Taller 1

Andrés García Montoya

Juan José Gómez Arenas

Técnicas de Aprendizaje de Maquina

PONTIFICIA UNIVERSIDAD JAVERIANA

BOGOTÁ D.C

2024

i. Introducción

El presente documento presenta el proceso de construcción y los resultados obtenidos en el taller. Este taller tiene como objetivo extraer, preparar y utilizar un conjunto de datos de un conjunto de datos de riesgo de crédito alemán, para hacer un análisis de clasificación diferentes y evidenciar su la diferencia de diferentes modelos en términos de su capacidad de clasificación.

ii. Comprensión del Dataset

El dataset contiene información relacionada con personas que toman un crédito en un banco, donde cada persona es clasificada como un buen o mal riesgo de crédito con respecto a los datos proporcionados.

Dentro del conjunto de datos se pueden evidenciar 1000 filas y 21 columnas, las cuales tienen las siguientes características:

Atributo 1: (cualitativo)

- Estado de la cuenta corriente existente
    - A11: ... < 0 DM
    - A12: 0 <= ... < 200 DM
    - A13: ... >= 200 DM / asignaciones salariales durante al menos 1 año
    - A14: sin cuenta corriente

Atributo 2: (numérico)

- Duración en meses

Atributo 3: (cualitativo)

- Historial de crédito
    - A30: no se han tomado créditos / todos los créditos pagados a tiempo

    - A31: todos los créditos en este banco pagados a tiempo
    - A32: créditos existentes pagados a tiempo hasta ahora
    - A33: retraso en el pago en el pasado
    - A34: cuenta crítica / otros créditos existentes (no en este banco)

Atributo 4: (cualitativo)

- Propósito
    - A40: coche (nuevo)
    - A41: coche (usado)
    - A42: muebles/equipamiento
    - A43: radio/televisión
    - A44: electrodomésticos
    - A45: reparaciones
    - A46: educación
    - A47: (vacaciones - no existe?)
    - A48: reciclaje profesional
    - A49: negocios
    - A410: otros

Atributo 5: (numérico)

- Monto del crédito

Atributo 6: (cualitativo)

- Cuenta de ahorros/bonos
    - A61: ... < 100 DM
    - A62: 100 <= ... < 500 DM

- A63: 500 <= ... < 1000 DM
- A64: ... >= 1000 DM
- A65: desconocido / sin cuenta de ahorros

Atributo 7: (cualitativo)

- Empleo actual desde
  - A71: desempleado
  - A72: ... < 1 año
  - A73: 1 <= ... < 4 años
  - A74: 4 <= ... < 7 años
  - A75: ... >= 7 años

Atributo 8: (numérico)

- Tasa de cuotas en porcentaje del ingreso disponible

Atributo 9: (cualitativo)

- Estado civil y sexo
  - A91: hombre: divorciado/separado
  - A92: mujer: divorciada/separada/casada
  - A93: hombre: soltero
  - A94: hombre: casado/viudo
  - A95: mujer: soltera

Atributo 10: (cualitativo)

- Otros deudores/avalistas
  - A101: ninguno
  - A102: co-solicitante

    - A103: avalista

Atributo 11: (numérico)

- Desde hace cuanto es dueño de la residencia actual

Atributo 12: (cualitativo)

- Propiedad
    - A121: bienes raíces
    - A122: si no A121: acuerdo de ahorro de sociedad constructora / seguro de vida
    - A123: si no A121/A122: coche u otro, no en el atributo 6
    - A124: desconocido / sin propiedad

Atributo 13: (numérico)

- Edad en años

Atributo 14: (cualitativo)

- Otros planes de cuotas
    - A141: banco
    - A142: tiendas
    - A143: ninguno

Atributo 15: (cualitativo)

- Vivienda
    - A151: alquilado
    - A152: propio
    - A153: gratis

Atributo 16: (numérico)

- Número de créditos existentes en este banco

Atributo 17: (cualitativo)

- Trabajo
    - A171: desempleado / no cualificado - no residente
    - A172: no cualificado - residente
    - A173: empleado cualificado / oficial
    - A174: gestión / autónomo / empleado altamente cualificado / oficial

