi proyecto. Gracias a Manuel Mora por impulsar mi desarrollo profesional. Gracias a mi familia, especialmente a mi hermano Ricardo, sin él no habría estudiado la licenciatura en esta institución.

RESUMEN

Este trabajo se centra en mejorar la evaluación del riesgo crediticio utilizando el conjunto de datos *South German Credit*. El objetivo principal es seleccionar el mejor modelo de clasificación entre Regresión Logística, *Support Vector Machine* (SVM), Árboles Aleatorios y Árboles de Decisión para predecir si un cliente es "bueno" o "malo" en términos de crédito.
Los objetivos específicos del trabajo son describir el conjunto de datos, encontrar el mejor modelo de clasificación utilizando el *Accuracy* como métrica principal. En el desarrollo del trabajo, se realiza una descripción detallada del conjunto de datos, incluyendo su origen, tamaño, atributos y objetivo. También se justifica la elección de los cuatro modelos de clasificación mencionados y se describe brevemente cada uno de ellos. Se realiza un análisis de correlación para identificar las relaciones entre las variables del conjunto de datos y su influencia en el riesgo crediticio. Se destaca qué variables, como la duración del crédito, el monto del crédito, la edad del solicitante y la tasa de interés, tienen correlación con el riesgo crediticio. Se opta por utilizar todas las variables del conjunto de datos. Se resalta la importancia de la medida de *Accuracy* para evaluar el rendimiento de los modelos de clasificación en este problema y se presentan los resultados obtenidos en siete experimentos. Los principales cambios entre los experimentos fueron los tratamientos de las variables categóricas, utilizando principalmente *LabelEncoder* y *OneHotEncoder* para evitar una jerarquización falsa. Se presentan las *Accuracy* obtenidas para cada modelo en todos los experimentos y se selecciona el mejor de ellos.
El mejor modelo fue el SVM con *grid search* para la afinación de los hiperparámetos, con lo cual se obtuvo un *Accuracy* de 83.92%.

TABLA DE CONTENIDO

1. INTRODUCCIÓN

En esta sección se presenta el contexto del trabajo, se plantea el problema a resolver y su justificación, se menciona la base de datos que se utilizará y su utilidad para responder al problema planteado, y se definen los objetivos particulares y generales del trabajo.

1.1. Contexto

Existen numerosos estudios y trabajos previos que abordan el tema de la evaluación del riesgo crediticio [1]. Revisar y analizar estas investigaciones puede proporcionar fundamentos teóricos y técnicas utilizadas en el campo, así como un impacto positivo en la salud financiera de países enteros.

El análisis de riesgo crediticio es un proceso importante para las instituciones financieras. Al evaluar el riesgo de impago de sus clientes, usando herramientas como análisis de datos o modelos predictivos de puntuación crediticia, las instituciones financieras pueden tomar decisiones informadas sobre préstamos y créditos. Esto puede ayudar a proteger a las instituciones financieras de pérdidas y ayudar a garantizar que los clientes tengan acceso a los fondos que necesitan. Utilizar un modelo predictivo basado datos, como el conjunto de *South German Credit,* puede mejorar la capacidad de estas instituciones para evaluar el riesgo crediticio, lo cual impacta en la rentabilidad y la estabilidad del sector financiero y en la economía global. En 2022 en México se observó un índice de morosidad del 1.8% para crédito automotriz, 2.2% para nómina, 4.5% crédito personal y 5.0% para adquisición de bienes de consumo [2], por lo que aún hay margen para mejorar modelos que permita reducir la morosidad, especialmente en crédito persona y adquisición de bienes de consumo.

1.2. Justificación

Desde una perspectiva económica, el análisis de riesgo crediticio es un tema crítico en la industria financiera. Los bancos y otras instituciones financieras necesitan evaluar con precisión el riesgo de impago de sus clientes para tomar decisiones informadas sobre préstamos y créditos. Utilizar un modelo predictivo basado en este conjunto de datos puede ayudar a estas instituciones a mejorar su capacidad para evaluar el riesgo crediticio, lo que puede tener un impacto significativo en la rentabilidad y la estabilidad del sector financiero.

Desde una perspectiva científica, puede ser utilizado como un caso de estudio para explorar y mejorar las técnicas de análisis predictivo. Los investigadores pueden utilizar este conjunto de datos para desarrollar y comparar diferentes modelos de aprendizaje automático y técnicas de análisis predictivo. Además, los resultados obtenidos pueden ser utilizados para mejorar los modelos predictivos en otras áreas, como el análisis de riesgo en la industria de seguros o el análisis de tendencias de mercado en la industria minorista.

Este conjunto de datos confiable y de alta calidad puede ayudar a mejorar la capacidad de las instituciones financieras para evaluar el riesgo crediticio y también puede ser utilizado como un caso de estudio para mejorar las técnicas de análisis predictivo en una variedad de áreas.

1.3. Problema

En la actualidad, la evaluación del riesgo crediticio es un tema crítico en la industria financiera y se realiza de forma manual o semiautomática. Aunque existen diversas soluciones que automatizan el proceso, la mayoría son soluciones privadas y no están disponibles para personas que se estén formando en esta área. Por lo tanto, implementar un modelo predictivo basado en el conjunto de datos puede tener un impacto significativo en la formación y capacitación de estudiantes y profesionales en el área de análisis de riesgo crediticio.

Se espera que al implementar un algoritmo de aprendizaje automático utilizando este conjunto de datos, se pueda mejorar la precisión y la eficiencia en la evaluación del riesgo crediticio. En lugar de depender de procesos manuales o semiautomáticos que pueden ser propensos a errores y subjetividades, un modelo predictivo basado en este conjunto de datos puede proporcionar resultados más precisos y objetivos. Además, este modelo puede ser utilizado como una herramienta de apoyo para los profesionales del área, lo que puede mejorar la calidad y la rapidez de sus decisiones.

Se espera que el uso de un modelo predictivo basado en el conjunto de datos pueda llevar a la industria financiera de A->B, es decir, de una evaluación del riesgo crediticio basada en procesos manuales o semiautomáticos, a una evaluación del riesgo crediticio más precisa y eficiente basada en técnicas de aprendizaje automático. Asimismo, se espera que la implementación de este algoritmo en el ámbito académico pueda mejorar la formación y capacitación de estudiantes y profesionales en el área de análisis de riesgo crediticio, lo que puede tener un impacto positivo en el sector financiero en general.

1.4. Objetivos

1.4.1 Objetivo General:

El objetivo de este trabajo es mejorar la evaluación del riesgo crediticio y proporcionar una mayor transparencia y comprensión en el proceso de solicitud de crédito para las compañías de crédito y los solicitantes de crédito mediante la selección del mejor modelo entre diferentes variaciones de clasificación: Regresión logística, *Support Vector Machine*, Árboles aleatorios y Árboles de decisión.

1.4.2 Objetivos Específicos:

1. Describir el conjunto de datos.

2. Realizar análisis exploratorio de las variables.
3. Encontrar el mejor modelo de clasificación que permita la experimentación efectuada utilizando como principal métrica el *Accuracy* entre cuatro modelos

2. METODOLOGÍA

En esta sección se describen los datos, se realiza una exploración de las variables categóricas y las numéricas, se describen los modelos y las métricas con las que serán evaluados y se explica la experimentación que fue realizada.

2.1. *Descripción de los datos*

El conjunto de datos *South German Credit* (Crédito del sur de Alemania) [3] es un conjunto de datos disponible en el repositorio de aprendizaje automático de UCI (*University of California, Irvine*). Fue recopilado por Hofmann y Klinkenberg en 1994 y está relacionado con la evaluación del riesgo crediticio de clientes bancarios en el sur de Alemania.

El conjunto de datos consta de 1000 instancias, donde cada instancia corresponde a un cliente bancario. Cada instancia tiene 20 atributos, incluyendo información demográfica del cliente, datos bancarios (por ejemplo, duración del crédito, cantidad del crédito, historial crediticio anterior) y datos sobre la situación laboral del cliente.

Este conjunto de datos puede ser usado para predecir si un cliente es "bueno" o "malo" en términos de crédito, según ciertos criterios específicos definidos por los investigadores. La variable objetivo es un atributo binario que indica si el cliente cumplió con sus obligaciones crediticias o no.

Este conjunto de datos se ha utilizado ampliamente en la evaluación de modelos de clasificación en el área de riesgo crediticio y es un ejemplo comúnmente utilizado en cursos de aprendizaje automático y minería de datos.

Como puede observarse más adelante en Figura 1, no hay valores faltantes y no fue necesaria una limpieza de datos, todas las columnas están completas.

```
Data columns (total 21 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   status_account          1000 non-null   object
 1   duration                1000 non-null   int64
 2   credit_history          1000 non-null   object
 3   purpose                 1000 non-null   object
 4   credit_amount           1000 non-null   int64
 5   savings_account         1000 non-null   object
 6   employment_since        1000 non-null   object
 7   installment_rate        1000 non-null   int64
 8   personal_status_sex     1000 non-null   object
 9   other_debtors           1000 non-null   object
 10  present_residence       1000 non-null   int64
 11  property                1000 non-null   object
 12  age                     1000 non-null   int64
 13  other_installment_plans 1000 non-null   object
 14  housing                 1000 non-null   object
 15  number_credits          1000 non-null   int64
 16  job                     1000 non-null   object
 17  people_liable           1000 non-null   int64
 18  telephone               1000 non-null   object
 19  foreign_worker          1000 non-null   object
 20  credit_risk             1000 non-null   int64
dtypes: int64(8), object(13)
```

Figura 1. Descripción general de las variables

Las variables incluidas en la base de datos son:

1. *Satus_account*: variable nominal que indica el estado de la cuenta corriente o el balance actual (ver Figura 2), donde los valores posibles son (no especifica la moneda):
    - A11: < 0
    - A12: 0 <= … < 200
    - A13: >= 200
    - A14: no hay cuenta corriente

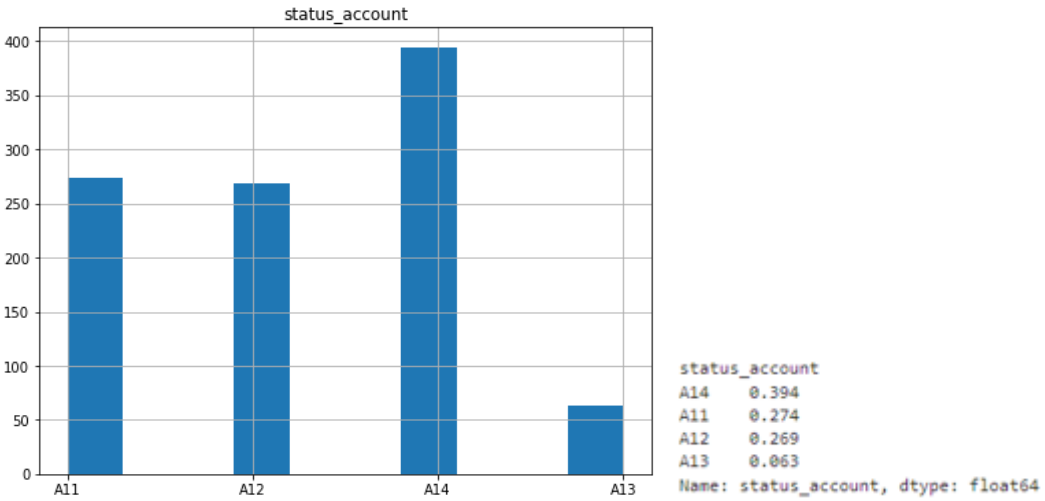


Figura 2. Status Account

2. *duration*: variable numérica que indica la duración del crédito en meses (ver Figura 3).

