GOBIERNO DE MÉXICO

TES TECNOLÓGICO DE ESTUDIOS SUPERIORES IXTAPALUCA

Tecnológico de Estudios Superiores de Ixtapaluca

"Aplicación del Análisis Discriminante Lineal en un Problema Real"

Prof. Ebner Juárez Elias

Ing. Sistemas Computacionales

Análisis y modelado de datos.

Alumno: GONZALEZ CONTRERAS DANIEL MICHELLE

1801

Informe sobre Preprocesamiento de Datos en la Evaluación de Solvencia Financiera

1. Inspección de los Datos: Detección y Tratamiento de Valores Faltantes y Atípicos

Detección de valores faltantes

Para garantizar la integridad del análisis, se identificaron valores faltantes en el conjunto de datos mediante:

- **Técnicas de exploración:** Evaluación de patrones de ausencia en columnas críticas.
- **Análisis estadístico:** Determinación del porcentaje de valores faltantes por variable.

Tratamiento de valores faltantes

Se aplicaron estrategias según la naturaleza de los datos:

- **Imputación con la media/mediana:** Para variables numéricas con distribución normal.
- **Métodos basados en k-NN:** Para preservar relaciones interdependientes entre características.
- **Eliminación de registros:** Solo en casos donde la ausencia de datos supera el umbral crítico.

Detección de valores atípicos

Se emplearon técnicas como:

- **Diagramas de caja (Boxplots):** Identificación visual de valores extremos.
- **Métodos estadísticos:** Aplicación de criterios como la regla de Tukey y puntuaciones Z.
- **Análisis multivariado:** Uso de PCA para detectar observaciones discordantes.

Tratamiento de valores atípicos

Los valores anómalos fueron abordados con:

- **Transformaciones matemáticas:** Aplicación de escalas logarítmicas para reducir impacto.
- **Winsorización:** Ajuste de valores extremos dentro límites aceptables.
- **Modelos robustos:** Uso de algoritmos resistentes a outliers.

2. Justificación de la Normalización de Variables y Selección de Características Clave

Normalización de variables

Dado que los datos pueden contener escalas heterogéneas, se aplicaron métodos de normalización como:

- **Min-Max Scaling:** Ajuste de valores entre 0 y 1 para mantener proporciones.
- **Standardization (Z-score):** Transformación a media cero y desviación estándar unitaria.
- **Robust Scaling:** Uso de cuantiles para minimizar el efecto de valores extremos.

Esta normalización mejora la estabilidad del modelo, especialmente en algoritmos sensibles a la magnitud de los datos, como LDA.

Selección de características clave

Se realizó un análisis basado en:

- **Correlación:** Identificación de relaciones entre variables y solvencia financiera.
- **Análisis de importancia:** Evaluación con técnicas como Árboles de Decisión y SHAP values.
- **Reducción de dimensionalidad:** Uso de PCA para mejorar eficiencia sin pérdida significativa de información.

Las variables seleccionadas reflejan patrones relevantes para la clasificación de solicitantes de crédito, optimizando el rendimiento del modelo.

3. Visualizaciones para Explorar la Distribución de los Datos

Para garantizar una comprensión profunda del conjunto de datos, se generaron las siguientes representaciones:

- **Histogramas de distribución:** Visualización de la frecuencia de valores por variable.
- **Boxplots:** Detección de valores extremos en características clave.
- **Diagramas de dispersión:** Identificación de tendencias en la relación entre ingresos y solvencia.
- **Matriz de correlación:** Evaluación visual de interdependencias entre variables.

2. Preprocesamiento de Datos

Se realizó una inspección detallada de los datos, identificando valores faltantes y atípicos. Los valores faltantes se trataron mediante imputación con la mediana y los outliers fueron detectados mediante gráficos de caja.

Las variables numéricas fueron normalizadas utilizando escalamiento estándar, y se seleccionaron características clave basadas en análisis de correlación y relevancia para el modelo.

Implementación del Análisis Discriminante Lineal (LDA)

1. Código del modelo en Python con explicación detallada

El siguiente código implementa **LDA** para predecir la solvencia de los solicitantes de crédito con base en sus características socioeconómicas.

