Programa de Maestría en Ciencia de los datos y analítica

Mejoramiento de la confiabilidad en equipos de potencia a través del análisis predictivo basado en datos históricos

Proyecto integrador

Semestre III

Ciencia de los datos y analítica

PRESENTA:
Cristian Castro Arias

Tutores Principales:
Carlos Alzate

Colombia,Medellin. (Noviembre) 2023

Contents

Abstract

Este trabajo investiga la aplicación de métodos estadísticos robustos y no paramétricos para la clasificación de maquinaria industrial basada en datos de rendimiento en un contexto de desequilibrio de clases. Frente a la limitación de los enfoques tradicionales que asumen distribuciones normales y están sujetos a influencia por outliers, se emplean técnicas que no se basan en supuestos de distribución subyacente. Utilizando el 'Performance Measurement (pfm)' como principal indicador, se desarrolla un modelo de clasificación que maneja eficazmente la prevalencia de clases mayoritarias y la escasez de datos para clases minoritarias. Se adoptan técnicas de sobremuestreo como SMOTE para equilibrar el conjunto de datos y se implementa un enfoque de machine learning utilizando Random Forest, optimizado a través de búsqueda en cuadrícula y validación cruzada. La evaluación del modelo se realiza mediante métricas que incluyen la precisión F1 ponderada y el índice de Jaccard, revelando un avance significativo sobre las métricas estándar y resaltando la importancia de un enfoque estadístico robusto en el análisis de datos industriales. Los resultados demuestran la eficacia de estos métodos en la mejora de la predicción y la relevancia de su aplicación práctica en el mantenimiento predictivo y la optimización operativa.

1 Introducción

En el campo de la ciencia de datos, la clasificación precisa de entidades basada en sus características inherentes es fundamental para la toma de decisiones informadas y la generación de conocimientos accionables. Este desafío se magnifica en presencia de conjuntos de datos desequilibrados, donde las clases predominantes eclipsan a las minoritarias, conduciendo a modelos predictivos sesgados. La tradicional dependencia de estadísticas paramétricas y modelos basados en suposiciones de normalidad a menudo falla ante la naturaleza compleja y heterogénea de los datos reales. Por consiguiente, surge la necesidad de métodos que sean robustos y no paramétricos, capaces de proporcionar resultados confiables incluso cuando las condiciones ideales de los modelos estadísticos no se cumplen.

La estadística robusta y no paramétrica se presenta como una solución a este dilema, permitiendo el análisis y modelado de datos sin depender de suposiciones previas sobre su distribución. Este enfoque es de particular importancia en el dominio de la ingeniería eléctrica y la fabricación de equipos, donde la clasificación de máquinas según su rendimiento – medido por indicadores como el 'Performance Measurement (pfm)' – es crucial para el mantenimiento predictivo y la optimización de procesos.

El presente estudio aborda la tarea de clasificar maquinaria industrial en categorías predeterminadas basadas en mediciones de rendimiento, utilizando técnicas estadísticas robustas y no paramétricas. Se enfrenta al reto de los datos desequilibrados a través de la implementación de métodos de sobremuestreo y modelos de aprendizaje automático avanzados, con el objetivo de alcanzar un equilibrio entre la precisión y la generalización. Mediante la adopción de métricas como el índice de Jaccard y la puntuación F1, este proyecto aspira a desarrollar un modelo que no solo sea preciso en la mayoría de las clases, sino también efectivo en la identificación de las clases minoritarias, proporcionando así una solución equitativa y robusta para la clasificación de equipos.

2 Estado del Arte

El uso de la estadística robusta y no paramétrica ha ganado terreno en diversas áreas de la ingeniería y la ciencia de datos. Estos métodos se han mostrado particularmente útiles en situaciones donde los modelos paramétricos tradicionales fallan debido a la presencia de outliers o distribuciones no normales de los datos. En el campo de la ingeniería eléctrica, estudios recientes han empezado a explorar el uso del PCA robusto, las pruebas de rangos de Wilcoxon y Mann-Whitney, y la regresión robusta para mejorar la precisión y confiabilidad en la predicción del estado de los equipos. Estos enfoques han demostrado ser prometedores, particularmente en la gestión de activos y en el mantenimiento predictivo, donde la calidad y naturaleza de los datos pueden variar considerablemente.

3 Marco Teórico

3.1 Estadística Robusta y No Paramétrica

La estadística robusta se ocupa del desarrollo de métodos estadísticos que proporcionan resultados fiables incluso cuando las suposiciones subyacentes de los modelos estándar se violan. Estos métodos son diseñados para ser insensibles a pequeñas desviaciones de las suposiciones teóricas, como la normalidad de los datos o la presencia de outliers. Dentro de este enfoque, se prioriza la robustez de las estimaciones y pruebas, buscando que los resultados sean válidos y consistentes en una amplia gama de situaciones.

Por otro lado, la estadística no paramétrica no hace suposiciones específicas sobre la forma de la distribución de los datos. Estos métodos son flexibles y aplicables a datos que no cumplen con las suposiciones de los modelos paramétricos, como la homogeneidad de varianzas o la linealidad. La estadística no paramétrica incluye pruebas de hipótesis, estimaciones de densidad, y métodos de clasificación que se basan únicamente en la información proporcionada por los datos, sin recurrir a un modelo de distribución previo.

