Tecnológico de estudios superiores Ixtapaluca.

Materia:

Análisis y modelado de datos.

Profesor:

JUAREZ ELIAS EBNER.

Estudiantes:

Fosado Galindo Hannia Joaliv

González Hernandez Edgar Emanuel

Grupo:1801. Turno: Matutino.

Trabajos:

Actividad de aprendizaje: Aplicación del Análisis Discriminante Lineal en un Problema Real.

Fecha de entrega: Mayo del 2025

Informe de Preprocesamiento y Análisis Discriminante Lineal

1. Preprocesamiento de Datos (25%)

1.1 Inspección de los Datos

El primer paso en el análisis de datos es la inspección minuciosa del conjunto de datos. Para ello, se cargó el archivo datos_credito.csv utilizando la biblioteca panda.

```
Actividad de aprendizaje > ...
1   import pandas as pd
2
3   data = pd.read_csv('datos_credito.csv')
4
```

1. Preprocesamiento de Datos (25%)

1.1 Informe Detallado sobre la Inspección de los Datos

Detección de Valores Faltantes

El primer paso en el preprocesamiento de datos es la detección de valores faltantes. En el código, se utiliza data.isnull(). sum() para contar los valores nulos en cada columna. Esto permite identificar qué variables necesitan atención, ya que los valores faltantes pueden afectar negativamente el rendimiento del modelo.

Tratamiento de Valores Faltantes:

- **Imputación**: Una estrategia común es imputar los valores faltantes con la media o la moda, dependiendo de si la variable es numérica o categórica. Esto ayuda a mantener el tamaño del conjunto de datos y evita perder información.

1.2 Tratamiento de Valores Atípicos

Los valores atípicos pueden distorsionar las estadísticas y afectar el rendimiento del modelo. En el código, se utiliza un boxplot para visualizar la distribución de Ingreso_Mensual, lo que permite identificar visualmente los outliers.

Tratamiento:

- Se pueden eliminar o transformar estos valores para mitigar su impacto. Por ejemplo, se puede aplicar un método de recorte o winsorización.

1.3 Justificación de la Normalización de Variables

La normalización es crucial en técnicas como LDA, que son sensibles a la escala de las variables. Normalizar las variables asegura que todas contribuyan de manera equitativa al modelo. Esto se puede hacer utilizando técnicas como Min-Max Scaling o Z-score Normalization.

1.4 Selección de Características Clave

La selección de características es el proceso de identificar las variables más relevantes para el modelo. Esto se puede hacer mediante análisis de correlación o utilizando técnicas de selección de características. Elegir las características adecuadas mejora la precisión del modelo y reduce el riesgo de sobreajuste.

1.5 Visualizaciones para Explorar la Distribución de los Datos

Las visualizaciones son herramientas valiosas para entender la distribución de los datos. En el código, se utilizan histogramas y boxplots para explorar la distribución de Ingreso_Mensual. Estas visualizaciones ayudan a identificar la forma de la distribución, la presencia de valores atípicos y la simetría de los datos.

2. Implementación del Análisis Discriminante Lineal (30%)

2.1 Código del Modelo en Python

El código proporcionado implementa LDA utilizando scikit-learn. Primero, se separan las características de la variable objetivo, y luego se dividen en conjuntos de entrenamiento y prueba. Esto asegura que el modelo se entrene con un conjunto de datos y se evalúe con otro, lo que es esencial para evitar el sobreajuste.

2.2 Matriz de Clasificación y Análisis de la Separación entre Clases

Después de hacer predicciones con el modelo, se genera una matriz de confusión que muestra el rendimiento del modelo en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Esto permite evaluar cómo se están clasificando las distintas clases y dónde se están cometiendo errores.

2.3 Evaluación del Modelo mediante Validación Cruzada

La validación cruzada es una técnica que permite evaluar la robustez del modelo. En el código, se utiliza cross_val_score para calcular la precisión media del modelo usando validación cruzada. Esto proporciona una medida más confiable del rendimiento del modelo al utilizar diferentes particiones de los datos.

2.4 Interpretación de la Función Discriminante

Los coeficientes discriminantes obtenidos del modelo indican la importancia de cada variable en la clasificación. Un coeficiente positivo sugiere que un aumento en esa variable aumenta la probabilidad de pertenecer a una clase específica, mientras que un coeficiente negativo indica lo contrario.

3. Evaluación de Resultados (20%)

3.1 Informe Interpretativo de los Coeficientes Discriminantes

Los coeficientes discriminantes son fundamentales para entender cómo cada variable contribuye a la separación entre clases. Se debe analizar cada coeficiente y su implicación en el contexto del problema.