Atributo 18: (numérico)

- Número de personas dependientes

Atributo 19: (cualitativo)

- Teléfono
    - A191: ninguno
    - A192: sí, registrado a nombre del cliente

Atributo 20: (cualitativo)

- Trabajador extranjero
    - A201: sí
    - A202: no

Además de estas 20 columnas, esta la columna objetivo "Category", que representa dos posibles categorías. La primera categoría representa a las personas que son un buen riesgo para el banco para concederles créditos, el cual esta dado por el valor **(1)**; por otra parte, la segunda categoría representa a las personas que son un mal riesgo para la concesión de créditos, la cual esta dada por el valor **(2).**

En el contexto del problema, se da a entender que es mas grave que una persona sea clasificada como (1) cuando en realidad es (2), que sea clasificada como (2) cuando en realidad es (1).

Para determinar las variables que se vayan a utilizar para el modelo, se verifico la distribución y la correlación que tienen las variables respecto a la variable objetivo “Category”

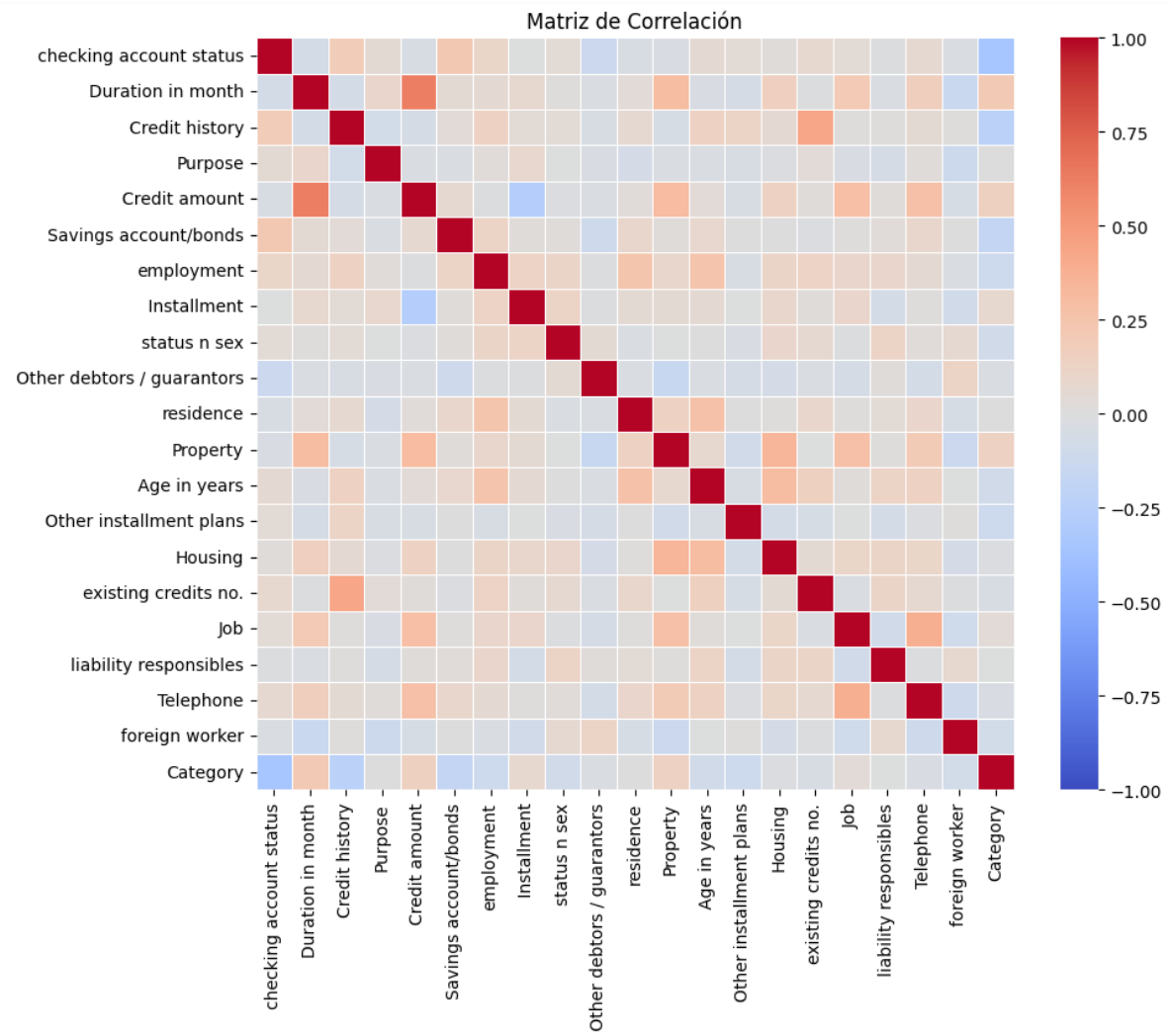

Fig. 1.1. Correlación de variables

Por otra parte, como se puede evidenciar en la Fig. 1.2., existe un mayor número de datos que contienen los valores de (1), lo cual indica que la mayoría de las personas dentro del