3.2 Modelos de Clasificación en Machine Learning

Los modelos de clasificación en machine learning buscan categorizar entidades en grupos predefinidos basándose en características observadas. Estos modelos pueden ser supervisados, donde se utilizan datos etiquetados para entrenar al modelo, y no supervisados, donde las estructuras se identifican sin etiquetas previas. En particular, los modelos supervisados como Random Forest o Gradient Boosting han demostrado ser efectivos para la clasificación de datos desequilibrados, especialmente cuando se combinan con técnicas de reequilibrio como SMOTE (Synthetic Minority Over-sampling Technique).

3.3 Métricas de Evaluación de Modelos

La precisión de un modelo de clasificación se evalúa típicamente a través de métricas como la precisión, el recall, la puntuación F1 y el índice de Jaccard. La precisión F1 es una medida armónica de la precisión y el recall, y es particularmente útil cuando las clases están desequilibradas. El índice de Jaccard, o la puntuación de Jaccard, mide la similitud entre las etiquetas verdaderas y las predichas, y es robusto en situaciones donde las clases minoritarias pueden ser ignoradas por otras métricas como la precisión global.

3.4 Desequilibrio de Clases

El desequilibrio de clases se refiere a situaciones en las cuales ciertas clases están subrepresentadas en el conjunto de datos, lo cual puede llevar a un sesgo en el modelo hacia las clases mayoritarias. Este fenómeno es común en muchos ámbitos prácticos, incluyendo la medicina, la detección de fraudes y la fabricación industrial. Abordar el desequilibrio de clases es crítico para desarrollar modelos de clasificación justos y equitativos que funcionen bien en todas las categorías de interés.

4 métodos y enfoques

4.1 Función de Distribución Empírica (ECDF)

La **Función de Distribución Empírica (ECDF)** es una herramienta no paramétrica fundamental en la estadística que proporciona la proporción acumulativa de observaciones por debajo o igual a un valor particular en un conjunto de datos. Representada formalmente por $F_n(x)$, para un conjunto de datos $x_1, x_2, \ldots, x_n$, la ECDF se define como:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

donde I es la función indicadora, que es 1 si $x_i \leq x$ y 0 de lo contrario. La ECDF es una estimación no paramétrica de la verdadera función de distribución subyacente que generó los datos y proporciona una visión visual e intuitiva de la distribución de los datos, lo que es particularmente útil en el análisis exploratorio de datos y en la verificación de suposiciones estadísticas.

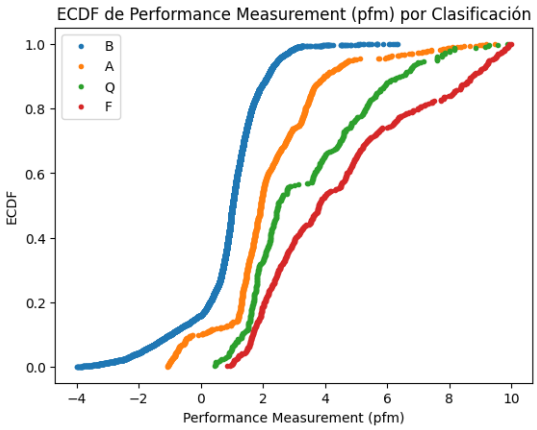


Figure 1: ECDF

4.1.1 Notación Científica de Métricas Estadísticas

Las métricas de evaluación de modelos en la estadística robusta y no paramétrica son fundamentales para medir el rendimiento de los modelos de clasificación, especialmente en presencia de clases desequilibradas. Estas métricas incluyen:

La **Precisión F1 (F1 Score)** compensa entre la precisión y la sensibilidad. La puntuación F1 se define como:

$$F1 = 2 \cdot \frac{\text{precisión} \cdot \text{recall}}{\text{precisión} + \text{recall}}$$

Donde la *precisión* es la proporción de identificaciones positivas que fueron realmente correctas, y el *recall* (sensibilidad) es la proporción de positivos reales que fueron identificados correctamente. La puntuación F1 es especialmente útil en situaciones donde se busca un equilibrio entre precisión y recall.

4.1.2 Índice de Jaccard (Jaccard Score)

Midiendo la similitud entre los conjuntos de etiquetas verdaderas y predichas, el índice de Jaccard, o la intersección sobre la unión (IoU), se calcula como:

$$J(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}$$

donde Y representa el conjunto de etiquetas verdaderas y $\hat{Y}$ el conjunto de etiquetas predichas. Esta fórmula es útil para medir la similitud y la divergencia entre dos conjuntos, como en el caso de las etiquetas verdaderas y las predichas en problemas de clasificación.

Ambas métricas son robustas frente a los datos desequilibrados y no requieren la suposición de una distribución particular de los datos, haciéndolas ideales para este estudio.

4.1.3 Estadística No Paramétrica en la Evaluación del Modelo

La estadística no paramétrica se extiende más allá de la estimación de distribuciones para incluir métodos de inferencia y evaluación que no dependen de una forma específica de la distribución de los datos. En el contexto de este proyecto, la estadística no paramétrica proporciona un marco para evaluar modelos de clasificación sin recurrir a supuestos sobre la normalidad de los datos o la igualdad de varianzas, lo que es esencial en la presencia de desequilibrios de clases y datos industriales que pueden no seguir patrones estadísticos tradicionales.