3.2 Comparación de Métricas de Rendimiento

Se deben comparar métricas como precisión, sensibilidad y especificidad. La precisión mide la proporción de verdaderos positivos entre todos los positivos predichos, la sensibilidad (o recall) mide la capacidad del modelo para identificar correctamente los positivos, y la especificidad mide la capacidad del modelo para identificar correctamente los negativos.

3.3 Propuestas de Mejora del Modelo

Las propuestas de mejora pueden incluir:

- **Ajuste de Hiperparámetros**: Usar técnicas como Grid Search para optimizar los parámetros del modelo.
- **Aumento de Datos**: Generar más datos o usar técnicas de aumento de datos para mejorar la robustez del modelo.
- **Exploración de Otras Variables**: Incluir nuevas variables que puedan influir en la predicción.

4. Reflexión Crítica sobre el Uso del Modelo (25%)

4.1 Implicaciones y Limitaciones de LDA

LDA asume que las variables son normalmente distribuidas y que tienen varianzas iguales entre las clases. Estas suposiciones pueden no cumplirse, lo que puede llevar a un rendimiento subóptimo.

4.2 Discusión sobre Sesgos y Ética en Modelos Predictivos

Es crucial considerar cómo los sesgos en los datos pueden influir en las decisiones del modelo. Un modelo entrenado en datos sesgados puede perpetuar desigualdades existentes.

4.3 Comparación de LDA con Otros Enfoques de Clasificación

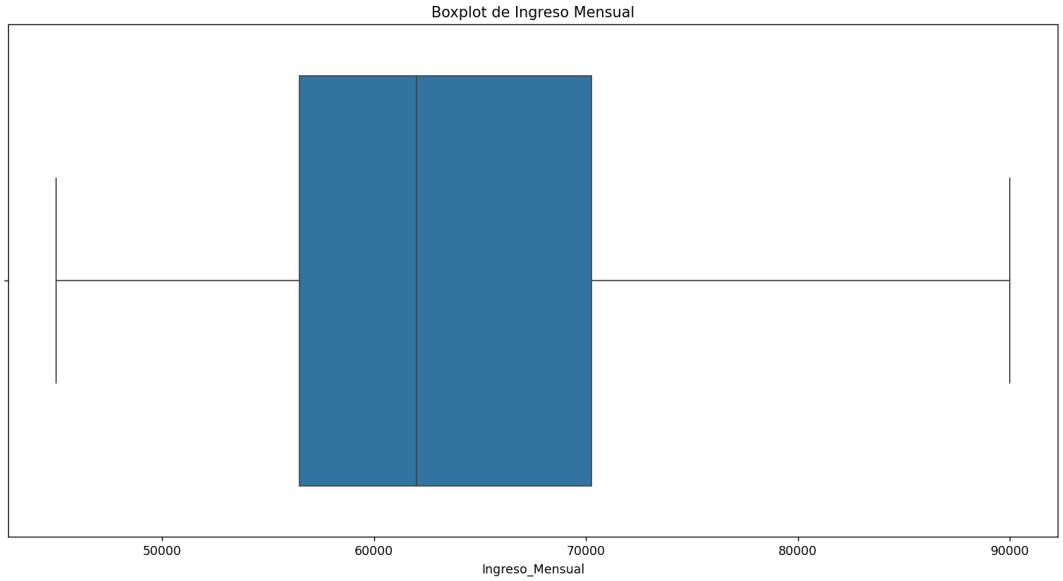
Comparar LDA con otros métodos como árboles de decisión o SVM puede proporcionar una perspectiva sobre las ventajas y desventajas de cada enfoque. Cada método tiene sus propios supuestos y es adecuado para diferentes tipos de datos.