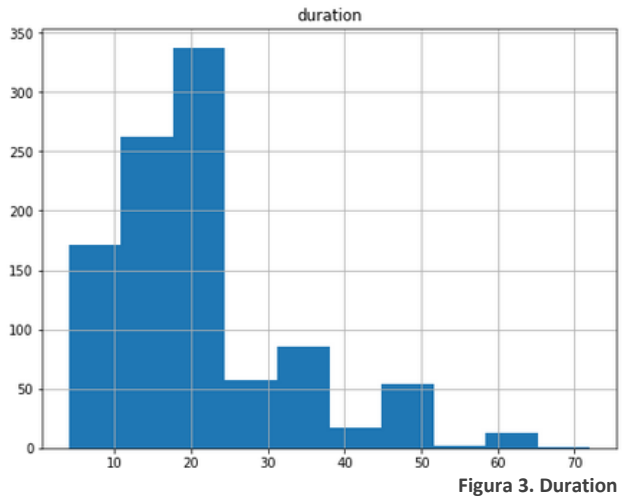
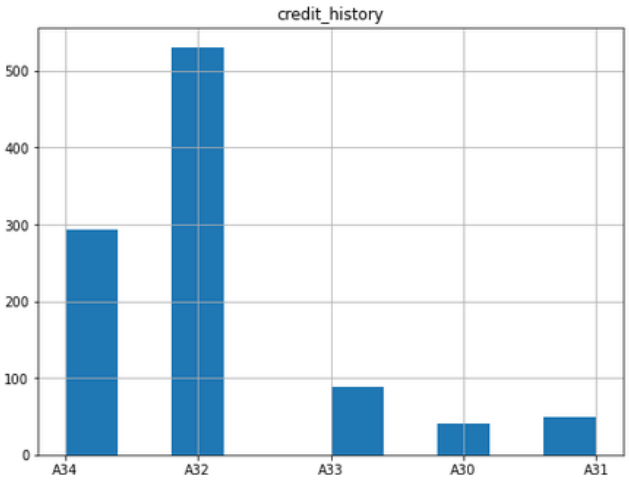


Figura 3. Duration

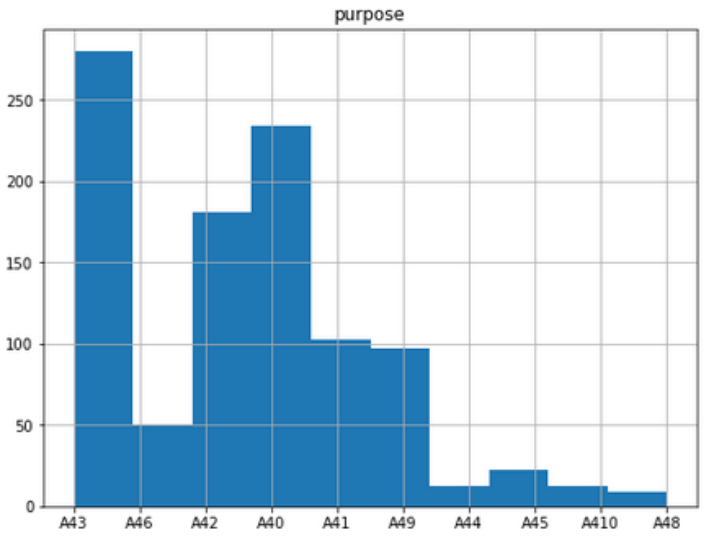
3. *credit_history*: variable nominal que indica el historial crediticio del solicitante (ver Figura 4), donde los valores posibles son:
    - A30: sin créditos tomados o créditos existentes pagados puntualmente
    - A31: todos los créditos existentes pagados puntualmente hasta ahora
    - A32: atraso en el pago de créditos existentes en el momento actual
    - A33: problemas críticos en el historial crediticio




```
credit_history
A32    0.530
A34    0.293
A33    0.088
A31    0.049
A30    0.040
```

Figura 4. Credit_history

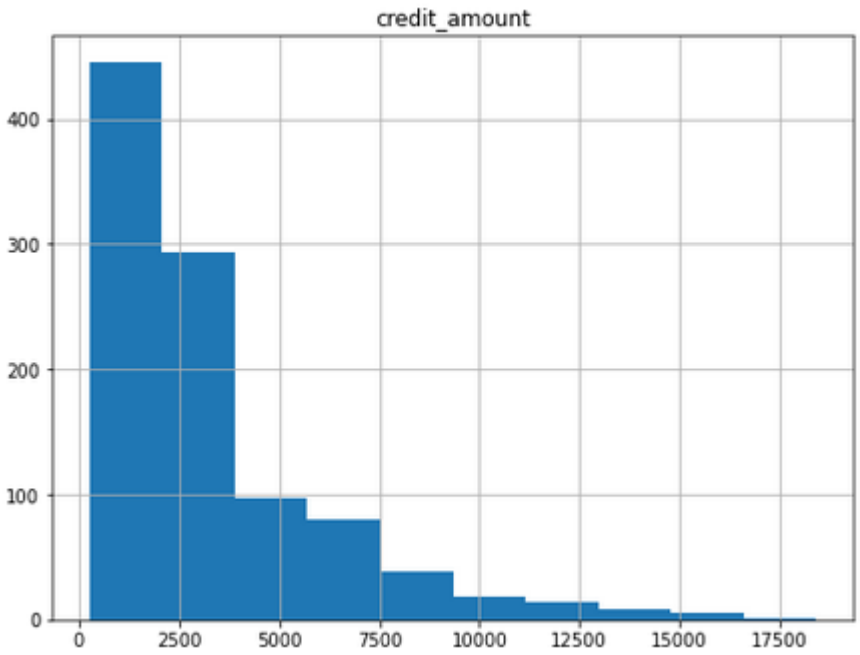
4. *purpose*: variable nominal que indica el propósito del crédito (ver Figura 5), donde los valores posibles son:
    - A40: coche nuevo
    - A41: coche usado
    - A42: muebles/equipos domésticos
    - A43: radio/televisión
    - A44: aparatos eléctricos
    - A45: reparaciones
    - A46: educación
    - A47: vacaciones
    - A48: capacitación
    - A49: otros




```
purpose
A43     0.280
A40     0.234
A42     0.181
A41     0.103
A49     0.097
A46     0.050
A45     0.022
A44     0.012
A410    0.012
A48     0.009
Name: purpose, dtype: float64
```

Figura 5 Purpose

5. *credit_amount*: variable numérica que indica el monto del crédito solicitado (No especifica la moneda; ver Figura 6):




```
credit_amount
1478    0.003
1262    0.003
1258    0.003
1275    0.003
1393    0.003
        ...
1459    0.001
882     0.001
3758    0.001
1136    0.001
4576    0.001
Name: credit_amount,
```

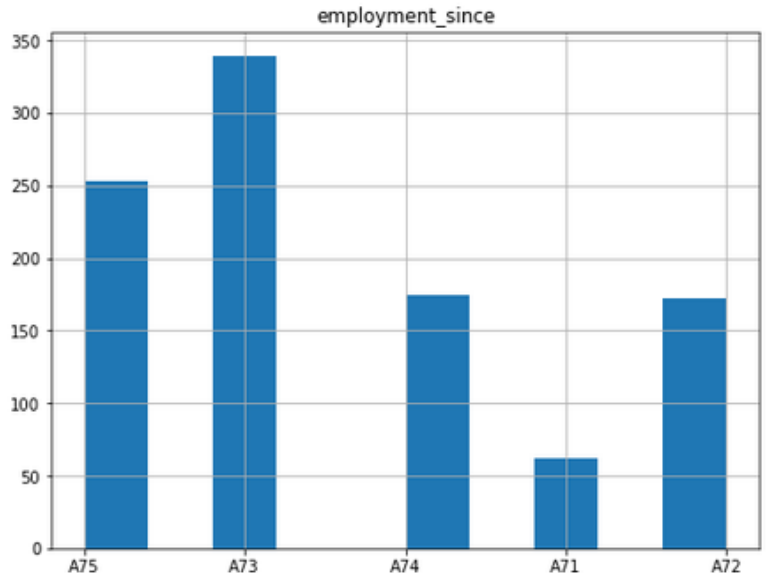
Figura 6.- Credit Amount

6. *savings_account*: variable nominal que indica la cuenta de ahorros o los títulos del solicitante, donde los valores posibles son:
    - A61: < 100
    - A62: 100 <= ... < 500
    - A63: 500 <= ... < 1000
    - A64: >= 1000
    - A65: desconocido/ninguno


```
savings_account
A61    0.603
A65    0.183
A62    0.103
A63    0.063
A64    0.048
Name: savings_account
```

Figura 7.- Savings_account

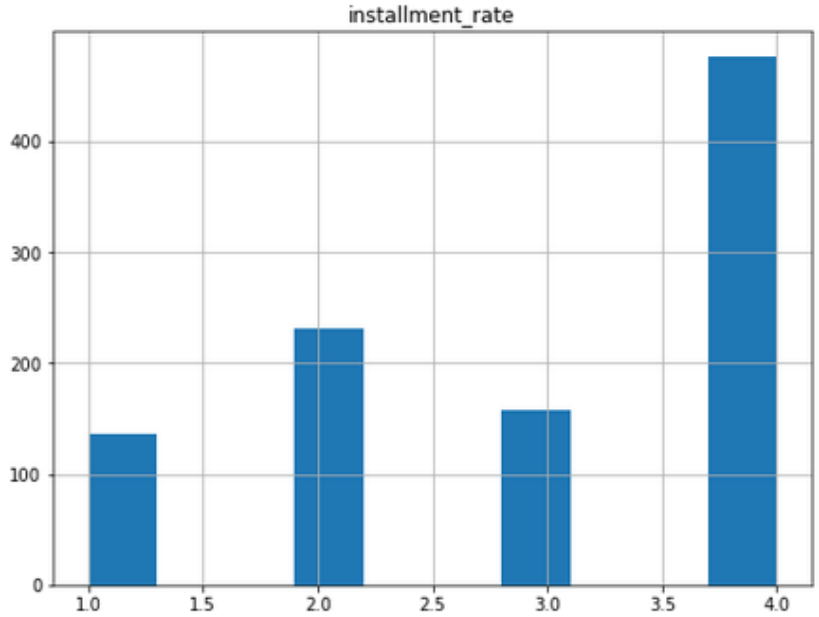
7. *employment_since:* variable nominal que indica la situación laboral del solicitante (ver Figura 7) donde los valores posibles son:
    - A71: desempleado/no residente
    - A72: empleado
    - A73: autónomo
    - A74: profesional/gerente




```
employment_since
A73    0.339
A75    0.253
A74    0.174
A72    0.172
A71    0.062
Name: employment_since
```

Figura 8.- employment_since

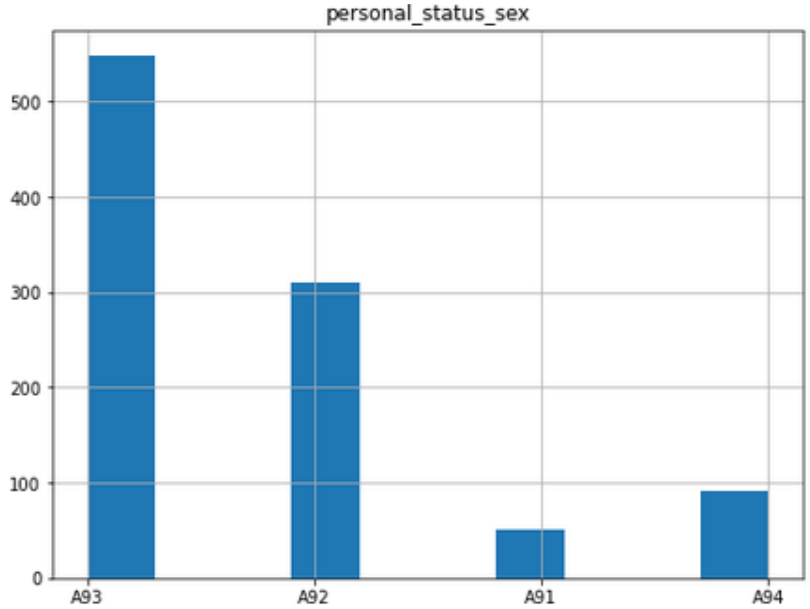
8. *installement_rate:* variable numérica que indica la tasa de interés en % de la ganancia o disposición del crédito (ver Figura 8).




```
installment_rate
4    0.476
2    0.231
3    0.157
1    0.136
Name: installment_rate,
```