4.1.4 Carga y Procesamiento del Archivo de Estadísticas

El archivo de estadísticas, que contiene las métricas descriptivas clave para cada grupo de datos, fue cargado en un entorno de análisis de datos Python. Se verificó la integridad de la información, incluyendo la media y la desviación estándar, para asegurar su aptitud para el proceso de clasificación subsiguiente.

memory usage: 10.5+ KB

	Unnamed: 0	stator_kv	mfr	group_number	requested_test_kv	media	mediana	desviacion_estandar	moda	B	A	Q	F
0	0	4.16	GE	GST	2.0	3.712238	3.696725	0.108619	3.604379	3.820856	3.929475	4.038094	False
1	1	4.16	GE	UST	2.0	3.612875	3.613676	0.060816	3.522913	3.673691	3.734508	3.795324	False
2	2	4.16	MIT	GST	2.0	1.102913	1.109547	0.014936	1.085808	1.117849	1.132785	1.147722	False
3	3	4.16	MIT	UST	2.0	1.321190	1.330177	0.076301	1.240793	1.397492	1.473793	1.550095	False
4	4	4.16	OTH	GST	2.0	4.668992	3.066276	2.864054	2.472069	7.533046	10.397099	13.261153	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
104	104	18.00	GE	UST	2.0	1.282481	0.800000	1.931212	0.400000	3.213693	5.144905	7.076118	False
105	105	18.00	GE	UST	4.0	1.279029	0.770000	1.996945	0.300000	3.275973	5.272918	7.269862	False
106	106	18.00	GE	UST	6.0	1.131672	0.629849	2.018955	0.170000	3.150627	5.169582	7.188537	False
107	107	18.00	GE	UST	8.0	1.053292	0.593617	1.855739	0.040000	2.909032	4.764771	6.620510	False
108	108	18.00	GE	UST	10.0	1.051402	0.580000	1.752816	0.220000	2.804218	4.557034	6.309851	False

109 rows × 13 columns

Figure 2: Carga y Procesamiento del Archivo de Estadísticas

4.1.5 Carga y Preparación del Conjunto de Datos Principal

El conjunto de datos principal, *'cleandataSets.csv'*, fue importado en el mismo entorno de análisis. Este archivo proporciona los valores de 'pfm' para cada instancia, los cuales son esenciales para la clasificación. Se realizó un preprocesamiento adecuado, incluyendo la limpieza de datos y la codificación de variables categóricas.

memory usage: 10.5+ KB

	Unnamed: 0	stator_kv	mfr	group_number	requested_test_kv	media	mediana	desviacion_estandar	moda	B	A	Q	F
0	0	4.16	GE	GST	2.0	3.712238	3.696725	0.108619	3.604379	3.820856	3.929475	4.038094	False
1	1	4.16	GE	UST	2.0	3.612875	3.613676	0.060816	3.522913	3.673691	3.734508	3.795324	False
2	2	4.16	MIT	GST	2.0	1.102913	1.109547	0.014936	1.085808	1.117849	1.132785	1.147722	False
3	3	4.16	MIT	UST	2.0	1.321190	1.330177	0.076301	1.240793	1.397492	1.473793	1.550095	False
4	4	4.16	OTH	GST	2.0	4.668992	3.066276	2.864054	2.472069	7.533046	10.397099	13.261153	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
104	104	18.00	GE	UST	2.0	1.282481	0.800000	1.931212	0.400000	3.213693	5.144905	7.076118	False
105	105	18.00	GE	UST	4.0	1.279029	0.770000	1.996945	0.300000	3.275973	5.272918	7.269862	False
106	106	18.00	GE	UST	6.0	1.131672	0.629849	2.018955	0.170000	3.150627	5.169582	7.188537	False
107	107	18.00	GE	UST	8.0	1.053292	0.593617	1.855739	0.040000	2.909032	4.764771	6.620510	False
108	108	18.00	GE	UST	10.0	1.051402	0.580000	1.752816	0.220000	2.804218	4.557034	6.309851	False

109 rows × 13 columns

Figure 3: Carga y Preparación del Conjunto de Datos Principal

4.1.6 Clasificación de Datos

Con los datos estadísticos a mano, se aplicaron las reglas de clasificación predefinidas para asignar cada valor de 'pfm' a una de las categorías designadas: B, A, Q o F. Este proceso se realizó utilizando operaciones vectorizadas para garantizar la eficiencia.

memory usage: 10.5+ KB

	Unnamed: 0	stator_kv	mfr	group_number	requested_test_kv	media	mediana	desviacion_estandar	moda	B	A	Q	F
0	0	4.16	GE	GST	2.0	3.712238	3.696725	0.108619	3.604379	3.820856	3.929475	4.038094	False
1	1	4.16	GE	UST	2.0	3.612875	3.613676	0.060816	3.522913	3.673691	3.734508	3.795324	False
2	2	4.16	MIT	GST	2.0	1.102913	1.109547	0.014936	1.085808	1.117849	1.132785	1.147722	False
3	3	4.16	MIT	UST	2.0	1.321190	1.330177	0.076301	1.240793	1.397492	1.473793	1.550095	False
4	4	4.16	OTH	GST	2.0	4.668992	3.066276	2.864054	2.472069	7.533046	10.397099	13.261153	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
104	104	18.00	GE	UST	2.0	1.282481	0.800000	1.931212	0.400000	3.213693	5.144905	7.076118	False
105	105	18.00	GE	UST	4.0	1.279029	0.770000	1.996945	0.300000	3.275973	5.272918	7.269862	False
106	106	18.00	GE	UST	6.0	1.131672	0.629849	2.018955	0.170000	3.150627	5.169582	7.188537	False
107	107	18.00	GE	UST	8.0	1.053292	0.593617	1.855739	0.040000	2.909032	4.764771	6.620510	False
108	108	18.00	GE	UST	10.0	1.051402	0.580000	1.752816	0.220000	2.804218	4.557034	6.309851	False

109 rows × 13 columns

Figure 4: Carga y Preparación del Conjunto de Datos Principal

4.1.7 Cálculo de la Función de Distribución Empírica (FDE)

Posteriormente, se calculó la Función de Distribución Empírica (ECDF) para cada categoría de clasificación, utilizando los valores de 'pfm' clasificados. La ECDF proporciona una representación visual de la distribución acumulativa de los datos, lo que facilita la identificación de patrones y anomalías.