dataset representan un buen riesgo crediticio. No obstante, este desbalanceo entre las clases puede afectar el rendimiento del modelo.

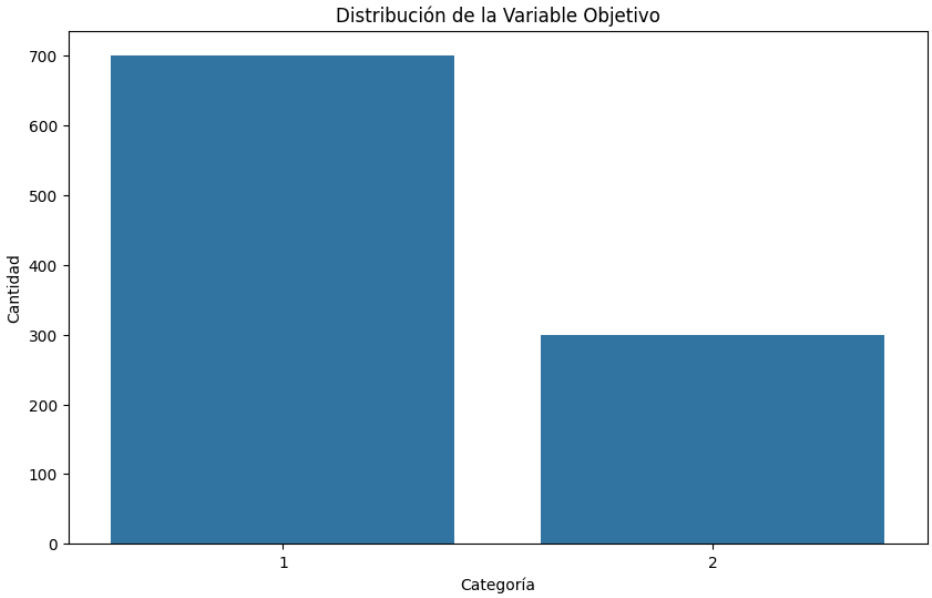

Fig. 1.2. Número de datos por categoría

Por último, en cuanto a las distribuciones de las variables que se aprecian en la Fig. 1.3., la mayoría de los préstamos tienen una duración entre 6 y 24 meses, con pocos superando los 60 meses.

La “Credit Amount” están sesgados a la derecha, predominando los montos bajos y con pocos créditos altos. Las cuotas suelen ser altas, especialmente de valor 4, y hay menos créditos con cuotas bajas.

La mayoría de los individuos han residido en su domicilio actual durante 4 años, seguidos por aquellos con 2 años; otros periodos son menos comunes.

La edad de los individuos también está sesgada a la derecha, con la mayoría entre 20 y 40 años y pocos mayores de 60 años.

Finalmente, la mayoría tiene un crédito existente, con menos personas con dos y aún menos con más de dos créditos. En cuanto a las responsabilidades financieras, la mayoría tiene una sola responsabilidad, y un número menor tiene dos.