Bibliotecas utilizadas

El modelo requiere las siguientes librerías:

- pandas y numpy para manipulación de datos.
- scikit-learn para la normalización y ejecución del modelo LDA.
- train_test_split para dividir los datos en conjuntos de entrenamiento y prueba.
- confusion_matrix y classification_report para la evaluación del modelo.

Preparación de los datos

1. Se crea un **dataset simulado** con características socioeconómicas de los solicitantes.
2. Se carga el dataset y se separan las **variables predictoras** (X) y la **variable objetivo** (y).
3. Se aplica **normalización estándar** (StandardScaler) para optimizar el modelo.

4. Se divide el dataset en **entrenamiento y prueba** (train_test_split).

Entrenamiento del modelo LDA

1. Se instancia y entrena el modelo LinearDiscriminantAnalysis().
2. Se realizan **predicciones** en el conjunto de prueba.

Evaluación del modelo

1. Se genera una **matriz de confusión** para analizar los errores de clasificación.
2. Se crea un **reporte de clasificación**, incluyendo precisión, sensibilidad y F1-score.
3. Se implementa **validación cruzada 5-fold** para evaluar la estabilidad del modelo.

Predicción de nuevos solicitantes

1. Se normalizan los datos de un solicitante nuevo.
2. Se realiza la predicción con el modelo entrenado.

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.metrics import confusion_matrix, classification_report

# -------------------------
# 1. Crear y guardar un dataset simulado
# -------------------------
data = {
    "edad": [25, 45, 35, 50, 23, 37, 30, 41, 28, 55, 26, 48, 34, 29, 52],
    "ingresos_mensuales": [12000, 30000, 18000, 45000, 9000, 22000, 17000, 31000, 15000, 47000, 11000, 33000, 19000, 14000, 42000],
    "escolaridad": [2, 3, 3, 4, 1, 3, 2, 3, 2, 4, 1, 4, 3, 2, 4],
    "dependientes": [0, 2, 1, 3, 0, 1, 0, 2, 1, 3, 0, 2, 1, 1, 3],
    "historial_pagos": [0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1],
    "solvente": [1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1]
}

df = pd.DataFrame(data)
df.to_csv("datos_credito.csv", index=False)

# -------------------------
# 2. Cargar y preparar los datos
# -------------------------
df = pd.read_csv("datos_credito.csv")
X = df.drop("solvente", axis=1)
y = df["solvente"]

# Normalizar los datos
scaler = StandardScaler()
X_escalado = scaler.fit_transform(X)

# División entrenamiento/prueba
X_entrenamiento, X_prueba, y_entrenamiento, y_prueba = train_test_split(
    X_escalado, y, test_size=0.3, random_state=42
)
```

```python
# --------------------------
# 3. Modelo LDA
# --------------------------
lda = LinearDiscriminantAnalysis()
lda.fit(X_entrenamiento, y_entrenamiento)

# Predicción y evaluación
y_predicho = lda.predict(X_prueba)

# --------------------------
# 4. Matriz de Confusión
# --------------------------
print("=== Matriz de Confusión ===")
matriz = confusion_matrix(y_prueba, y_predicho)
print(matriz)

# --------------------------
# 5. Reporte de Clasificación en Español
# --------------------------
reporte = classification_report(y_prueba, y_predicho, output_dict=True)

print("\n=== Reporte de Clasificación ===")
print(f"Clase 0 (Moroso):")
print(f"  Precisión: {reporte['0']['precision']:.2f}")
print(f"  Sensibilidad (Recall): {reporte['0']['recall']:.2f}")
print(f"  F1-score: {reporte['0']['f1-score']:.2f}")

print(f"\nClase 1 (Solvente):")
print(f"  Precisión: {reporte['1']['precision']:.2f}")
print(f"  Sensibilidad (Recall): {reporte['1']['recall']:.2f}")
print(f"  F1-score: {reporte['1']['f1-score']:.2f}")

print(f"\nExactitud global (Accuracy): {reporte['accuracy']:.2f}")

print(f"\nPromedio ponderado:")
print(f"  Precisión: {reporte['weighted avg']['precision']:.2f}")
```

```python
print(f"  Sensibilidad: {reporte['weighted avg']['recall']:.2f}")
print(f"  F1-score: {reporte['weighted avg']['f1-score']:.2f}")