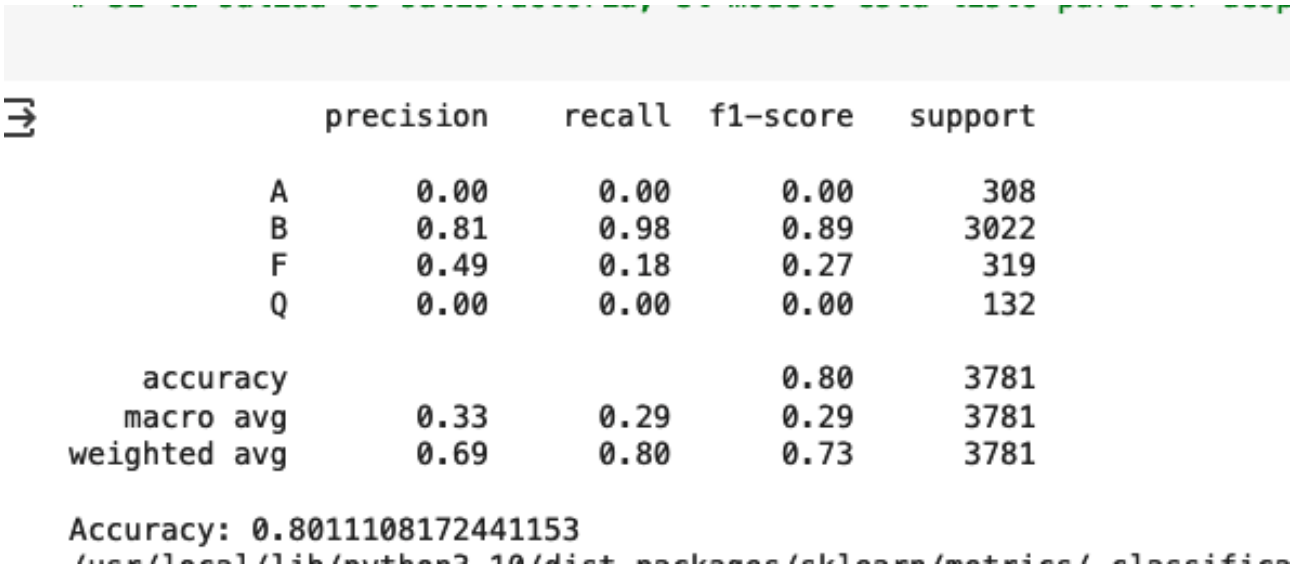
```python
import numpy as np
import matplotlib.pyplot as plt

def ecdf(data):
    """Calcular la ECDF para un conjunto de datos."""
    x = np.sort(data)
    y = np.arange(1, len(x) + 1) / len(x)
    return x, y

# Graficar la ECDF para cada clasificación
classifications = ['B', 'A', 'Q', 'F']
for classification in classifications:
    x, y = ecdf(_data[_data['classification'] == classification]['pfm'])
    plt.plot(x, y, marker='.', linestyle='none', label=classification)

plt.legend()
plt.xlabel('Performance Measurement (pfm)')
plt.ylabel('ECDF')
plt.title('ECDF de Performance Measurement (pfm) por Clasificación')
plt.show()
```

Figure 5: ECDF

4.1.8 Evaluación del Modelo

Se entrenó un modelo de clasificación utilizando el algoritmo Random Forest, aprovechando los datos equilibrados a través de la técnica SMOTE. Se aplicaron procedimientos de validación cruzada y optimización de hiperparámetros

para afinar el rendimiento del modelo. La evaluación se basó en métricas robustas, incluyendo la precisión F1 ponderada y el índice de Jaccard, para proporcionar una medida integral del rendimiento del modelo.



	precision	recall	f1-score	support
A	0.00	0.00	0.00	308
B	0.81	0.98	0.89	3022
F	0.49	0.18	0.27	319
Q	0.00	0.00	0.00	132
accuracy			0.80	3781
macro avg	0.33	0.29	0.29	3781
weighted avg	0.69	0.80	0.73	3781

Accuracy: 0.8011108172441153

Figure 6: Evaluación del Model

```
Jaccard por clase: [0.         0.79994615 0.1530343  0.        ]
Jaccard promedio macro: 0.238245112623845
Jaccard promedio ponderado: 0.6522759075264378
              precision    recall  f1-score   support

           A       0.00      0.00      0.00       308
           B       0.81      0.98      0.89      3022
           F       0.49      0.18      0.27       319
           Q       0.00      0.00      0.00       132

    accuracy                           0.80      3781
   macro avg       0.33      0.29      0.29      3781
weighted avg       0.69      0.80      0.73      3781

Accuracy: 0.8011108172441153
Jaccard Index - Macro Average: 0.238245112623845
Jaccard Index - Weighted Average: 0.6522759075264378
```