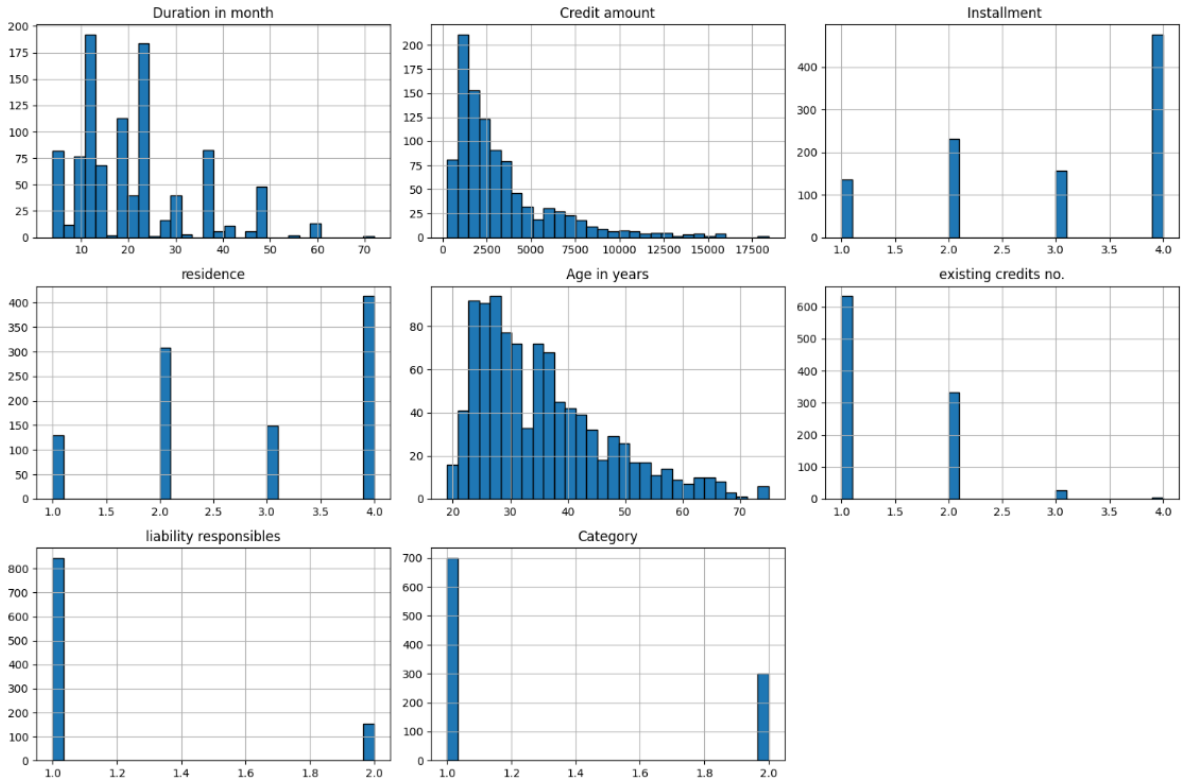

Fig. 1.3. Distribución general de las variables numéricas

iii. Limpieza de datos

Para el modelo de Bayes, se decidió utilizar todas las variables disponibles en el conjunto de datos. Aunque las correlaciones con la variable objetivo 'Category' varían entre las diferentes características, se justificó la inclusión de todas las variables para aprovechar la capacidad del modelo de Bayes de manejar múltiples características y suposiciones de independencia. Esto permite que el modelo evalúe cada característica de manera individual y contribuya al proceso de clasificación, donde posiblemente contribuya a la precisión global del modelo.

iv. Construcción del dataset

Para la construcción del modelo de clasificación, se empleará la técnica del clasificador de Bayes. Dependiendo de la naturaleza del conjunto de datos, se debe elegir el tipo apropiado de clasificador de Bayes. A continuación, se describen los tipos de clasificadores de Bayes y cuándo es apropiado utilizar cada uno:

Clasificador Gaussiano

El clasificador Gaussiano asume que las características continuas siguen una distribución normal (gaussiana). Es ideal para datos donde las características numéricas se distribuyen de manera normal alrededor de un valor medio. Se usa cuando las características son continuas y se distribuyen normalmente, algunos ejemplos típicos incluyen características como altura, peso, temperatura, etc.

