Universidad de La Laguna

Tratamiento Inteligente de Datos

Nombre de la Práctica

09/03/2024

Eric Angueta Rogel
alu0101335339@ull.edu.es

Universidad de La Laguna

Índice

Universidad de La Laguna

Preparación de Datos

Valores ausentes

Para tratar los valores nulos o valores ausentes he escogido la técnica de k-Means Clustering Imputation en esta estrategia, las observaciones se agrupan en clusters utilizando el **algoritmo k-means**. Luego, los valores ausentes se imputan utilizando la media o mediana del clúster al que pertenece la observación. Puede ser útil cuando hay patrones estructurados en los datos.

Explicaré brevemente en qué consisten las demás técnicas.

- **Valor Más Común en el Concepto (Most Common Value in Concept):** Similar a la estrategia anterior, pero aquí se considera el valor más común dentro de un subconjunto específico o concepto relacionado con la observación que tiene el valor faltante. Puede ser útil en situaciones donde la variabilidad de los datos es alta.
- **k-Nearest Neighbor Imputation (kNNI):** Esta técnica utiliza la información de las observaciones más cercanas (vecinos más cercanos) para imputar el valor faltante. Se calcula la similitud entre observaciones y se asigna el valor basándose en los k vecinos más cercanos. Es eficaz para datos continuos y puede preservar patrones más complejos.
- **Valor Más Común (Most Common Value):** En esta técnica, los valores ausentes se llenan con el valor más frecuente presente en la variable correspondiente. Es una opción sencilla y rápida, adecuada cuando la variable es categórica o discreta.

Conversión de tipos de datos

Para las variables categóricas se ha utilizado la clase LabelEncoder de scikit-learn para transformar los tipos de datos.

Tratamiento de Outliers

El tratamiento de *outliers* utilizado en esta práctica es la eliminación de los mismos.

Clase balanceada

Al tratarse de clases desbalanceadas se implementa en el cuaderno el algoritmo SMOTE.

Para llegar a esta conclusión tenemos en cuenta lo siguiente:
Calculamos la frecuencia de cada clase y luego evaluamos si la diferencia en las proporciones con respecto a la clase mayoritaria es menor que el umbral especificado. Si todas las clases tienen proporciones cercanas a la clase mayoritaria, se considera que la clase está balanceada.

Bases de Datos

Accuracy (Precisión Global):

- Definición: El accuracy es la proporción de predicciones correctas respecto al total de instancias en el conjunto de datos.
- Interpretación: Indica la tasa general de aciertos del modelo. Sin embargo, puede ser engañosa en conjuntos de datos desbalanceados, donde la precisión en la clase mayoritaria puede inflar el valor del accuracy.

Precisión:

- Definición: La precisión mide la proporción de instancias positivas correctamente clasificadas entre todas las instancias predichas como positivas.
- Interpretación: Evalúa la exactitud de las predicciones positivas del modelo. Es útil cuando los falsos positivos son críticos y deben minimizarse.

Recall (Sensibilidad o Tasa de Verdaderos Positivos):

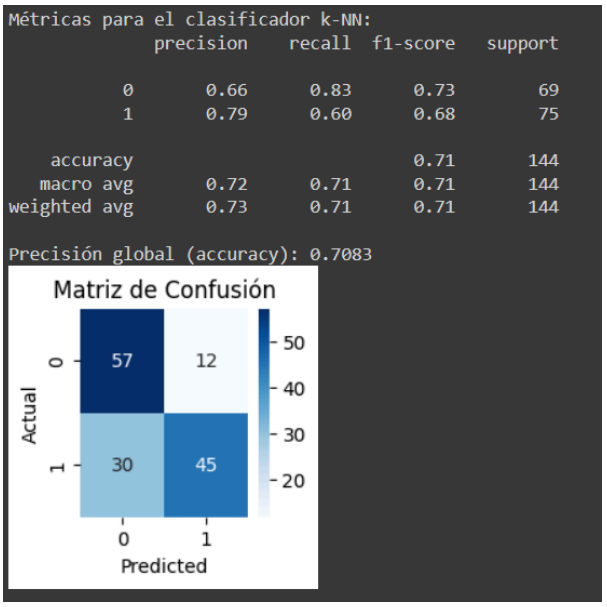
- Definición: El recall mide la proporción de instancias positivas correctamente clasificadas entre todas las instancias reales positivas.
- Interpretación: Evalúa la capacidad del modelo para capturar todas las instancias positivas en el conjunto de datos. Es útil cuando los falsos negativos son críticos y deben minimizarse.

F1-score:

- Definición: El F1-score es la media armónica de la precisión y el recall, proporcionando un equilibrio entre ambas métricas.
- Interpretación: Es especialmente útil en conjuntos de datos desbalanceados, donde tanto los falsos positivos como los falsos negativos son importantes. Un F1-score alto indica un buen equilibrio entre precisión y recall.