Figure 7: Evaluación del Model

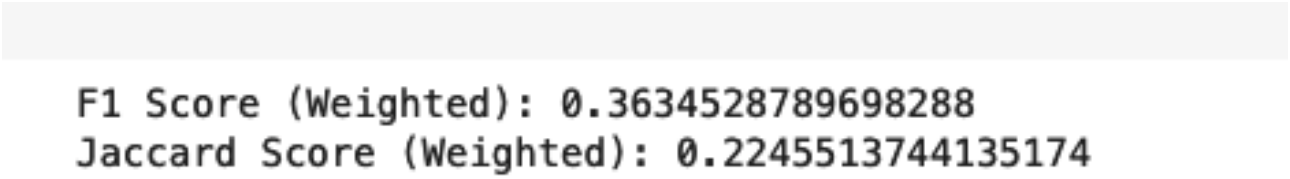


Figure 8: Evaluación del Model

5 Enfoque mineria de datos

5.1 Aplicación de Técnicas de Clustering

se aplicaron dos algoritmos de clustering ampliamente reconocidos en la minería de datos: K-Means y el Modelo de Mezcla Gaussiana (GMM). Cada uno tiene características únicas y es adecuado para diferentes tipos de estructuras de datos.

5.2 K-Means

5.2.1 Descripción

Es un algoritmo de clustering particional que divide el conjunto de datos en K clusters. El objetivo es minimizar la varianza dentro de cada cluster. Cada punto de datos se asigna al centroide del cluster más cercano.

5.2.2 Aplicación

Se utilizó K-Means debido a su simplicidad y eficacia en grandes conjuntos de datos. Es especialmente útil cuando se espera que los clusters tengan tamaños y densidades similares.

5.3 Modelo de Mezcla Gaussiana (GMM)

5.3.1 Descripción:

GMM es un algoritmo de clustering basado en la probabilidad que asume que los datos se distribuyen como una mezcla de varias distribuciones gaussianas. Cada cluster se modela como una distribución gaussiana.

5.3.2 Aplicación:

Se eligió GMM para este proyecto debido a su flexibilidad en la forma de los clusters, lo que permite identificar grupos con formas elípticas y tamaños variables.

5.4 K-Means - Método del Codo

5.4.1 Concepto

El método del codo implica trazar la varianza explicada en función del número de clusters y buscar un punto donde el aumento de la varianza explicada no justifica el aumento en el número de clusters.

5.4.2 Implementación

Se aplicó este método trazando la suma de las distancias cuadradas de los puntos de datos a su centroide más cercano. Se buscó un "codo" en el gráfico donde un aumento adicional en el número de clusters no resultó en una mejora significativa en la varianza total explicada.