Clasificador MultinomialNB

El clasificador MultinomialNB es adecuado para datos discretos, especialmente cuando las características representan conteos o frecuencias de eventos. Es comúnmente utilizado en problemas de clasificación de texto, donde las características son las frecuencias de palabras en documentos. Se usa **c**uando las características son discretas y representan conteos.

Clasificador BernoulliNB

El clasificador BernoulliNB es apropiado para datos binarios o booleanos, donde las características toman valores de 0 o 1. Al igual que el clasificador MultinomialNB, también se utiliza en problemas de clasificación de texto, pero es más adecuado cuando los datos están representados por la presencia o ausencia de palabras.

Preparación de los datos

Durante el análisis exploratorio de datos, se observó que una gran parte de las variables en el conjunto de datos son categóricas; específicamente, 13 de las 20 variables se clasifican como categóricas. Debido a esta alta proporción, se determinó que el clasificador **MultinomialNB** es el más adecuado para modelar estos datos, ya que es particularmente eficiente para datos categóricos.

No obstante, dado que no todas las variables son categóricas, se realizó una transformación en las características numéricas. Esta transformación consistió en discretizar estas variables creando intervalos de igual tamaño y asignando cada valor a un grupo específico. De este modo, todas las variables del conjunto de datos se convirtieron en categóricas.

Una vez transformadas todas las variables en categóricas, se procedió a aplicar **One-Hot Encoding**. Este proceso convierte las variables categóricas en representaciones binarias, lo que facilita el manejo de los datos para el clasificador. Además, el uso de One-Hot Encoding ayuda a evitar problemas asociados con la suposición de peso implícito que podrían surgir si se usaran valores numéricos continuos para las variables categóricas, lo que podría llevar a interpretaciones erróneas o sesgos en el modelo.

v. Elaboración del Modelo

Una vez que el conjunto de datos fue adecuadamente transformado y adaptado para ser compatible con el modelo **MultinomialNB**, se procedió a dividir los datos en conjuntos de entrenamiento y prueba. La división se realizó utilizando una proporción del 70% para los datos de entrenamiento y el 30% restante para los datos de prueba. Esta partición asegura que el modelo se entrene con una cantidad suficiente de datos y, al mismo tiempo, se evalúe su rendimiento de manera robusta sobre un conjunto independiente.

vi. Verificación de resultados

Al realizar la preparación de las variables independientes y la dependiente, se hicieron las diferentes predicciones de las Y, para obtener los siguientes resultados:

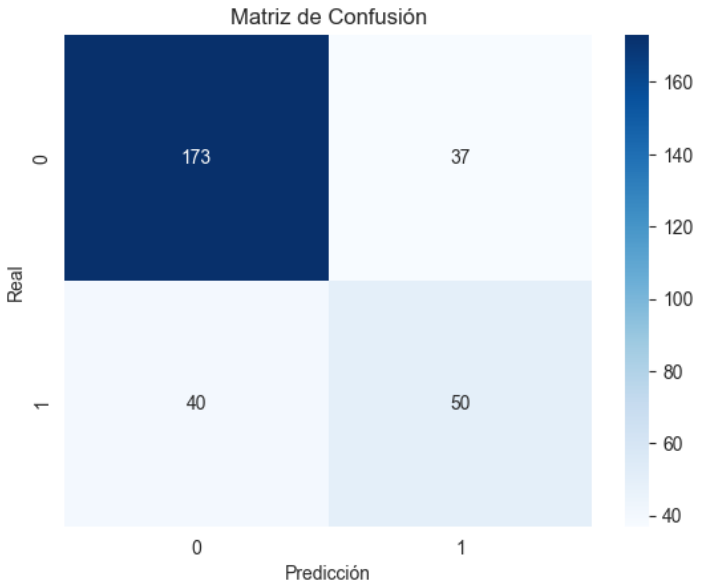

Fig. 6.1. "Matriz de Confusión del clasificador de bayes"

	precision	recall	f1-score	support
1	0.81	0.82	0.82	210
2	0.57	0.56	0.56	90
accuracy			0.74	300
macro avg	0.69	0.69	0.69	300
weighted avg	0.74	0.74	0.74	300

Fig. 6.2. Reporte de clasificación del clasificador de bayes

En términos generales, el modelo alcanzó una exactitud global del 74%, lo que indica que es razonablemente efectivo para predecir el riesgo de crédito, como se ve en la Fig. 6.2.