Figura 9.- Installment_rate

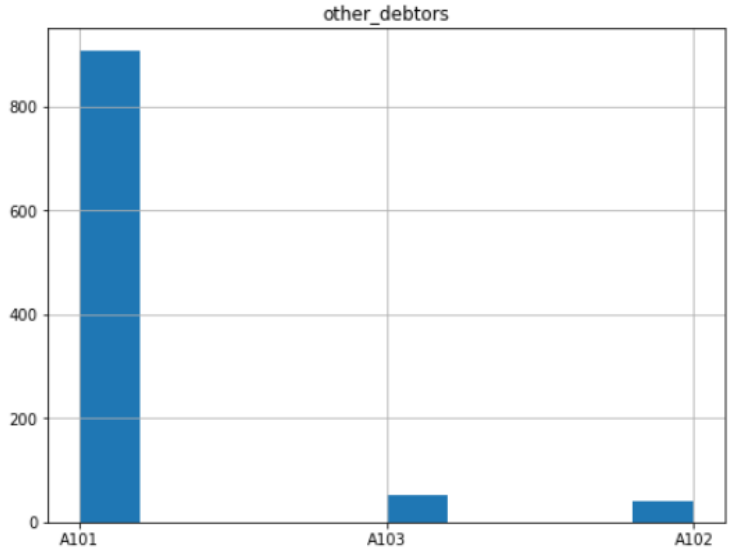
9. *personal_status_sex:* variable nominal que indica el sexo y estado civil del solicitante (ver Figura 9)., donde los valores posibles son:
    - A91: masculino - divorciado/separado
    - A92: femenino - divorciado/separado/viudo
    - A93: masculino - soltero
    - A94: masculino - casado/viudo
    - A95: femenino – soltero




```
personal_status_sex
A93    0.548
A92    0.310
A94    0.092
A91    0.050
Name: personal_status_sex
```

Figura 10.-personal_status_sex

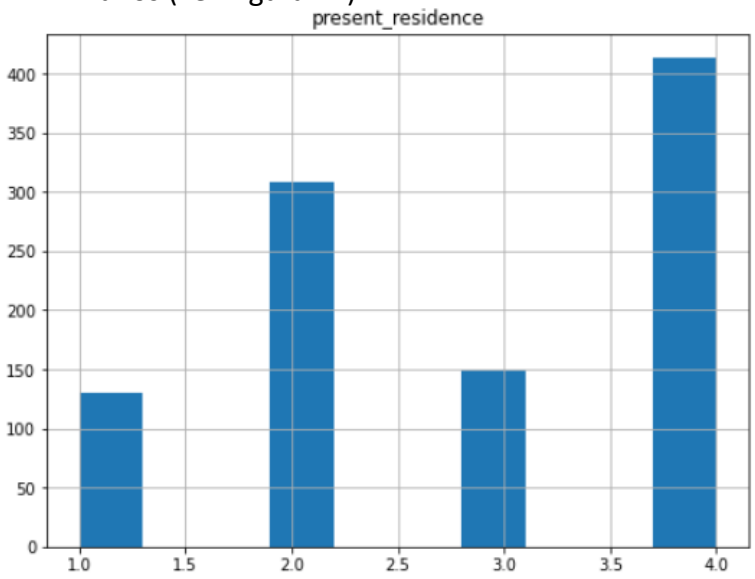
10. *other_debtors:* variable nominal que indica si hay otros deudores o garantes presentes (ver Figura 10) donde los valores posibles son:
    - A101: sin otros deudores o garantes
    - A102: co-solicitante presente
    - A103: garante presente




```
other_debtors
A101    0.907
A103    0.052
A102    0.041
Name: other_debtors,
```

Figura 11.-Other_debtors

11. *present_residence* variable numérica que indica la duración en la dirección actual en años (ver Figura 11).

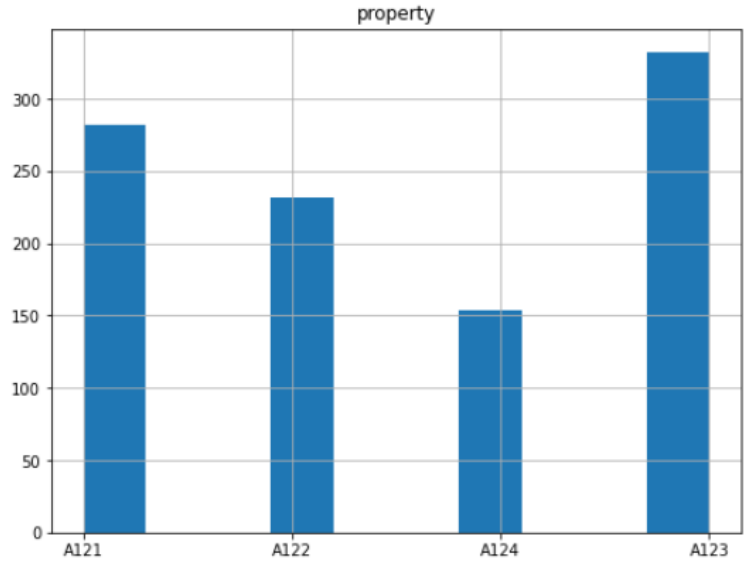



```
present_residence
4    0.413
2    0.308
3    0.149
1    0.130
Name: present residence
```

Figura 12.-Present_residence

12. *property*: variable nominal que indica si el solicitante posee propiedad (ver Figura 12), donde los valores posibles son:
    - A121: propiedad totalmente pagada
    - A122: propiedad en proceso de ser pagada

- A123: no hay propiedad




```
property
A123    0.332
A121    0.282
A122    0.232
A124    0.154
Name: property,
```

Figura 13.-property

13. *age*: variable numérica que indica la edad del solicitante en años (ver Figura 13).

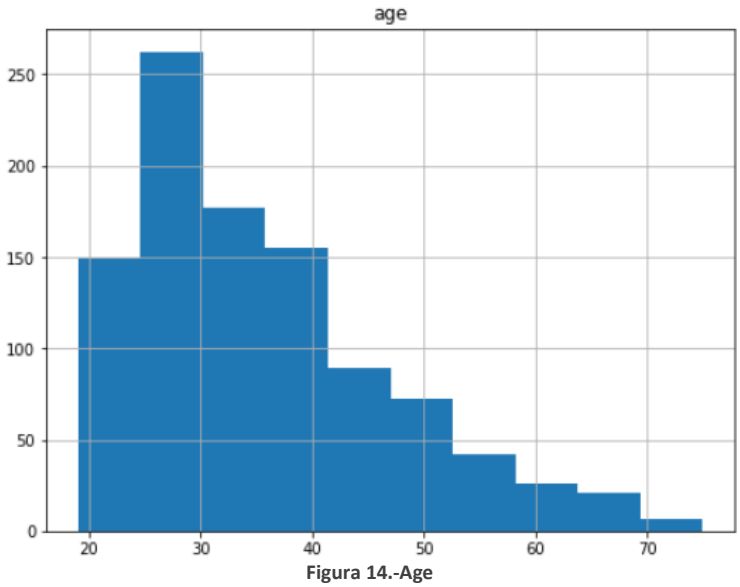
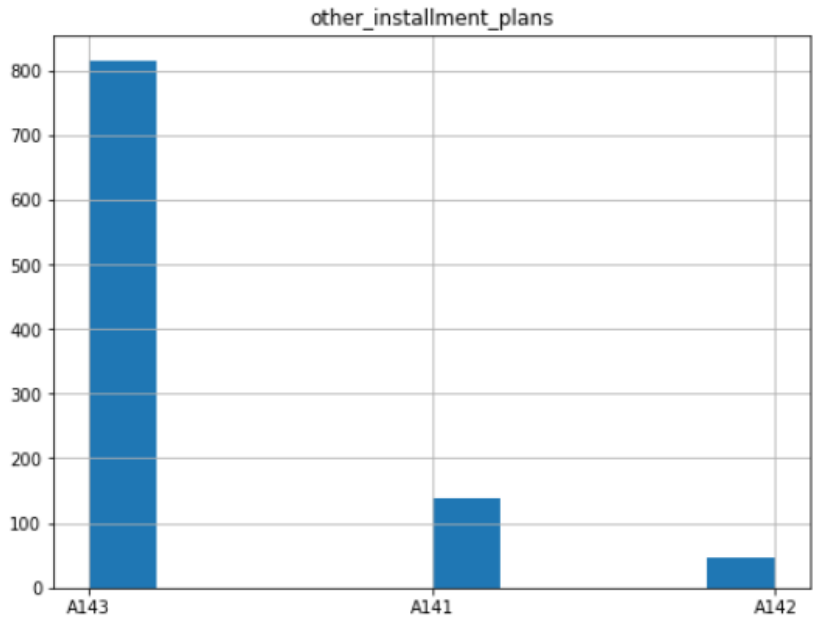


Figura 14.-Age

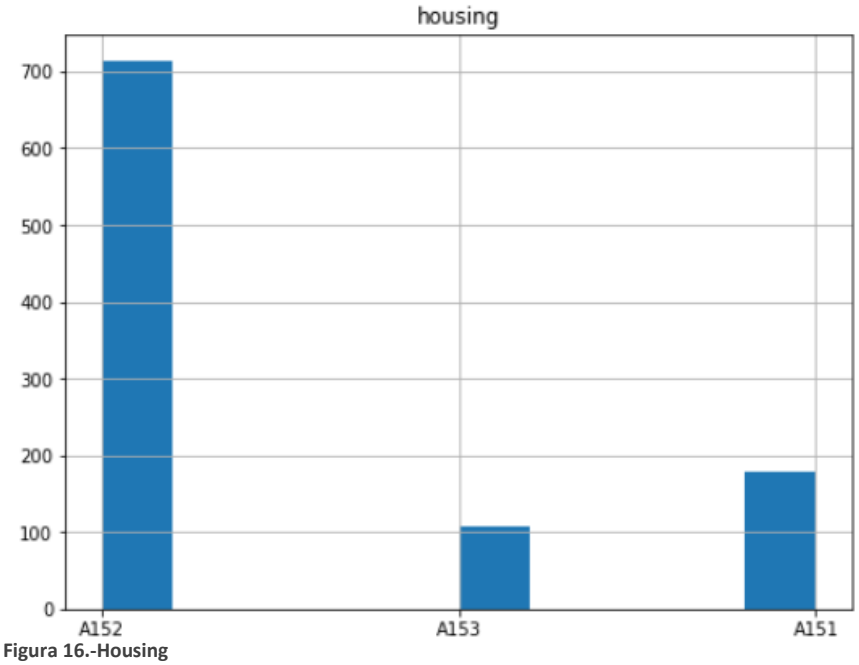
14. *other_installment_plans*: variable nominal que indica si hay otros planes de financiación presentes (ver Figura 14), donde los valores posibles son:
    - A141: banco existente
    - A142: crédito para nuevo propósito
    - A143: ningún plan existente