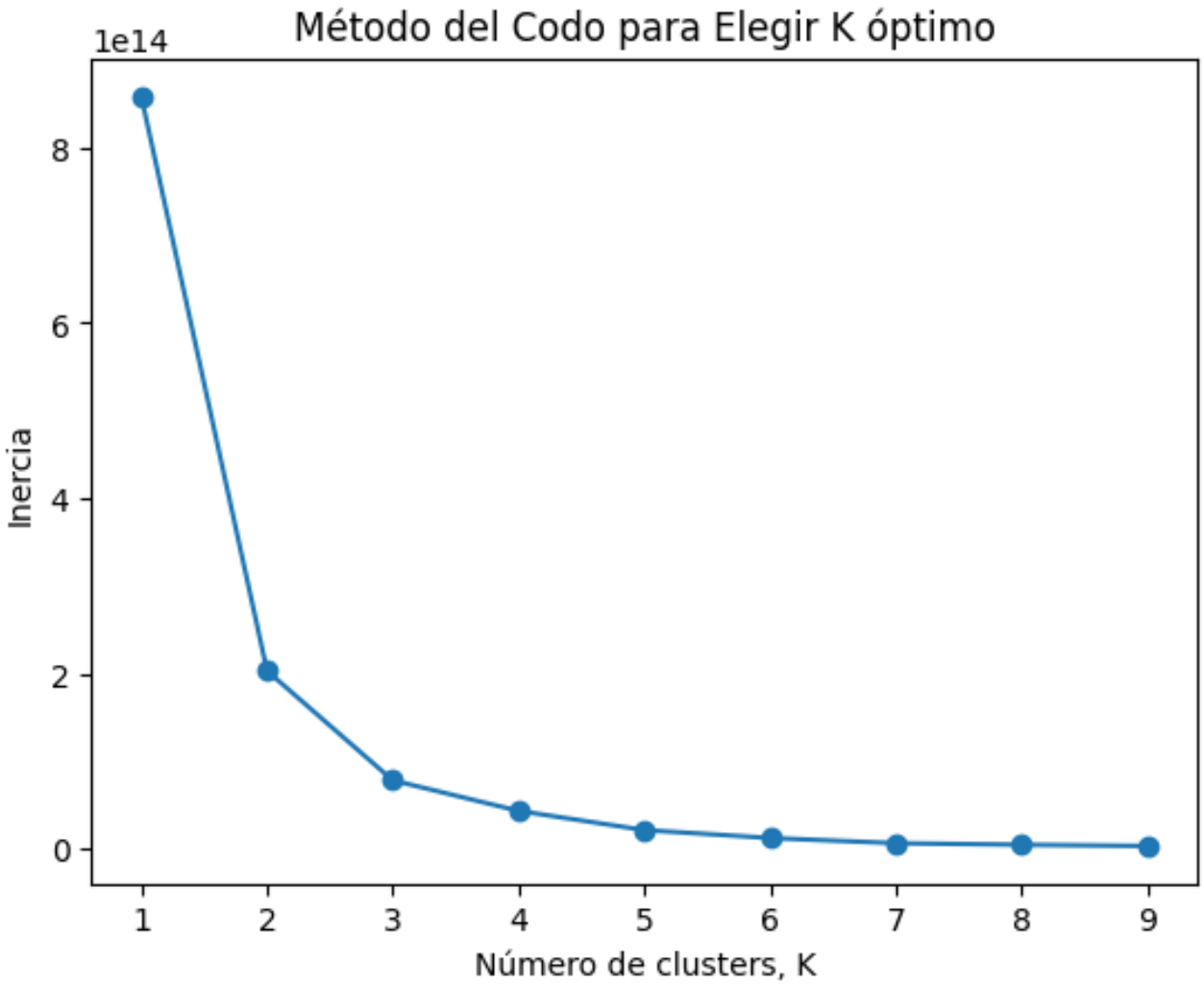


Figure 9: Evaluación del Model

La gráfica del método del codo que has generado muestra que la inercia disminuye significativamente a medida que aumentamos el número de clusters (K) de 1 a 2 y sigue disminuyendo a un ritmo más lento para valores mayores de K. La elección de K óptimo generalmente se hace donde se observa un cambio en la pendiente de la inercia que ya no es tan pronunciado, lo que a menudo se describe como el "punto de inflexión" o "codo".

este punto parece estar alrededor de K=3, donde la disminución de la inercia comienza a aplanarse.

5.5 GMM - Criterios BIC y AIC

5.5.1 Concepto:

BIC y AIC son medidas que penalizan la complejidad del modelo y recompensan el buen ajuste del modelo a los datos. Ambos criterios buscan el equilibrio entre la complejidad del modelo y la explicación de la varianza.

5.5.2 Implementación:

Se calcularon los valores de BIC y AIC para diferentes números de componentes en GMM. Se eligió el número de clusters que minimizó estos valores, proporcionando así una medida de calidad del modelo ajustado.

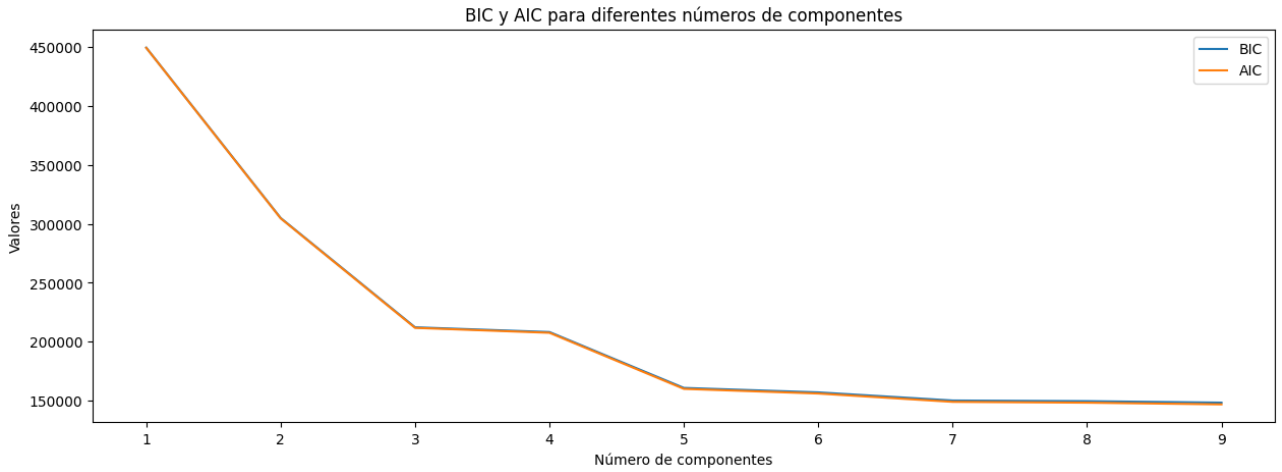


Figure 10: Criterios BIC y AIC

Estos criterios ayudaron a determinar el número óptimo de clusters en ambos algoritmos, asegurando así que el modelo de clustering fuera ni demasiado simple ni excesivamente complejo para la estructura inherente de los datos.

Tendencia: Ambos, el BIC (Criterio de Información Bayesiano) y el AIC (Criterio de Información de Akaike), disminuyen a medida que el número de componentes en el Modelo de Mezcla Gaussiana (GMM) aumenta. Esto es típico, ya que agregar más componentes generalmente mejora el ajuste del modelo a los datos.

Selección de Componentes: Se busca el punto donde el BIC y el AIC ya no disminuyen sustancialmente con componentes adicionales. Esto indica un equilibrio entre la complejidad del modelo y el ajuste a los datos. En el análisis gráfico, este punto de equilibrio parece ser alrededor de 2 o 3 componentes, ya que después de eso, la disminución en los valores de BIC y AIC es menos pronunciada.

Número Óptimo de Componentes: Basándonos en este análisis gráfico, se podría elegir 2 o 3 como el número óptimo de componentes para el modelo GMM. La elección específica entre 2 o 3 podría depender de consideraciones adicionales, como la interpretación de los clusters en el contexto de la aplicación específica y el deseo de un modelo más simple frente a uno más detallado.

5.6 Evaluación de Modelos de Clustering

La evaluación de modelos de clustering es un paso crucial para entender la efectividad y la pertinencia de los clusters identificados. En este proyecto, se utilizaron varios métodos para evaluar los modelos de clustering, incluyendo el Puntaje de Silueta, el Índice Rand Ajustado (ARI) y la Información Mutua Normalizada (NMI). Además, se realizó una comparación con etiquetas reales cuando estuvieron disponibles.