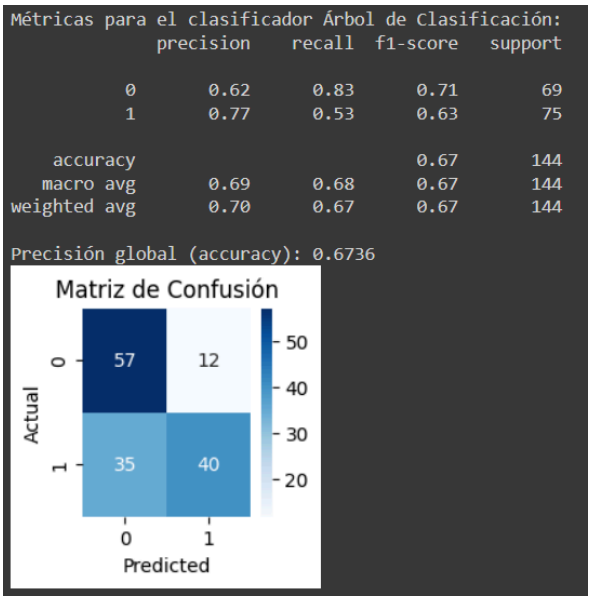
Base de Datos Tratada:

k-NN:



- Accuracy: 0.71
- Precisión y recall equilibrados para ambas clases (0 y 1).
- F1-score promedio del 71%.
- Buen rendimiento general, aunque la precisión global podría mejorarse.

Árbol de Clasificación:



- Accuracy: 0.67
- Mejor recall para la clase 0, pero mejor precisión para la clase 1.
- F1-score promedio del 67%.

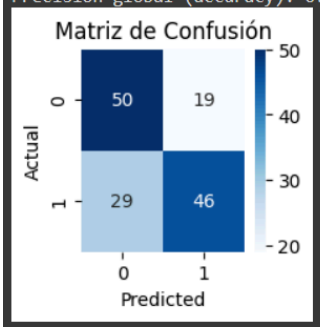
- Rendimiento aceptable, pero hay margen para mejoras en la precisión global.

Naive Bayes:

Métricas para el clasificador Naive Bayes:

	precision	recall	f1-score	support
0	0.63	0.72	0.68	69
1	0.71	0.61	0.66	75
accuracy			0.67	144
macro avg	0.67	0.67	0.67	144
weighted avg	0.67	0.67	0.67	144

Precisión global (accuracy): 0.6667



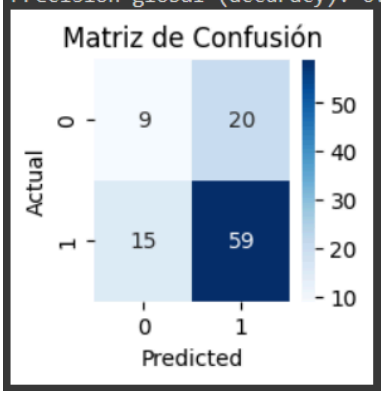
- Accuracy: 0.67
- Equilibrio en precisiones y recalls para ambas clases.
- F1-score promedio del 67%.
- Rendimiento similar al Árbol de Clasificación, pero con menos variación entre las clases.

Universidad de La Laguna

Base de Datos sin Tratar:

k-NN:

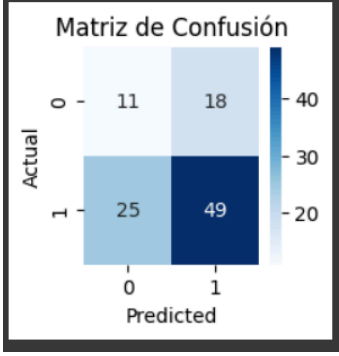
```
Métricas para el clasificador k-NN:
              precision    recall  f1-score   support

           0       0.38      0.31      0.34        29
           1       0.75      0.80      0.77        74

    accuracy                           0.66       103
   macro avg       0.56      0.55      0.56       103
weighted avg       0.64      0.66      0.65       103

Precisión global (accuracy): 0.6602
```



- Accuracy: 0.66
- Mayor precisión para la clase 1, pero menor recall.
- F1-score promedio del 65%.
- Rendimiento moderado, con espacio para mejoras en la precisión global.

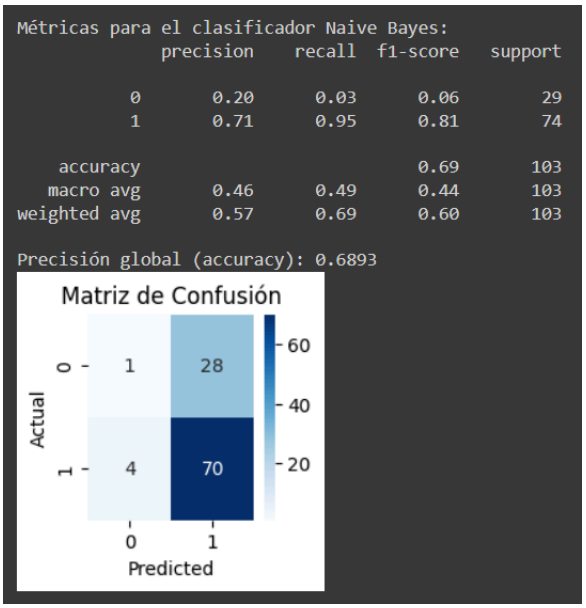
Árbol de Clasificación:

```
Métricas para el clasificador Árbol de Clasificación:
              precision    recall  f1-score   support

           0       0.31      0.38      0.34        29
           1       0.73      0.66      0.70        74

    accuracy                           0.58       103
   macro avg       0.52      0.52      0.52       103
weighted avg       0.61      0.58      0.59       103

Precisión global (accuracy): 0.5825
```



- Accuracy: 0.58
- Mayor recall para la clase 1, pero menor precisión.
- F1-score promedio del 59%.
- Rendimiento más bajo en comparación con k-NN.

Naive Bayes:



- Accuracy: 0.69
- Mayor recall y precisión para la clase 1.
- F1-score promedio del 60%.
- Mejor rendimiento entre los tres clasificadores en la base de datos sin tratar.

Enlace a cuader Jupyter

[Enlace a Cuaderno Jupyter]