other_installment_plans
A143 0.814
A141 0.139
A142 0.047

Figura 15.-Other_installment_plans

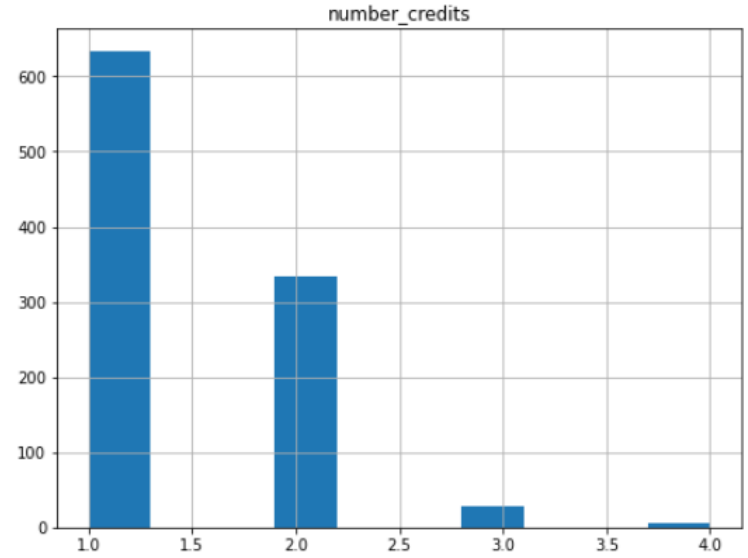
15. *housing*: variable nominal que indica el tipo de vivienda del solicitante (ver Figura 15), donde los valores posibles son:
    - A151: alquiler
    - A152: propiedad propia
    - A153: gratis



housing
A152 0.713
A151 0.179
A153 0.108

Figura 16.-Housing

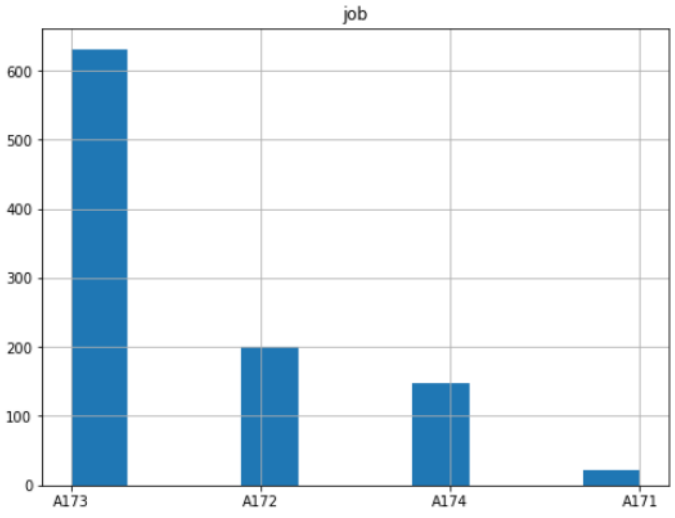
16. *number_credits*: variable numérica que indica el número de créditos existentes en este banco (ver Figura 16).




```
number_credits
1    0.633
2    0.333
3    0.028
4    0.006
Name: number_credits
```

Figura 17.-Number_credits

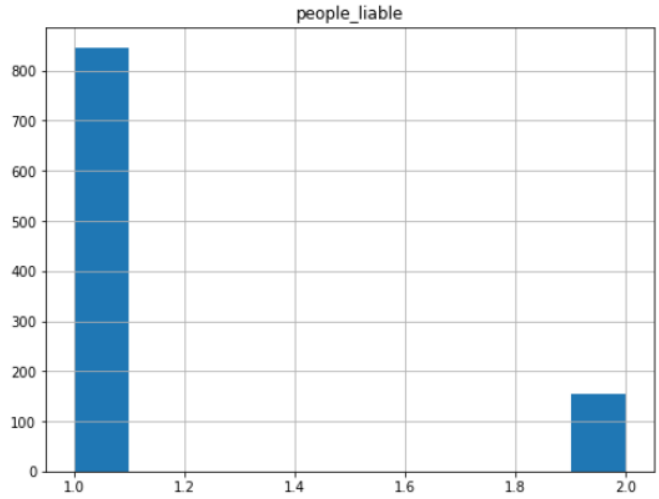
17. *job*: variable nominal que indica el nivel de trabajo del solicitante (ver Figura 17), donde los valores posibles son:
    - A171: desempleado/no residente
    - A172: no calificado/no especializado
    - A173: calificado/especializado
    - A174: gerente/alto nivel




```
job
A173    0.630
A172    0.200
A174    0.148
A171    0.022
```

Figura 18.-Job

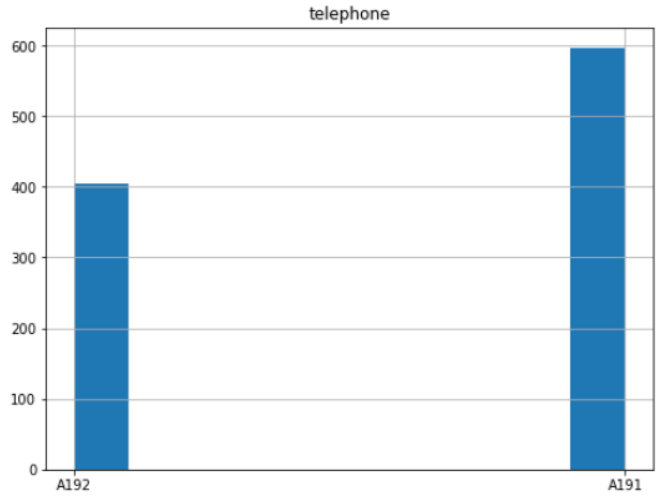
18. *People_liable*: variable numérica que indica el número de personas mantenidas por el solicitante (ver Figura 18).




```
people_liable
1    0.845
2    0.155
Name: people_liable
```

Figura 19.-People_liable

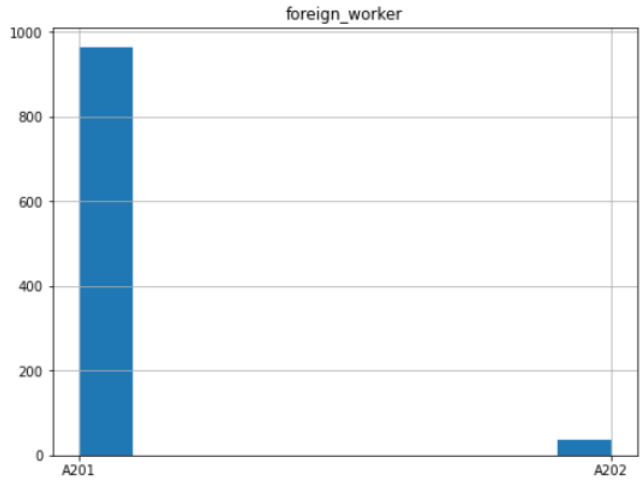
19. *telephone*: variable nominal que indica si el solicitante tiene un teléfono registrado (ver Figura 19), donde los valores posibles son:
    - A191: no registrado
    - A192: registrado+




```
telephone
A191    0.596
A192    0.404
Name: telephone,
```

Figura 20.-Telephone

20. *foreign_worker:* variable nominal que indica si el solicitante es un trabajador extranjero (ver Figura 20), donde los valores posibles son:
    - A201: sí
    - A202: no

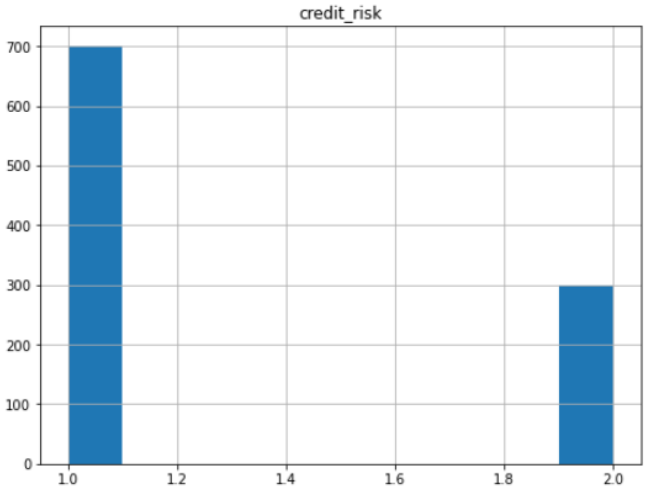



```
foreign_worker
A201    0.963
A202    0.037
```

Figura 21.-Foreign_worker

21. *credit_risk* (ver Figura 21)*:*

- *credit_risk* igual a 1 significa que el préstamo ha sido concedido y el cliente ha cumplido con todas sus obligaciones de pago.
- *credit_risk* igual a 2 significa que el préstamo ha sido concedido, pero el cliente no ha cumplido con todas sus obligaciones de pago o ha cumplido con ellas de manera insuficiente.




```
credit_risk
1    0.7
2    0.3
Name: credit_risk,
```

Figura 22.-Credit_risk

En resumen, esta base de datos es una fuente de información relevante para el análisis de riesgo crediticio. Los datos fueron recopilados de solicitudes de crédito de un banco alemán y contienen información detallada sobre los solicitantes, incluyendo su historial crediticio,

situación laboral, propósito del crédito, monto del crédito, entre otros. Las variables incluyen tanto variables nominales como numéricas, y cada valor de las variables está codificado de forma específica en la base de datos.

2.2. Análisis exploratorio

Se inicia la exploración con una visualización básica de todos los datos, adicional a las gráficas que ya fueron presentadas en la sección anterior.

Se presenta la información de los encabezados de las 21 columnas o variables totales, Figura 23.

```
  status_account  duration credit_history purpose  credit_amount  \
0            A11         6            A34     A43           1169
1            A12        48            A32     A43           5951
2            A14        12            A34     A46           2096
3            A11        42            A32     A42           7882
4            A11        24            A33     A40           4870

  savings_account employment_since  installment_rate personal_status_sex  \
0             A65              A75                 4                 A93
1             A61              A73                 2                 A92
2             A61              A74                 2                 A93
3             A61              A74                 2                 A93
4             A61              A73                 3                 A93

  other_debtors  ...  property age  other_installment_plans housing  \
0          A101  ...       A121  67                     A143    A152
1          A101  ...       A121  22                     A143    A152
2          A101  ...       A121  49                     A143    A152
3          A103  ...       A122  45                     A143    A153
4          A101  ...       A124  53                     A143    A153

  number_credits   job people_liable  telephone foreign_worker credit_risk
0              2  A173             1       A192           A201           1
1              1  A173             1       A191           A201           2
2              1  A172             2       A191           A201           1
3              1  A173             2       A191           A201           1
4              2  A173             2       A191           A201           2
```

Figura 23.-Encabezados

Se describen los tipos de datos de cada una de las 20 variables explicativas y la variable objetivo, Figura 24.

```
[5 rows x 21 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   status_account          1000 non-null   object
 1   duration                1000 non-null   int64
 2   credit_history          1000 non-null   object
 3   purpose                 1000 non-null   object
 4   credit_amount           1000 non-null   int64
 5   savings_account         1000 non-null   object
 6   employment_since        1000 non-null   object
 7   installment_rate        1000 non-null   int64
 8   personal_status_sex     1000 non-null   object
 9   other_debtors           1000 non-null   object
 10  present_residence       1000 non-null   int64
 11  property                1000 non-null   object
 12  age                     1000 non-null   int64
 13  other_installment_plans 1000 non-null   object
 14  housing                 1000 non-null   object
 15  number_credits          1000 non-null   int64
 16  job                     1000 non-null   object
 17  people_liable           1000 non-null   int64
 18  telephone               1000 non-null   object
 19  foreign_worker          1000 non-null   object
 20  credit_risk             1000 non-null   int64
dtypes: int64(8), object(13)
```

Figura 24.-Tipo de datos

Se describen estadísticos básicos para las variables numéricas, Figura 25.

```
          duration  credit_amount  installment_rate  present_residence  \
count  1000.000000    1000.000000       1000.000000        1000.000000
mean     20.903000    3271.258000          2.973000           2.845000
std      12.058814    2822.736876          1.118715           1.103718
min       4.000000     250.000000          1.000000           1.000000
25%      12.000000    1365.500000          2.000000           2.000000
50%      18.000000    2319.500000          3.000000           3.000000
75%      24.000000    3972.250000          4.000000           4.000000
max      72.000000   18424.000000          4.000000           4.000000

               age  number_credits  people_liable  credit_risk
count  1000.000000     1000.000000    1000.000000  1000.000000
mean     35.546000        1.407000       1.155000     1.300000
std      11.375469        0.577654       0.362086     0.458487
min      19.000000        1.000000       1.000000     1.000000
25%      27.000000        1.000000       1.000000     1.000000
50%      33.000000        1.000000       1.000000     1.000000
75%      42.000000        2.000000       1.000000     2.000000
max      75.000000        4.000000       2.000000     2.000000
```

Figura 25.-Descripción inicial

2.2.1. *Variables categóricas*

Además de la exploración básica detallada que se realizó por cada columna en la exploración de los datos, se exploraron cada una de las variables categóricas analizadas contra la variable objetivo con el cálculo del porcentaje de crédito favorable por cada uno de los niveles mediante (2,1):

$$(\%\ \text{Credit risk} = 1)\ \frac{\text{creditrisk} = 1}{\text{creditrisk} = 1 + \text{creditrisk} = 2} \tag{2,1}$$

En gráficos de barras se muestran en las Figuras 25-38.