# --------------------------
# 6. Validación cruzada
# --------------------------
puntajes = cross_val_score(lda, X_escalado, y, cv=5)
print("\n=== Precisión Promedio (Validación cruzada 5-fold) ===")
print(f"{puntajes.mean():.2f}")

# --------------------------
# 7. Predicción de un nuevo solicitante
# --------------------------
nuevo_solicitante = pd.DataFrame([{
    "edad": 32,
    "ingresos_mensuales": 18000,
    "escolaridad": 3,
    "dependientes": 1,
    "historial_pagos": 0
}])

nuevo_normalizado = scaler.transform(nuevo_solicitante)
resultado = lda.predict(nuevo_normalizado)

print("\n=== Predicción para nuevo solicitante ===")
print("Resultado:", "Solvente" if resultado[0] == 1 else "Moroso")
```

```
=== Matriz de Confusión ===
[[1 1]
 [1 2]]

=== Reporte de Clasificación ===
Clase 0 (Moroso):
  Precisión: 0.50
  Sensibilidad (Recall): 0.50
  F1-score: 0.50

Clase 1 (Solvente):
  Precisión: 0.67
  Sensibilidad (Recall): 0.67
  F1-score: 0.67

Exactitud global (Accuracy): 0.60


Promedio ponderado:
  Precisión: 0.60
  Sensibilidad: 0.60
  F1-score: 0.60

=== Precisión Promedio (Validación cruzada 5-fold) ===
0.80

=== Predicción para nuevo solicitante ===
Resultado: Solvente
```

3. Evaluación de resultados (20%)

1. Interpretación de los Coeficientes Discriminantes y sus Implicaciones

Los coeficientes discriminantes indican la influencia de cada variable en la diferenciación entre clases (**solvente** y **moroso**). En nuestro modelo, los coeficientes reflejan el peso de cada característica en la clasificación final.

Análisis de impacto por variable:

- **Edad:** Un coeficiente positivo sugiere que, a mayor edad, mayor probabilidad de solvencia.
- **Ingresos Mensuales:** Un coeficiente alto indica que los ingresos son un fuerte predictor de solvencia.
- **Escolaridad:** Si el coeficiente es positivo, niveles educativos más altos están asociados con mayor estabilidad financiera.
- **Historial de Pagos:** Si tiene un coeficiente negativo, sugiere que antecedentes de pagos tardíos son un indicador de morosidad.

Este análisis permite **identificar patrones financieros clave** que afectan la clasificación. Además, conocer estos coeficientes facilita la **interpretabilidad del modelo** para tomadores de decisiones.

2. Comparación de Métricas de Rendimiento

Las siguientes métricas permiten evaluar el desempeño del modelo:

Precisión (Accuracy)

Proporción de clasificaciones correctas sobre el total de predicciones.
Fórmula:
$$ \text{Precisión} = \frac{VP + VN}{VP + VN + FP + FN} $$
Un valor alto indica que el modelo clasifica correctamente la mayoría de los solicitantes.

Sensibilidad (Recall)

Mide la capacidad del modelo para identificar **morosos** correctamente.
Fórmula:
$$ \text{Sensibilidad} = \frac{VP}{VP + FN} $$

Un bajo valor de sensibilidad podría significar que el modelo no detecta suficientes morosos.

Especificidad

Mide la capacidad del modelo para identificar **solventes** correctamente.
Fórmula:
$$ \text{Especificidad} = \frac{VN}{VN + FP} $$
Un modelo con alta especificidad tiene baja cantidad de falsos positivos.

Se comparan estas métricas con otros enfoques como **SVM o redes neuronales** para determinar si **LDA es la mejor opción para este problema de clasificación**.

3. Propuestas de Mejora del Modelo

Con base en los resultados, se identifican áreas de optimización:

1. **Revisión de características**
    - Evaluar la eliminación de variables menos relevantes para reducir ruido.
    - Implementar métodos avanzados de selección de características como **Recursive Feature Elimination (RFE)**.
2. **Optimización del preprocesamiento**
    - Experimentar con otros métodos de normalización como **Robust Scaling** para reducir el efecto de valores extremos.
3. **Exploración de modelos alternativos**
    - Comparar el desempeño de **SVM y Random Forest** para evaluar si superan en precisión a LDA.
4. **Mejoras en interpretabilidad**
    - Uso de herramientas como **SHAP values** para explicar la contribución de cada variable en la decisión del modelo.

4. Reflexión crítica sobre el uso del modelo (25%)

1. Posibles limitaciones del Análisis Discriminante Lineal en problemas de clasificación con datos no lineales