5.7 Puntaje de Silueta

5.7.1 Descripción

Este puntaje mide cuán similar es un objeto a su propio cluster en comparación con otros clusters. Varía de -1 a +1, donde un valor alto indica que los objetos están bien emparejados a su propio cluster y mal emparejados a los vecinos.

5.7.2 Uso en el Proyecto

El Puntaje de Silueta se utilizó para evaluar la cohesión y la separación de los clusters. Un puntaje cercano a +1 sugiere clusters bien definidos y separados, mientras que un puntaje cercano a 0 o negativo indica una superposición significativa entre los clusters.

5.8 Índice Rand Ajustado (ARI)

5.8.1 Descripción

ARI mide la similitud entre dos asignaciones de clustering, considerando todas las parejas de muestras y contando pares asignados en el mismo o en diferentes clusters en las asignaciones predichas y verdaderas.

5.8.2 Uso en el Proyecto

ARI se utilizó para comparar las etiquetas de clustering con un conjunto de etiquetas reales. Un valor de 1 indica una correspondencia perfecta, mientras que un valor cercano a 0 o negativo sugiere un acuerdo aleatorio o pobre entre las asignaciones.

5.9 Información Mutua Normalizada (NMI)

5.9.1 Descripción

NMI es una normalización del puntaje de Información Mutua que mide la cantidad de información compartida por las asignaciones de clustering y las etiquetas reales.

5.9.2 Uso en el Proyecto

NMI se utilizó para evaluar la cantidad de información compartida entre las etiquetas de clustering y las etiquetas reales. Al igual que con ARI, un valor más alto indica una mejor correspondencia entre las asignaciones de clustering y las etiquetas reales.

5.10 Proceso de Comparación con Etiquetas Reales

5.10.1 Cuando se Disponen de Etiquetas Reales

- En casos donde se disponía de etiquetas reales, se realizó una comparación directa entre las etiquetas generadas por los algoritmos de clustering y estas etiquetas reales.
- Se utilizaron ARI y NMI para cuantificar la correspondencia entre las asignaciones de clustering y las categorizaciones reales de los datos. Esto proporcionó una medida objetiva de la efectividad del modelo de clustering en replicar agrupaciones conocidas o esperadas en el conjunto de datos.

```python
import pandas as pd
import numpy as np

# Asegúrate de que los nombres de las columnas en 'estadisticas_df' coincidan con estos
media = estadisticas_df['media'].iloc[0]
desviacion_estandar = estadisticas_df['desviacion_estandar'].iloc[0]

# Definir la función de clasificación
def clasificar_pfm(row):
    if row['pfm'] <= media - desviacion_estandar:
        return 'B'
    elif row['pfm'] > media - desviacion_estandar and row['pfm'] <= media - 2 * desviacion_estandar:
        return 'A'
    elif row['pfm'] > media - 2 * desviacion_estandar and row['pfm'] <= media - 3 * desviacion_estandar:
        return 'Q'
    else:  # Aquí se asume que cualquier valor mayor a 'media - 3 * desviacion_estandar' es categoría 'F'
        return 'F'

# Aplicar la función de clasificación
clean_dataSets_df['Etiqueta'] = clean_dataSets_df.apply(clasificar_pfm, axis=1)

# Mostrar los resultados
print(clean_dataSets_df[['pfm', 'Etiqueta']].head())
```

```
        pfm Etiqueta
0 -0.122773        B
1 -0.011785        B
2  0.085386        B
3  0.186858        B
4 -0.125803        B
```

Figure 11: etiquetas

5.10.2 Importancia de la Comparación

- La comparación con etiquetas reales es fundamental para validar la utilidad del clustering en aplicaciones prácticas, especialmente en campos donde las categorizaciones previamente definidas son críticas, como en la medicina, la biología o la segmentación de clientes.
- Permite a los analistas y científicos de datos entender mejor si los patrones identificados por los modelos de clustering son significativos y si reflejan estructuras subyacentes conocidas o hipótesis sobre los datos.

5.11 Resultados del Clustering

Los resultados obtenidos de los modelos de clustering proveen una visión valiosa sobre la estructura y las características subyacentes de los datos.

5.11.1 Visualizaciones

- Las visualizaciones son herramientas esenciales para interpretar los resultados del clustering. Se pueden utilizar gráficos como scatter plots, gráficos de barras, o mapas de calor para mostrar cómo se agrupan los datos en diferentes clusters.
- En el caso de datos multidimensionales, técnicas de reducción de dimensionalidad como PCA (Análisis de Componentes Principales) o t-SNE pueden ser utilizadas para visualizar los clusters en dos o tres dimensiones.