Para la clase de buen riesgo (1), el modelo mostró una alta precisión de 0.81 y un recall de 0.982. Esto sugiere que el modelo es bastante confiable para identificar correctamente a los buenos riesgos de crédito. El F1-score para esta clase es de 0.82, consolidando la efectividad del modelo en esta área.

No obstante, para la clase de mal riesgo (2), los resultados no son positivos. La precisión fue tan solo de 0.57 y el recall de 0.44, con un F1-score de 0.56. Esto indica que el modelo tiene

dificultades para identificar correctamente a los malos riesgos de crédito, con una tendencia a clasificarlos incorrectamente como buenos riesgos.

La Fig.6.1. matriz de confusión muestra que hubo 173 verdaderos positivos y 50 verdaderos negativos. Por otro lado, se registraron 37 falsos positivos y 40 falsos negativos. Este análisis revela que, aunque el modelo es bastante bueno para identificar a los buenos riesgos, tiende a clasificar incorrectamente a una proporción significativa de malos riesgos como buenos riesgos, lo cual es problemático dado el contexto del problema, en el que es más grave clasificar a un mal riesgo como buen riesgo.

Teniendo en cuenta las limitaciones presentadas por este modelo, se considera utilizar otro modelo de clasificación en busca de conseguir mejores resultados, lo cual se detallará más adelante.

vii. Depuración del modelo

Al realizar la validación cruzada utilizando 5 pliegues, se obtuvieron los siguientes resultados en cuanto a la precisión: 0.775, 0.740, 0.780, 0.720 y 0.780. La precisión media calculada fue de 0.760. Estos resultados son consistentes con las métricas de precisión obtenidas previamente, lo que reafirma que el modelo **MultinomialNB** es razonablemente preciso. Sin embargo, también indica que existe un margen significativo para mejorar el rendimiento del modelo. Esto sugiere que, aunque el modelo tiene una base sólida, se podrían explorar técnicas adicionales o ajustes para optimizar su precisión y capacidad predictiva.

viii. Elaboración Segundo modelo

Para el segundo modelo se buscó realizar un regresor logístico, esto ya que la regresión logística es una técnica de clasificación lineal que puede manejar tanto características continuas como categóricas. El regresor logístico es una buena alternativa que permite contrastar con el modelo de Bayes, pues es posible que los supuestos de independencia de características del modelo de Bayes no se cumplen bien en los datos.

Tomando en consideración la correlación de las variables y el contexto se decidió remover algunas de las variables independientes que tienen menos correlación con la variable objetivo

y además esto se realiza para evitar algún tipo de sesgo por condición que pueda tener el modelo, de manera que las variables independientes seleccionadas son:

- Checking account status
- Duration in month
- Credit History
- Credit Amount
- Employment
- Savings account/bonds
- Property
- Other installment plans

Al realizar la preparación de las variables independientes y la dependiente, se hicieron las diferentes predicciones de las Y, para obtener los siguientes resultados:

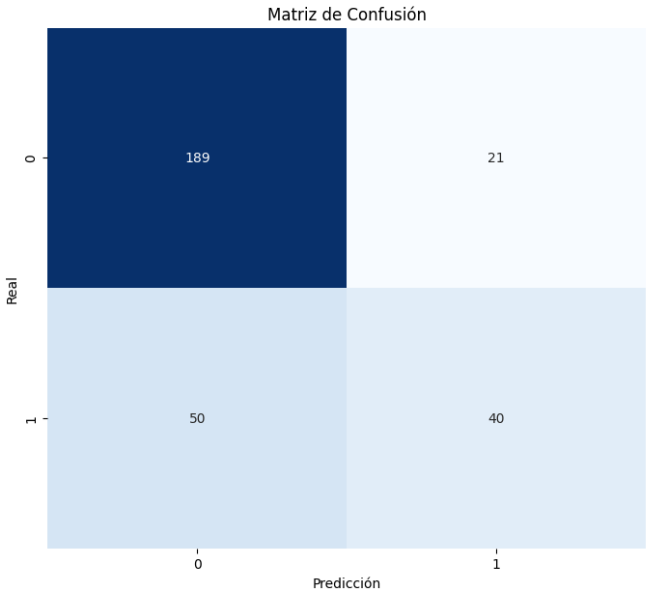

Fig. 7.1. "Matriz de Confusión del regresor logístico"

	precision	recall	f1-score	support
1	0.79	0.90	0.84	210
2	0.66	0.44	0.53	90
accuracy			0.76	300
macro avg	0.72	0.67	0.69	300
weighted avg	0.75	0.76	0.75	300

Fig. 7.2. Reporte de clasificación del regresor logístico

En términos generales, el modelo alcanzó una exactitud global del 76%, lo que indica que es razonablemente efectivo para predecir el riesgo de crédito, como se ve en la Fig. 7.2.

Para la clase de buen riesgo (1), el modelo mostró una alta precisión de 0.79 y un recall de 0.90. Esto sugiere que el modelo es bastante confiable para identificar correctamente a los buenos riesgos de crédito. El F1-score para esta clase es de 0.84, consolidando la efectividad del modelo en esta área.

Sin embargo, para la clase de mal riesgo (2), los resultados no son tan buenos. La precisión fue de 0.66 y el recall de 0.44, con un F1-score de 0.53. Esto indica que el modelo tiene dificultades para identificar correctamente a los malos riesgos de crédito, con una tendencia a clasificarlos incorrectamente como buenos riesgos.

La Fig.7.1. matriz de confusión muestra que hubo 189 verdaderos positivos y 40 verdaderos negativos. Por otro lado, se registraron 21 falsos positivos y 50 falsos negativos. Este análisis revela que, aunque el modelo es bastante bueno para identificar a los buenos riesgos, tiende a clasificar incorrectamente a una proporción significativa de malos riesgos como buenos riesgos, lo cual es problemático dado el contexto del problema, en el que es más grave clasificar a un mal riesgo como buen riesgo, los resultados indican que el modelo de regresión logística podría beneficiarse de ajustes adicionales.

Teniendo esto en cuenta y comparándolo con el modelo de Bayes, el modelo de regresión logística muestra un mejor rendimiento general con una mayor precisión, recall y f1-score para la clase 1 en comparación con el modelo de Naive Bayes, lo que sugiere una mejor capacidad para identificar correctamente la clase 1 (buen riesgo) en este contexto. Sin

embargo, el Naive Bayes presenta un mejor equilibrio en la identificación de la clase 2 (mal riesgo) con una precisión ligeramente mayor y un f1-score más consistente.

ix. Depuración Adicional

Para mejorar el rendimiento del modelo, se podría:

Manejar el desbalance de clases: Dado que hay un desbalance significativo entre las clases, debería tomarse en consideración balancear las clases para que no se clasifique mejor una sobre otra. Métodos como el sobre-muestreo de la clase minoritaria o el sub-muestreo de la clase mayoritaria podrían ayudar a equilibrar el conjunto de datos.

Probar diferentes características: Evaluar la importancia de las características y realizar selección de características puede mejorar el rendimiento del modelo. Además de verificar que conjunto de características podrían ayudar a determinar al modelo la clase a la que pertenece una persona.

Ajuste de Umbrales de Clasificación: Dado que la clasificación incorrecta de la clase 2 como clase 1 es más grave, ajustar el umbral de decisión del modelo para favorecer la detección de la clase 2 podría ayudar a reducir los falsos negativos, mejorando el recall para esta clase, y evitando un error que en este contexto es más delicado.