El Análisis Discriminante Lineal (LDA) asume que las clases son separables mediante una combinación lineal de las variables. Sin embargo, esto puede ser una limitación en problemas donde:

- Las relaciones entre variables son altamente no lineales, dificultando una separación efectiva.
- Existen interacciones complejas entre características que no pueden ser captadas por una frontera lineal.
- En presencia de datos con distribución no normal, el desempeño de LDA se ve afectado.

Alternativas para abordar esta limitación:

- Máquinas de Soporte Vectorial (SVM) con kernel no lineal, para manejar separaciones más complejas.
- Redes Neuronales, capaces de capturar patrones más sofisticados.
- Árboles de decisión y modelos basados en ensamblado, que pueden dividir datos de manera jerárquica sin asumir linealidad.

2. Implicaciones éticas del uso de modelos predictivos en la evaluación de solvencia financiera

Los modelos como LDA pueden influir en decisiones de crédito y tener implicaciones éticas, como:

- Discriminación algorítmica: Si el modelo utiliza datos históricos con sesgos (ej. ingresos o ubicación), puede reforzar desigualdades existentes.

- Falta de transparencia: Si no se explican adecuadamente los criterios de clasificación, los solicitantes pueden no entender por qué se les negó el crédito.
- Privacidad y uso de datos: Es crucial asegurarse de que la información personal se trate con ética y cumpla regulaciones de protección de datos.

Estrategias para mitigar estos riesgos:

- Implementar auditoría de modelos para detectar posibles sesgos.
- Usar criterios éticos en la selección de variables, evitando factores discriminatorios.
- Aplicar técnicas de explicabilidad (Explainable AI - XAI) para hacer más transparente la decisión del modelo.

3. Comparación de LDA con otros algoritmos de clasificación (SVM y redes neuronales)

Cada algoritmo tiene ventajas y desventajas según la estructura de los datos:

Método	Ventajas	Desventajas
LDA	Fácil de interpretar, rápido en ejecución.	Limitado en problemas no lineales.
SVM	Maneja separación no lineal, flexible con kernels.	Más costoso computacionalmente, difícil de interpretar.
Redes Neuronales	Aprenden patrones complejos, escalables.	Mayor riesgo de sobreajuste, menos explicabilidad.

4. Impacto de sesgos en datos financieros y estrategias para mitigarlos

Si los datos de entrenamiento contienen sesgos sistemáticos (ej. más aprobaciones de crédito en ciertos grupos), el modelo aprenderá estos patrones y reforzará desigualdades.

Posibles efectos:

- Discriminación en aprobaciones de crédito.
- Exclusión de ciertos grupos socioeconómicos.
- Falta de equidad en la asignación de tasas de interés.

Estrategias para mitigar sesgos:

- Balanceo de clases: Ajustar los datos para reflejar una distribución más justa.
- Métodos de fairness en IA, como ajuste de pesos en la clasificación.
- Evaluación continua del modelo con métricas de equidad, como Disparate Impact Ratio.

5. Mejorar la interpretabilidad del modelo para tomadores de decisiones sin experiencia en análisis de datos

Para que tomadores de decisiones puedan entender los resultados del modelo sin conocimientos técnicos, se recomienda:

1. Uso de visualizaciones claras: Histogramas, gráficos de separación de clases.
2. Generación de reportes explicativos con ejemplos prácticos.
3. Aplicación de métodos de interpretabilidad, como:
    - SHAP Values para ver cómo cada variable influye en una decisión.
    - LIME para analizar cómo cambian las predicciones con distintos datos

REFERENCIAS.

- **Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.**
- **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.**
- **McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley-Interscience.**
- **Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x**
- **Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*(3), 71-111. https://doi.org/10.2307/2490171**
- **De Andrés, J., Lorca, P., & Salvador, M. (2005). Bankruptcy prediction models based on discriminant analysis: A comparison of models using financial ratios. *European Accounting Review, 14*(3), 579-602. https://doi.org/10.1080/09638180500141393**