Esto permite analizar si existe alguna variación entre cada uno de los niveles de las categorías, generando una visualización básica de la importancia de cada una de estas variables. La variable categórica se encuentra en el título del eje horizontal, las barras están ordenadas de menor a mayor y se presenta una leyenda con los máximos y mínimos para determinar el rango de % de crédito favorable por cada uno de los niveles de las variables categóricas. Un rango más alto implica una mayor importancia o impacto en la variable de respuesta.

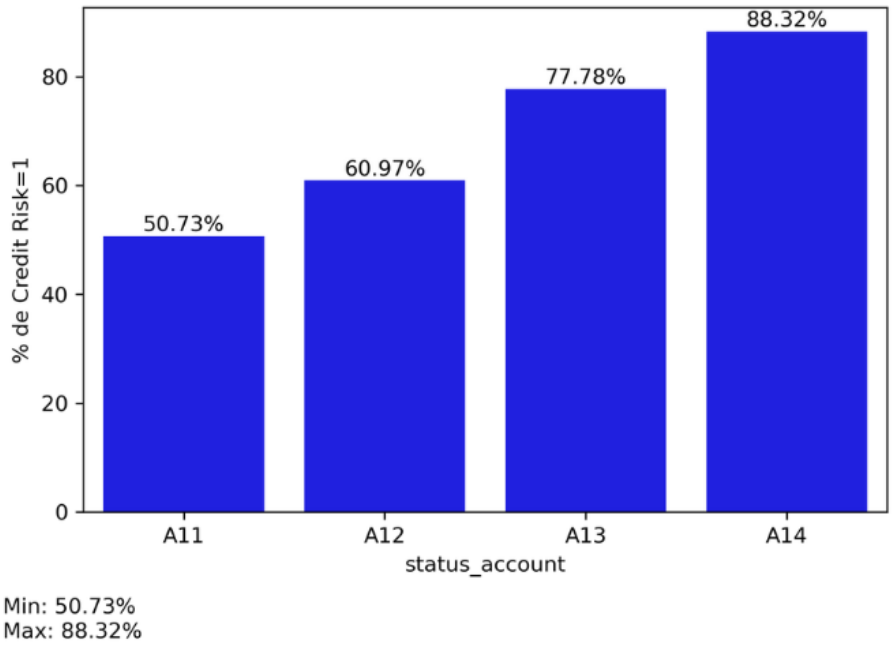
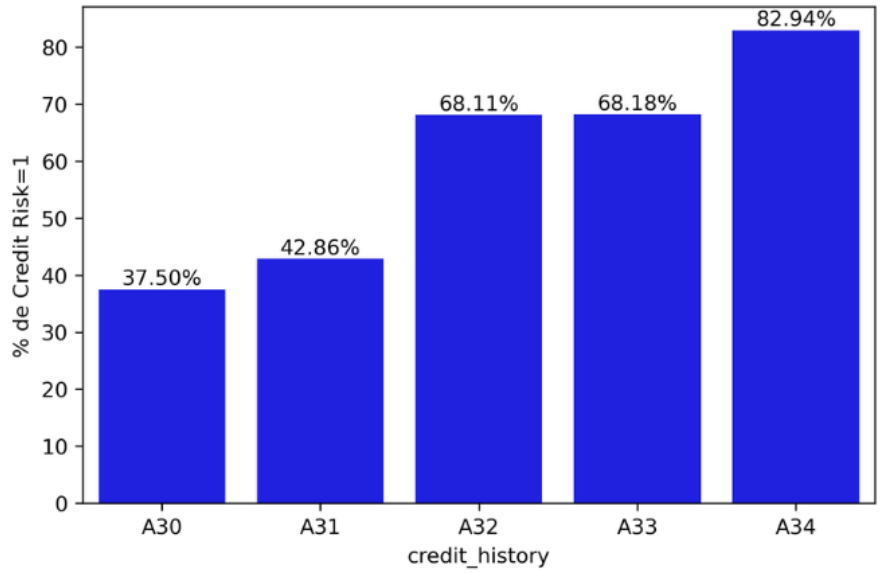
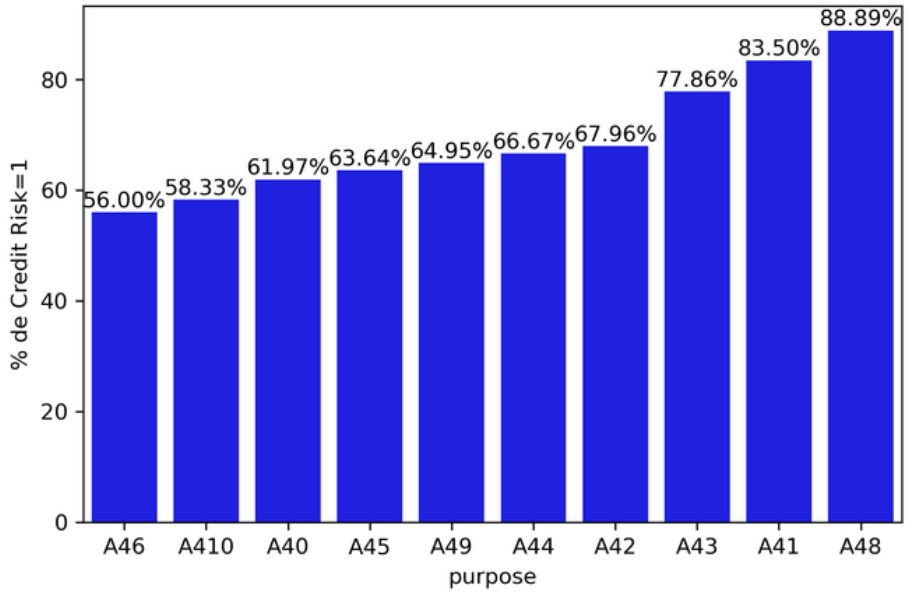