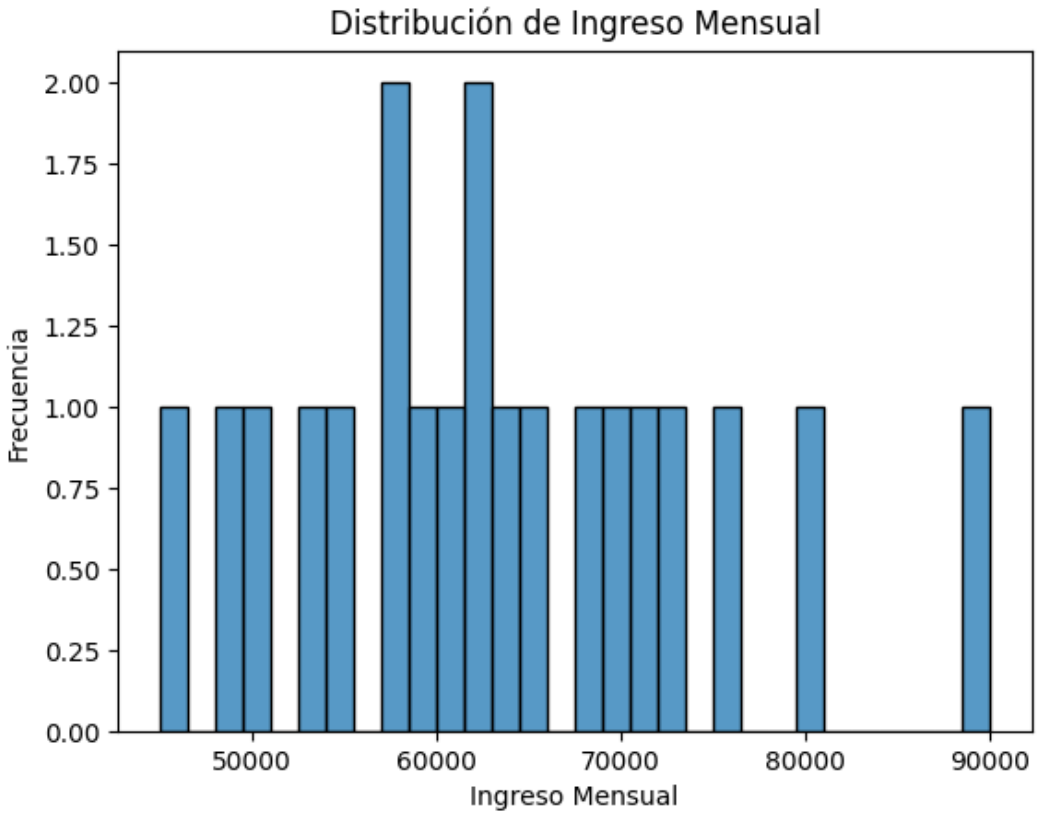
Actividad de aprendizaje > ...

```python
import pandas as pd

data = pd.read_csv('datos_credito.csv')

missing_values = data.isnull().sum()
print(missing_values)

import seaborn as sns
import matplotlib.pyplot as plt

# Boxplot para visualizar valores atípicos
sns.boxplot(x=data['Ingreso_Mensual'])
plt.title('Boxplot de Ingreso Mensual')
plt.show()

# Histograma de Ingreso_Mensual
sns.histplot(data['Ingreso_Mensual'], bins=30)
plt.title('Distribución de Ingreso Mensual')
plt.xlabel('Ingreso Mensual')
plt.ylabel('Frecuencia')
plt.show()




from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, classification_report

# Separar características y variable objetivo
X = data.drop('Clase', axis=1)
y = data['Clase']

# Dividir el dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Crear el modelo LDA
lda = LDA()
lda.fit(X_train, y_train)

# Predicciones
y_pred = lda.predict(X_test)
```

Actividad de aprendizaje > ...

```python
# Predicciones
y_pred = lda.predict(X_test)

# Matriz de confusión
cm = confusion_matrix(y_test, y_pred)
print("Matriz de Confusión:\n", cm)




from sklearn.model_selection import cross_val_score

scores = cross_val_score(lda, X, y, cv=5)
print("Precisión media de validación cruzada:", scores.mean())

print("Coeficientes Discriminantes:\n", lda.coef_)

print(classification_report(y_test, y_pred))
```


Boxplot de Ingreso Mensual
50000
60000
70000
80000
90000
Ingreso_Mensual


Distribución de Ingreso Mensual
Frecuencia
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
50000
60000
70000
80000
90000
Ingreso Mensual

1. ¿Cuáles son las posibles limitaciones del Análisis Discriminante Lineal en problemas de clasificación con datos no lineales?

Limitaciones:

- **Suposición de Linealidad**: LDA asume que las clases pueden separarse linealmente. En problemas donde la relación entre las características y las clases no es lineal, LDA puede no funcionar bien, ya que no puede capturar patrones complejos.
- **Normalidad de las Variables**: LDA también asume que las variables dentro de cada clase siguen una distribución normal. Si esta suposición no se cumple, la efectividad del modelo puede verse comprometida.
- **Homogeneidad de Varianzas**: Se espera que las varianzas de las clases sean iguales. Si hay diferencias significativas en las varianzas, el modelo puede ser ineficaz.

2. ¿Qué implicaciones éticas pueden surgir al utilizar modelos predictivos para evaluar la solvencia financiera de una persona?