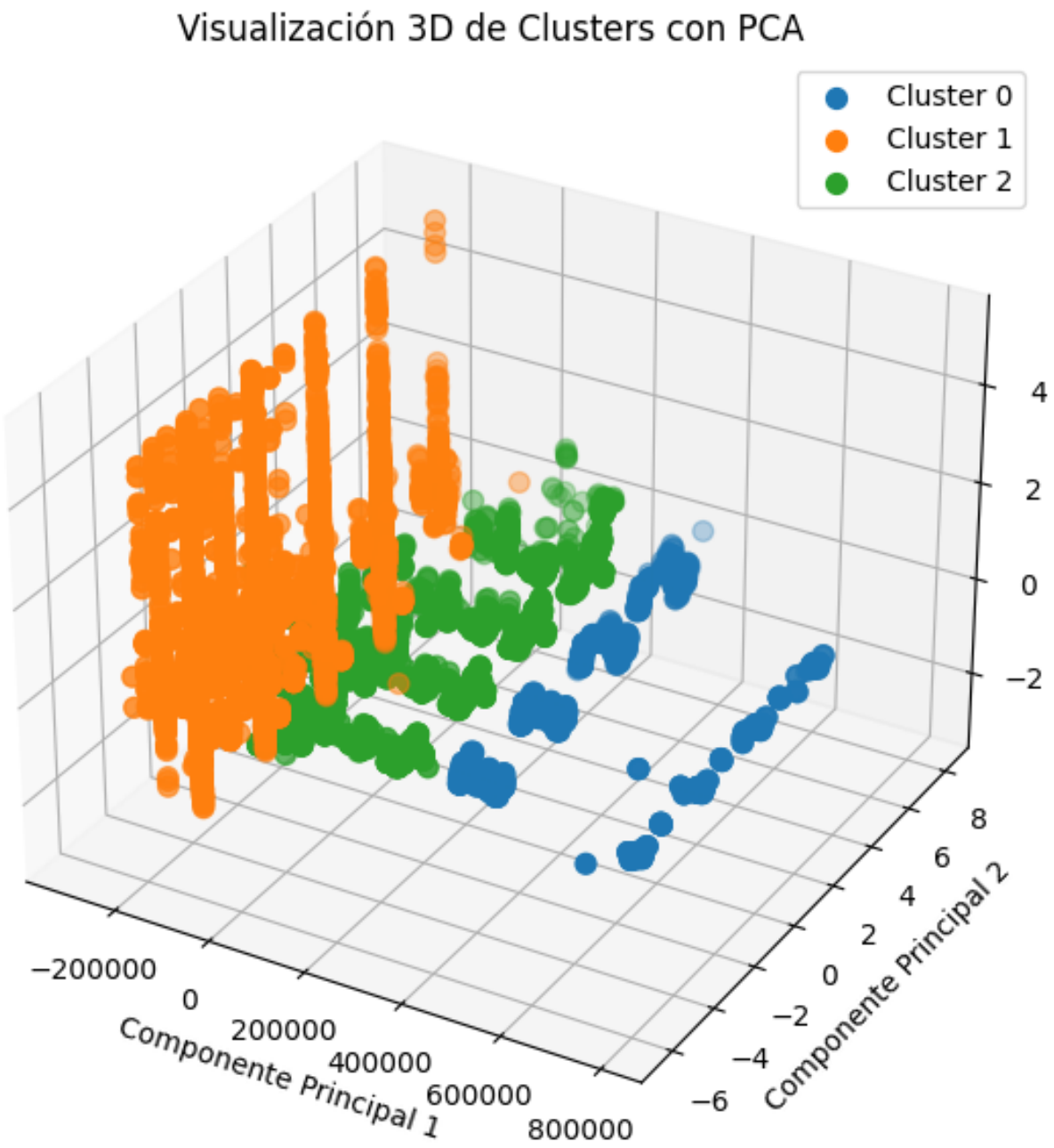


Figure 12: Evaluación del Model

hay una separación clara entre el cluster azul y los otros dos clusters a lo largo del eje del primer componente principal (PC1). Sin embargo, los clusters naranja y verde parecen tener cierta superposición a lo largo de PC1 y PC2. Esto puede indicar que el cluster azul tiene características distintivas que lo separan de los otros dos, mientras que los clusters naranja y verde podrían ser más difíciles de diferenciar basándose en las características seleccionadas para el clustering.

5.11.2 Descripción de los Clusters

- Se presenta una descripción detallada de cada cluster, incluyendo el número de observaciones, las características medias y otras estadísticas relevantes.
- Se identifican y describen las características distintivas de cada cluster, basándose en las variables más significativas que contribuyen a la formación del cluster.

6 Conclusiones

- F1 Score (Weighted): 0.3634528789698288

  Jaccard Score (Weighted): 0.2245513744135174

  Los resultados obtenidos para el F1 Score Ponderado y el Índice de Jaccard Ponderado son significativamente más bajos de lo que se esperaría para un modelo bien ajustado. Un F1 Score Ponderado de aproximadamente 0.36 y un Índice de Jaccard Ponderado de aproximadamente 0.22 sugieren que el modelo no está realizando una buena clasificación general, particularmente en cuanto a la precisión y el recall de las clases minoritarias.

  Estos resultados sugieren que es necesario un ajuste adicional del modelo o una exploración más profunda de los datos.

- Índice de Jaccard por Clase: Los valores son 0 para las clases A y Q, lo que indica que el modelo no pudo identificar correctamente ninguna instancia de estas clases. Para la clase B, el índice es bastante alto (aproximadamente 0.80), lo que sugiere que el modelo clasifica bien esta clase. La clase F tiene un índice bajo (aproximadamente 0.15), lo que indica un rendimiento pobre en esta clase.

- Índice de Jaccard Promedio Macro: El promedio ponderado es aproximadamente 0.65, que es más alto que el promedio macro debido a la influencia de la clase mayoritaria (B).

- Índice de Jaccard Promedio Ponderado: El promedio ponderado es aproximadamente 0.65, que es más alto que el promedio macro debido a la influencia de la clase mayoritaria (B).

- Precisión y Recall: La precisión y el recall son altos para la clase B, pero bajos o nulos para las clases A y Q, y moderados para la clase F.

- Accuracy: A pesar de que la precisión general es del 80%, este número está influenciado en gran medida por la clase B, que es la clase mayoritaria.