Figura 26.-Split Status Account



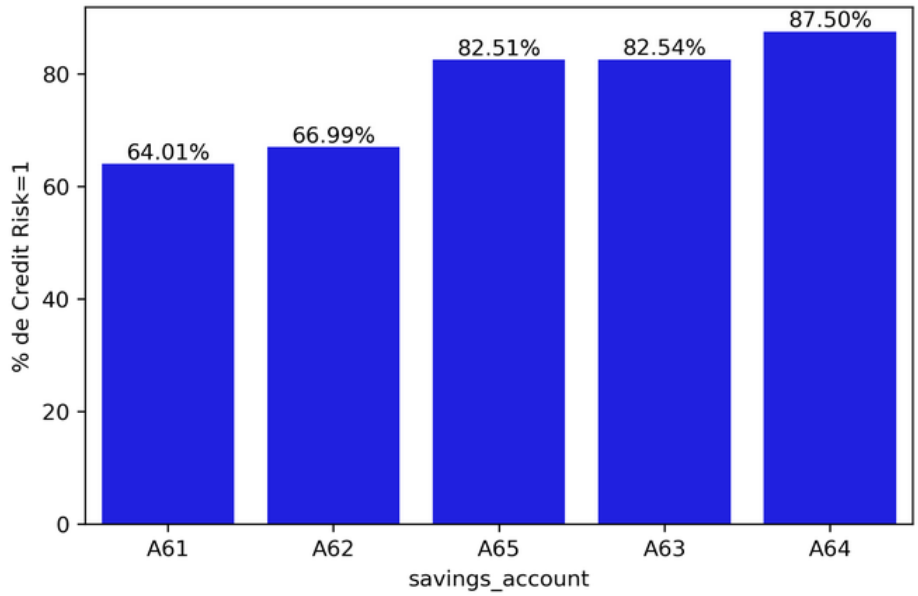
Min: 37.50%
Max: 82.94%

Figura 27.-Split *Credit History*



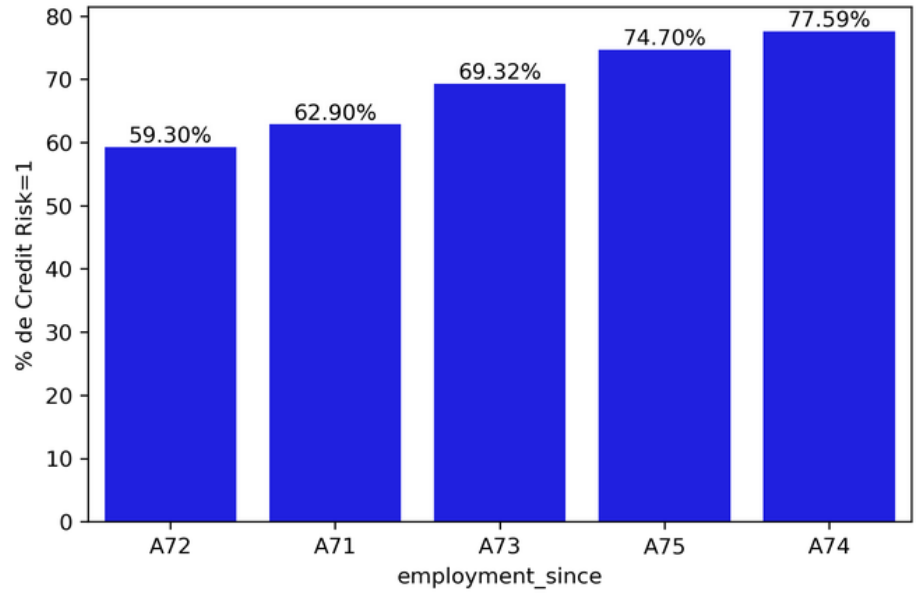
Min: 56.00%
Max: 88.89%

Figura 28.-*Split Purpose*



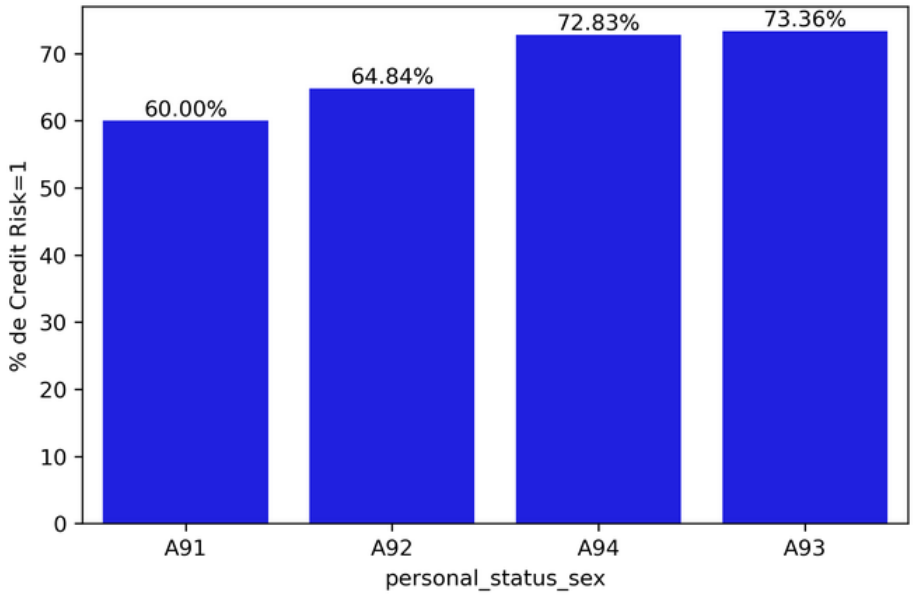
Min: 64.01%
Max: 87.50%

Figura 29.-*Split Savings_account*



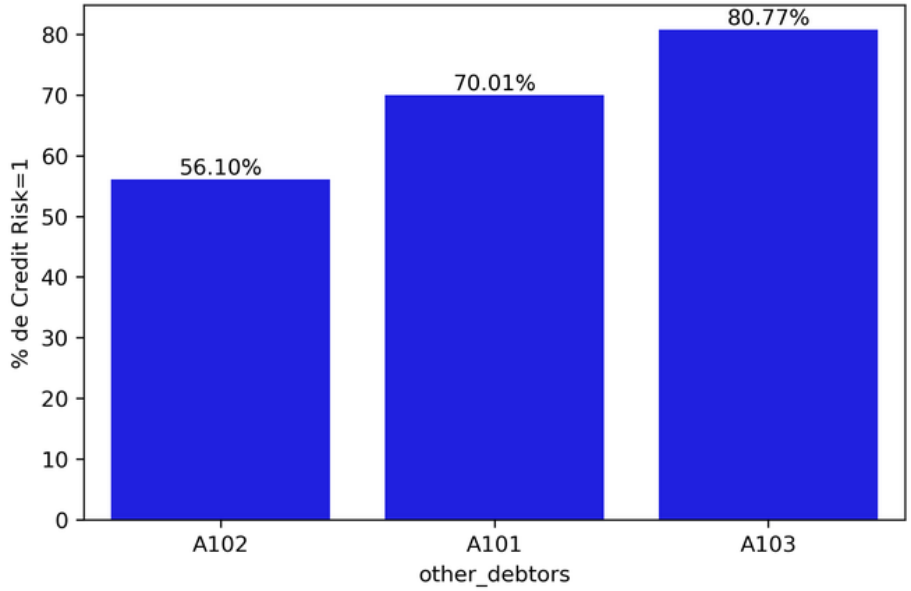
Min: 59.30%
Max: 77.59%

Figura 30.- *Split Employment_Since*



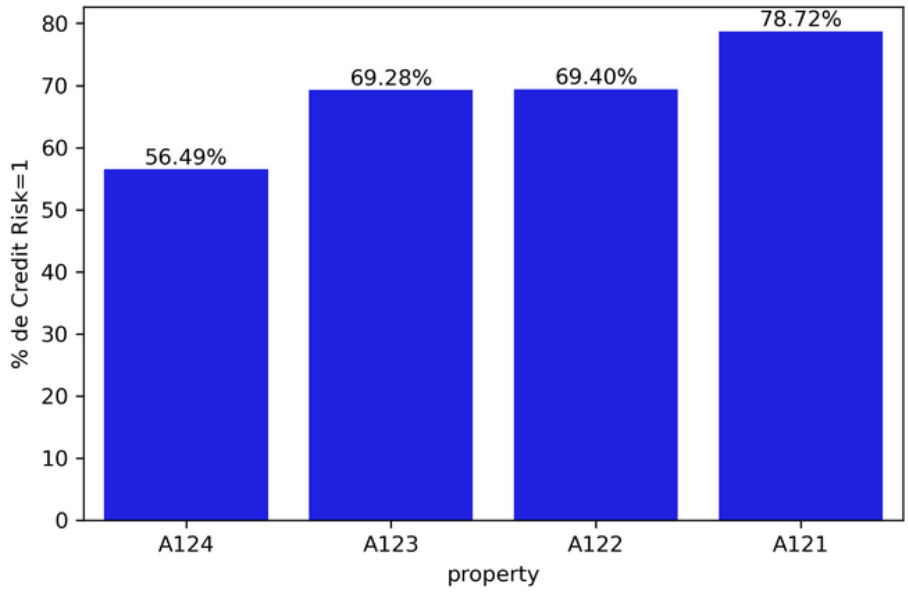
Min: 60.00%
Max: 73.36%

Figura 31.- ***Split Personal Status Sex***



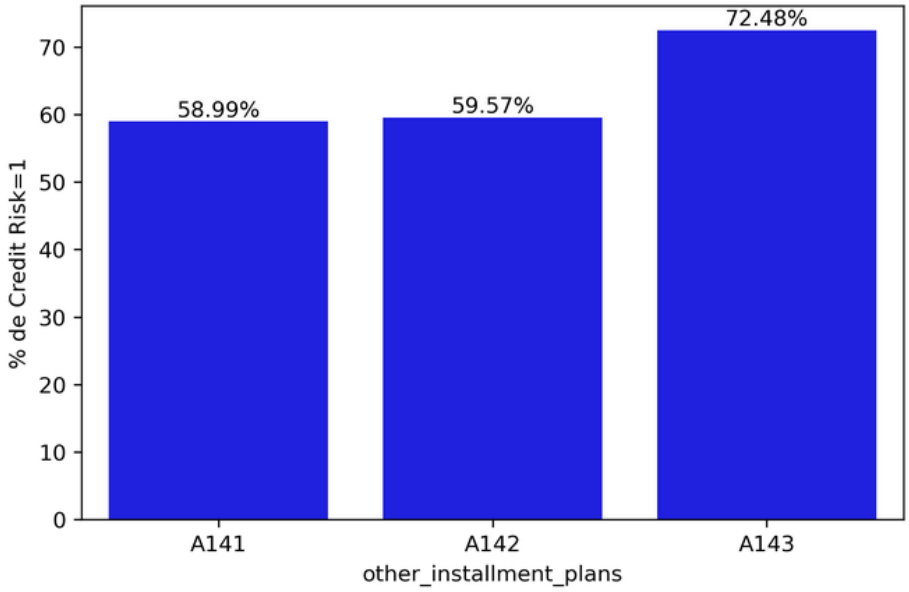
Min: 56.10%
Max: 80.77%

Figura 32.-*Split Other debtors*



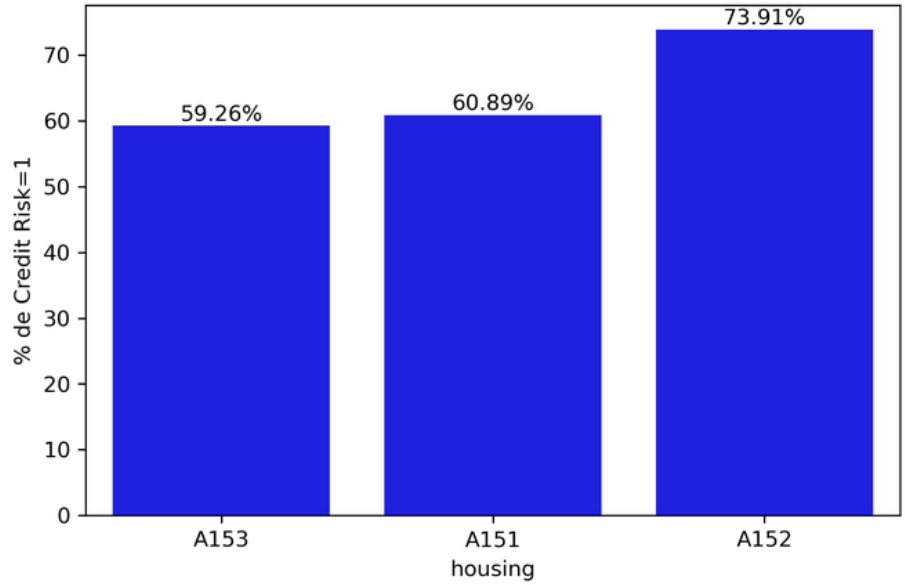
Min: 56.49%
Max: 78.72%

Figura 33.- ***Split property***



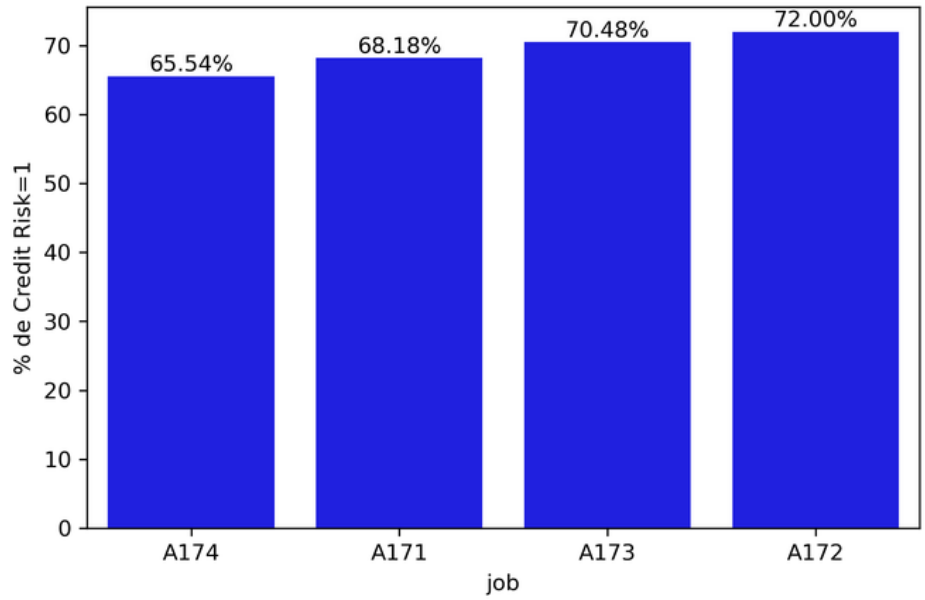
Min: 58.99%
Max: 72.48%

Figura 34.- ***Split Other Installment Plans***



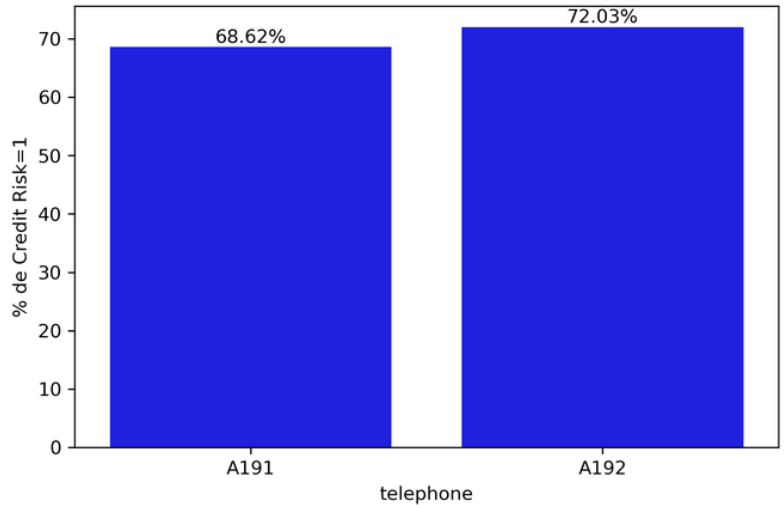
Min: 59.26%
Max: 73.91%

Figura 35.-*Split housing.*



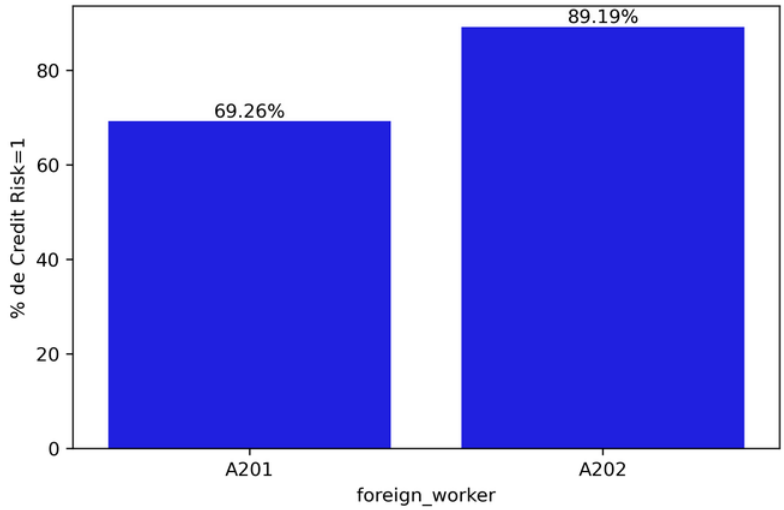
Min: 65.54%
Max: 72.00%

Figura 36.- *Split job*



Min: 68.62%
Max: 72.03%

Figura 37.- ***Split Telephone***



Min: 69.26%
Max: 89.19%

Figura 38.-*Split Foreign Worker*

En conclusión, todas las variables categóricas mostraron una diferencia en el porcentaje de crédito favorable en todos sus niveles, por lo que se utilizarán todas para la generación de un modelo predictivo.

2.2.2. Variables numéricas

Se realizó una matriz de correlación para revisar la importancia de las variables respecto a la variable objetivo, ver Figura 39:

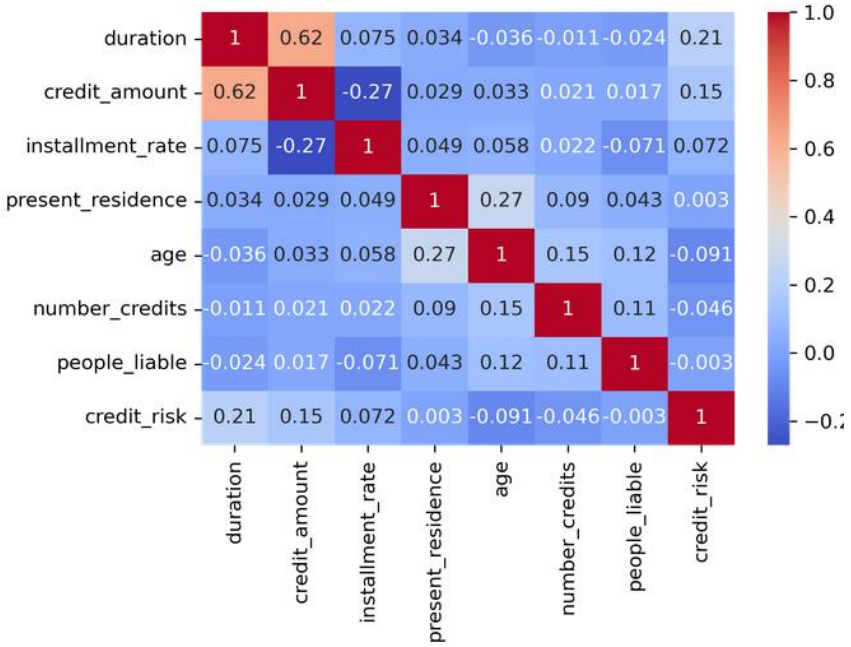


Figura 39.-Matriz de correlación

En la matriz de correlación se puede observar que la variable objetivo *credit_risk* está positivamente correlacionada con la duración del crédito *duration*, el monto del crédito *credit_amount* y la edad del solicitante *age*. Esto significa que a medida que la duración, el monto del crédito y la edad del solicitante aumentan, también lo hace la probabilidad de que se produzca un riesgo crediticio.

Por otro lado, la variable objetivo *credit_risk* está negativamente correlacionada con la tasa de interés del crédito *installment_rate.* Esto significa que a medida que la tasa de interés del crédito aumenta, disminuye la probabilidad de que se produzca un riesgo crediticio. También se puede observar que las variables numéricas tienen una correlación moderada entre sí. Por ejemplo, la duración del crédito *duration* y el monto del crédito *credit_amount* tienen una correlación positiva moderada. Esto sugiere que las personas que solicitan créditos de mayor duración también tienden a solicitar mayores montos de crédito.

En general, la matriz de correlación proporciona información valiosa sobre la relación entre las variables en el conjunto de datos y nos ayuda a identificar posibles patrones y relaciones que pueden ser relevantes para nuestro análisis. En el caso específico de la variable objetivo *credit_risk,* la matriz de correlación sugiere que la duración del crédito, el monto del crédito, la edad del solicitante y la tasa de interés del crédito son variables importantes para considerar en la evaluación del riesgo crediticio.

En conclusión, se opta por sí utilizar todas las variables, pues por la correlación se descarta colinealidad o una correlación muy fuerte que permita descartar alguna variable.

2.2.3. Ingeniería de características

En la experimentación, se utilizaron 3 principales herramientas de ingeniería de características:

- Escalado de variables predictoras.
- Para el tratamiento de variables categóricas:
    - *Label encoder*: Que permite sustituir los niveles de variables categóricas por etiquetas, sin embargo, puede generar una jerarquización falsa.
    - *One hot encoder:* Sustituye todas las variables categóricas por *dummies*, sin embargo, genera columnas adicionales, k-1 variables donde k es la cardinalidad de cada una de las variables categóricas.

2.3. *Descripción de los modelos*

Dada la naturaleza de la base de datos, se pueden utilizar varios modelos de aprendizaje automático para predecir la probabilidad de que una solicitud de crédito sea aprobada o rechazada. A continuación, se justifica el uso de cuatro modelos diferentes para analizar esta base de datos:

1. Regresión logística (*sklearn.linear_model.LogisticRegression*): La regresión logística es una técnica de aprendizaje supervisado utilizada para predecir la probabilidad de una variable binaria (0 o 1). En este caso, la variable binaria sería la aprobación o rechazo de una solicitud de crédito. La regresión logística se utiliza comúnmente en problemas de clasificación binaria como este, ya que proporciona una buena interpretación de los coeficientes y es relativamente rápida en términos de tiempo de entrenamiento y predicción.
2. *Support vector machine (*sklearn.svm.SVC): Las máquinas de vectores de soporte (SVM) son un algoritmo de aprendizaje supervisado que se utiliza para la clasificación y regresión. SVM es una buena opción para conjuntos de datos que tienen muchas características y cuando se requiere alta precisión. En el caso de la base de datos de crédito, SVM podría ser útil debido a la gran cantidad de características que contiene.
3. *Random forest* (*sklearn.ensemble.RandomForestClassifier*): Los bosques aleatorios son una técnica de aprendizaje automático que combina varios árboles de decisión para construir un modelo predictivo. Los bosques aleatorios son conocidos por su capacidad para manejar datos de alta dimensionalidad, conjuntos de datos con muchos atributos y pueden ser eficaces en problemas de clasificación, como la aprobación o rechazo de solicitudes de crédito.