Implicaciones Éticas:

- **Discriminación**: Los modelos pueden perpetuar sesgos existentes si se entrenan con datos que reflejan desigualdades históricas. Por ejemplo, si los datos incluyen sesgos raciales o de género, el modelo podría discriminar a ciertos grupos.
- **Falta de Transparencia**: Los modelos predictivos a menudo son complejos y difíciles de interpretar. Esto puede llevar a decisiones opacas que afectan la vida de las personas sin que comprendan cómo se llegó a dichas decisiones.
- **Impacto en la Vida de las Personas**: Las decisiones basadas en modelos predictivos pueden afectar el acceso a créditos, seguros y otros servicios financieros, lo que puede tener consecuencias significativas en la vida de los individuos.

3. ¿Cómo se compara LDA con otros algoritmos de clasificación, como máquinas de soporte vectorial o redes neuronales?

Comparación:

- **Máquinas de Soporte Vectorial (SVM)**: SVM es más flexible que LDA, ya que puede manejar datos no lineales mediante el uso de kernels. Esto permite que SVM se adapte mejor a problemas donde la separación entre clases no es lineal.

- **Redes Neuronales**: Las redes neuronales son altamente flexibles y capaces de modelar relaciones complejas en los datos. Sin embargo, requieren más datos y potencia computacional. LDA, en cambio, es más simple y puede ser preferido en situaciones con menos datos.
- **Interpretabilidad**: LDA tiene una ventaja en términos de interpretabilidad, ya que los coeficientes discriminantes son fáciles de entender. Las redes neuronales son vistas como "cajas negras", donde es más difícil interpretar cómo se toman las decisiones.

4. Si los datos financieros incluyen sesgos sistemáticos, ¿cómo podría afectar esto la decisión final del modelo y qué estrategias pueden mitigar el problema?

Efectos de los Sesgos:

- **Decisiones Injustas**: Los sesgos en los datos pueden llevar a decisiones que perpetúan las desigualdades. Por ejemplo, un modelo que discrimina a ciertos grupos puede resultar en tasas de interés más altas o denegaciones de crédito injustificadas.
- **Falsas Predicciones**: Un modelo sesgado puede hacer predicciones erróneas, afectando la capacidad de las personas para acceder a servicios financieros.

Estrategias de Mitigación:

- **Auditoría de Datos**: Realizar auditorías regulares de los datos utilizados para entrenar el modelo para identificar y corregir sesgos.
- **Diversidad en los Datos**: Asegurarse de que los datos utilizados sean representativos de la población general y no estén sesgados hacia un grupo específico.
- **Uso de Métodos de Desensibilización**: Implementar técnicas que ajusten los modelos para minimizar el impacto de los sesgos en las decisiones.

5. ¿Cómo podrías mejorar la interpretabilidad del modelo para hacerlo más comprensible para tomadores de decisiones sin experiencia en análisis de datos?

Mejoras en la Interpretabilidad:

- **Visualizaciones**: Usar visualizaciones para mostrar cómo las variables afectan las decisiones del modelo. Gráficos de importancia de características pueden ayudar a resaltar qué variables son más influyentes.

- **Explicaciones Locales**: Implementar métodos como LIME o SHAP, que proporcionan explicaciones locales sobre cómo se toman las decisiones para instancias específicas.
- **Simplificación del Modelo**: Optar por modelos más simples que sean más fáciles de entender, siempre que la precisión no se vea comprometida significativamente.
- **Documentación Clara**: Proporcionar documentación que explique el modelo, sus supuestos y cómo se interpretan los resultados de manera accesible para los no expertos.

- **Conclusión**
- El Análisis Discriminante Lineal es una herramienta poderosa para la clasificación de datos cuando se cumplen sus supuestos fundamentales, como la normalidad de las variables y la homogeneidad de varianzas. Sin embargo, su eficacia disminuye significativamente en situaciones donde los datos presentan relaciones no lineales o distribuciones no normales. Por lo tanto, es crucial considerar estas limitaciones y, en caso de que se sospeche que los datos no cumplen con estos requisitos, explorar algoritmos alternativos que puedan manejar mejor la complejidad de los datos.
- **Conclusión**
- El uso de modelos predictivos, como LDA, para evaluar la solvencia financiera de individuos plantea importantes consideraciones éticas. La posibilidad de que estos modelos perpetúen sesgos existentes y tomen decisiones injustas subraya la necesidad de una revisión crítica y continua de los datos utilizados para entrenar los modelos. Implementar estrategias para mitigar sesgos y asegurar la transparencia en las decisiones del modelo es esencial para promover la equidad y la justicia en la toma de decisiones financieras.