4. Árbol de decisión (*sklearn.tree.DecisionTreeClassifier*): Los árboles de decisión son una técnica de aprendizaje automático que se utiliza para clasificación y regresión. El árbol de decisión se puede utilizar en el análisis de la base de datos de crédito para identificar los factores que tienen un mayor impacto en la aprobación o rechazo de una solicitud de crédito. El árbol de decisión también puede proporcionar una visión general de las diferentes características y su importancia para el problema en cuestión.

Los modelos elegidos para analizar la base de datos de crédito se seleccionaron por su capacidad para manejar grandes conjuntos de datos, su capacidad para manejar datos de alta dimensionalidad y la interpretación de sus resultados. Cada modelo tiene sus propias ventajas y desventajas, y es posible que se requiera más de un modelo para obtener resultados óptimos. La documentación y detalles de cada modelo puede consultarse en [4]-[7].

2.4. Descripción de las métricas

La medida de *Accuracy* es una medida comúnmente utilizada para evaluar el desempeño de modelos de clasificación en los que se tienen dos clases y ambas tienen la misma importancia. En el conjunto de datos el objetivo es predecir si un solicitante de crédito es "bueno" o "malo" basándose en características financieras y personales. La variable objetivo, tiene dos valores posibles: 1 si el solicitante es "bueno" y 2 si el solicitante es "malo".

Dado que el objetivo principal de este problema es clasificar a los solicitantes de crédito correctamente, la medida de *Accuracy* es una métrica adecuada para evaluar el rendimiento del modelo. La medida de *Accuracy* se define como la proporción de predicciones correctas en relación con el total de predicciones realizadas. En el caso de este conjunto de datos, un alto valor de *Accuracy* indicaría que el modelo es capaz de clasificar correctamente a la mayoría de los solicitantes de crédito como "buenos" o "malos".

Dado que el objetivo principal del conjunto de datos es clasificar correctamente a los solicitantes de crédito como "buenos" o "malos", la medida de *Accuracy* es una métrica adecuada para evaluar el rendimiento de los modelos de clasificación.

En el contexto de comparar cuatro modelos de clasificación para este conjunto de datos, la métrica de *Accuracy* es suficiente para evaluar y seleccionar el mejor modelo. Si un modelo tiene una mayor tasa de precisión que los otros, significa que es mejor para clasificar correctamente a los solicitantes de crédito. Además, dado que los cuatro modelos son modelos de clasificación y el problema es binario, es apropiado utilizar la medida de *Accuracy* para comparar su desempeño (2,2).

$$Accuracy = \frac{(True\ negatives + True\ positives)}{True\ Positive + False\ Negative + True\ Negaive + False\ \ Negative} \quad (2,2)$$

2.5. Descripción de los experimentos / simulaciones

Experimento 1:
Se realizaron los siguientes pasos generales:

1. Se codificaron las variables categóricas usando la función *LabelEncoder* de *sklearn*.
2. Se separaron las variables predictoras (X) y la variable objetivo (y).
3. Se dividió el conjunto de datos en entrenamiento y prueba usando la función *train_test_split* de *sklearn* con un tamaño de prueba del 30% y una semilla aleatoria de 42.
4. Se escalaron las variables predictoras usando la función *StandardScaler* de *sklearn*.
5. Se ajustaron cuatro modelos de clasificación diferentes: Regresión Logística, Árboles de Decisión, Bosques Aleatorios y SVM, con sus respectivos hiperparámetros por defecto.
6. Se realizó la predicción sobre el conjunto de prueba y se calculó la exactitud (*Accuracy*) de cada modelo.

```
Accuracy SVM: 0.7566666666666667
Accuracy LR: 0.78
Accuracy DT: 0.6933333333333334
Accuracy RF: 0.77
```

Experimento 2:
Mismo procedimiento que el experimento anterior, pero se utilizó *OneHotEncoder* en lugar de *label encoder* para codificar las variables categóricas para evitar una jerarquización falsa generada por *label encoder.*

```
Accuracy SVM: 0.7233333333333334
Accuracy LR: 0.73
Accuracy DT: 0.6666666666666666
Accuracy RF: 0.7566666666666667
```

Experimento 3:
Se realizó también el escalamiento de los datos y la codificación mediante *label encoder* (porque tuvo mejores resultados ese tratamiento de variables categóricas que el *OneHorEncoder*)

Se realizó un *grid search* con los mismos 4 modelos para poder encontrar los mejores hiperparámetros posibles.

```
Logistic Regression:
Best Score: 75%
Best Params: {'C': 18, 'penalty': 'l1', 'solver': 'liblinear'}
Decision Tree:
Best Score: 74.42%
```

```
Best Params: {'criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 1,
'min_samples_split': 2}
```
Random Forest:
```
Best Score: 76.8%
Best Params: {'bootstrap': False, 'criterion': 'entropy', 'max_depth': 8,
'min_samples_leaf': 4, 'min_samples_split': 16, 'n_estimators': 97}
```
SVM:
```
Best Score: 76.14%
Params: {'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}
```

Experimento 4:

Se realizó una prueba con clases balanceadas enfocado exclusivamente *Random Forest* con *SMOTE* (*Synthetic Minority Oversampling Technique)* y preprocesando variables categóricas con *get_dummies*

Accuracy Random Forest: 75%

Experimento 5:

Se realizó nuevamente una prueba con *Random Forest*, se estandarizaron con *StandardScaler*, se balancearon con *resample* de la biblioteca *sklearn.utils.* y se realizó una búsqueda de hiperparámetros, se agregó una validación cruzada con 5 *folds*, y se utilizó una partición de entrenamiento y prueba de 80-20%.

Accuracy: 79.5%

Experimento 6:

Se utilizan las mismas características del experimento 5, pero el balanceo se realiza con *SMOTE*

Accuracy: 81.78%

Experimento 7:

Con el propósito de comprobar si el resto de los modelos pueden funcionar de la misma manera, se ejecutaron con la misma preparación y metodología del experimento 6 también con un *grid_search.*

Random Forest:
```
Best Score: 81.78%
Random Forest best params: {'n_estimators': 50, 'min_samples_split': 2,
'min_samples_leaf': 1, 'max_depth': 15}
```

SVM:
```
Best Score: 83.92%
SVM best params: {'kernel': 'rbf', 'gamma': 'auto', 'C': 100}
```

Logistic Regression:
```
Best Score: 75.35%
Logistic Regression best params: {'solver': 'newton-cg', 'C': 0.1}
```

Decision Tree:
```
Best Score: 71.07%
Decision Tree best params: {'min_samples_split': 10, 'min_samples_leaf':
2, 'max_depth': 25, 'criterion': 'entropy'}
```

3. RESULTADOS Y DISCUSIÓN

En esta sección se presentan los resultados obtenidos de la metodología aplicada y se contextualizan los resultados con respecto a sus carencias.

3.1. Resultados y discusión

El principal propósito de la regresión logística, por ser el modelo más simple y con una de las mejores interpretabilidades al conocer los coeficientes de cada variable y su participación directa en el resultado, fue funcionar como marco de referencia para el resto de los con un preprocesamiento básico de datos.

Tabla 1.-Resultados de experimentación. Se muestra el *Accuracy* por cada modelo y experimento realizado.

Accuracy	Modelo			
Número de experimento:	Regresión logística	*Support Vector Machine*	Árbol de decisión	Bosque aleatorio
Experimento 1	78.00%	75.66%	69.33%	77.00%
Experimento 2	73.00%	72.33%	66.66%	75.66%
Experimento 3	75.00%	76.14%	74.42%	76.80%
Experimento 4	No participó	No participó	No participó	75.00%
Experimento 5	No participó	No participó	No participó	79.50%
Experimento 6	No participó	No participó	No participó	81.78%
Experimento 7	**75.35%**	**83.92%**	**71.07%**	**81.78%**

Como se observa en la Tabla 1, en el primer experimento fue posible alcanzar un *Accuracy* de 78% usando regresión logística. En los experimentos 2 y 3 el *Accuracy* no mejoró, sin embargo, se notó que con cada paso añadido en cada experimento el *Accuracy* de Bosque Aleatorio comenzó a mejorar más. En el experimento 5, el modelo de bosque aleatorio mejoró su desempeño en un 4.5 % con respecto al experimento 4 (*Accuracy* de 79.5% con respecto a 75 %), y en que en el experimento 6 llegó a 81.78%. En el experimento 7 se determinó correr el resto de los modelos con el nuevo conjunto preprocesado de datos. Con el *Support Vector Machine* con un *kernel RBF (Radial basis function),* que es uno de los más utilizados y populares *Kernels* para clasificación*,* se logró un *Accuracy* de 83.92%.

Respecto al rendimiento de los modelos, como puede observarse en la Tabla 2, el modelo que más consume tiempo y procesamiento fue *Random Forest,* con tiempo de ejecución de 3.59 segundos. Además de su *Accuracy* superior al resto de modelos a través de los experimentos, el modelo de *Support Vector Machine* tuvo mejor rendimiento computacional (ver Tabla 2), dado que, en segundos, fue más rápido y consumió menos recursos computacionales que el segundo mejor modelo (*Random Forest)*. Esto pudiese ser importante en términos de escalabilidad, dado que, con bases de datos más grandes, un modelo que use hasta un 21 % menos recursos puede preferible incluso si su rendimiento es ligeramente peor que otro modelo más costoso, aunque no fue el caso.

Tabla 2.- Rendimiento de los modelos de acuerdo con su tiempo de ejecución y recursos computacionales en el experimento 7.

Rendimiento	% de uso de CPU	Tiempo en segundos	% con base en el modelo más lento
Árboles Aleatorios	82%	3.59	0%
SVM	54%	2.84	-21%
Regresión Logística	58%	0.12	-97%
Árboles de Decisión	87%	0.13	-96%

Como se especifica en la descripción de los experimentos, del número 4 al 6 (ver Tabla 1), el propósito fue mejorar el preprocesamiento de datos para mejorar el bosque aleatorio. Cuando se encontró el mejor resultado en el experimento 6 se decidió aplicar los otros modelos para poder hacer una comparación más precisa. Es por esto por lo que en los experimentos 4-6 no se realizaron pruebas con el resto de los modelos (que se muestran como etiqueta “No participó” en la Tabla 1).

En las conclusiones se harán referencia a los experimentos por número, los específicos pueden ser consultados en la sección de experimentación donde fueron descritos los elementos que cambiaron.

4. CONCLUSIONES

En esta sección se presentan las conclusiones de los resultados de la experimentación y del trabajo en general, así como las recomendaciones del trabajo futuro.

4.1. Conclusiones

En este trabajo se buscó mejorar la evaluación del riesgo crediticio utilizando el conjunto de datos *South German Credit* y se seleccionó el mejor modelo de clasificación entre Regresión Logística, *Support Vector Machine*, Árboles Aleatorios (*Random Forest)* y Árboles de Decisión (Decision Trees).

Se analizó un conjunto de datos focalizado en riesgo crediticio, identificando variables clave como duración del crédito, monto, edad del solicitante y tasa de interés. A través de siete experimentos que variaron en el tratamiento de variables categóricas, se evaluó la efectividad de varios modelos de clasificación utilizando la métrica de *Accuracy*. El modelo de *Support Vector Machine* ajustado con *grid search* sobresalió con un *Accuracy* del 83.92%, seguido de cerca por *Random Forest* con un 81.78%.

El siguiente aspecto para considerar es el uso de recursos, el *Support Vector Machine con* tiempo de ejecución 21% menor que *Random Forest*, 2.84 segundos contra 3.59 segundos (Tabla 2). Donde sí hay una diferencia notable en uso de recursos fue en los modelos de *Logistic Regression* y *Decision Tree* con un tiempo de ejecución de 0.12 y 0.13 segundos, 97% y 96% respectivamente menor tiempo de ejecución contra el modelo más lento, *Random Forest*. Considerando el uso de recursos como factor importante, podría optarse por el modelo de *Logistic Regression,* el de mejor *Accuracy* de los dos modelos más rápidos, sin embargo, tendría un impacto importante en el *Accuracy*, dado que bajaría a 75.35%. El mejor modelo de alto *Accuracy* fue *Support Vector Machine.* Considerando que en eficiencia de recursos computacionales es muy superior, en caso de que pueda sacrificarse el *Accuracy*, el modelo *Logistic Regression* sería buena alternativa. Para propósito de este trabajo, el mejor modelo encontrado definitivamente fue *Support Vector Machine.*

4.2. Trabajo Futuro

Aunque esta es una base de datos que no es reciente, y por ende probablemente no pueda ser utilizada en una base de datos bancaria actual; es importante el precedente que genera este respecto a la importancia del balanceo de las clases para lograr un mejor *Accuracy*, especialmente para los modelos de SVM y *Random Forest*. Es también muy importante previo a considerar solo la métrica de *Accuracy* como el evaluador de los modelos, si es relevante o no hacer el análisis de la matriz de confusión para determinar los falsos positivos y falsos negativos.

El modelo predictivo basado en el conjunto de datos South German Credit puede llevar a una evaluación del riesgo crediticio más precisa y eficiente, reemplazando los procesos manuales o semiautomáticos propensos a errores y subjetividades, esto último especialmente en instituciones pequeñas que no tengan los modelos avanzados de datos a los que sí podrían tener acceso las más grandes, y, desde otra perspectiva, también podría ser aplicado

localmente a empresas que otorguen créditos a sus clientes, adaptando el modelo a sus necesidades particulares.

El potencial de mejora en este modelo es significativo y se puede lograr a través de varias estrategias. Uno de los aspectos clave que podría ser objeto de mejora, es explorar la codificación de variables, además de las ya realizadas previamente, para evaluar si alguna de ellas mejorar el rendimiento del modelo. La elección de la técnica de codificación adecuada puede tener un impacto importante en la capacidad del modelo para capturar relaciones entre las variables y por ende en su *Accuracy*.

Para evaluar la robustez del modelo y su capacidad de adaptación a diferentes contextos, sería adecuado utilizar otras bases de datos crediticias además del conjunto de datos South *German Credit*. La inclusión de bases de datos más amplias y variadas podría proporcionar información valiosa sobre cómo los modelos se desempeñan en diferentes situaciones y podría revelar patrones que no se observaron en este estudio.

BIBLIOGRAFÍA

[1] Springer, “Responsible Credit Risk Assessment with Machine Learning and Knowledge Acquisition”, 2023 [Online]. Disponible en: https://link.springer.com/article/10.1007/s42786-020-00020-3 [Acceso en: 21 de agosto de 2023]

[2] Banco de México, "Reporte de estabilidad financiera",2022 [Online]. Disponible: [https://www.banxico.org.mx/publicaciones-y-prensa/reportes-sobre-el-sistema-financiero/%7BC91285A1-2305-6839-FCD2-D310D5D70749%7D.pdf]. [Acceso en: 27 de junio de 2023].

[3] UCI Machine Learning Repository, "Statlog (German Credit Data) Data Set", 1994. [Online]. Disponible en: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29 [Acceso en: 26 de abril de 2023].

[4] *scikit-learn, "LogisticRegression", versión 1.2.2, 2023. [Online]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html [Acceso en: 26 de abril de 2023].*

[5] *scikit-learn, "SVC", versión 1.2.2, 2023. [Online]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html [Acceso en: 26 de abril de 2023].*

[6] *scikit-learn, "RandomForestClassifier", versión 1.2.2, 2023. [Online]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [Acceso en: 26 de abril de 2023].*

[7] *scikit-learn, "DecisionTreeClassifier", versión 1.2.2, 2023. [Online]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html [Acceso en: 26 de abril de 2023].*