Universidad Panamericana

Estudiante: Javier Horacio Pérez Ricárdez
Asignatura: Aprendizaje Máquina Aplicado
Profesor: Omar Velázquez López

Actividad: Análisis de Cáncer de Seno (Wisconsin)

Fecha:

Objetivo:

El objetivo de esta actividad es analizar el conjunto de datos de cáncer de seno de Wisconsin utilizando dos modelos de aprendizaje automático: Naïve Bayes y Regresión Logística. Se evalúa el rendimiento de ambos modelos en términos de precisión, recall, F1-Score y matrices de confusión, con el fin de determinar cuál es más adecuado para este conjunto de datos.

Contenido:

El análisis se realizó utilizando un conjunto de datos de cáncer de seno de Wisconsin, que contiene características como el grosor del tumor, la uniformidad del tamaño de las células, la adhesión marginal, entre otras. El conjunto de datos se dividió en conjuntos de entrenamiento, validación y prueba. Se entrenaron dos modelos: Naïve Bayes y Regresión Logística, y se evaluaron en los conjuntos de validación y prueba.

Resultados:

A continuación se presentan los resultados obtenidos por ambos modelos:

Cuadro 1: Resultados de los Modelos

Modelo	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naïve Bayes	Validación	0.9416	0.8824	0.9574	0.9184
Naïve Bayes	Prueba	0.9635	0.9138	1	0.95
Regresión Logística	Validación	0.9416	0.9149	0.9149	0.9149
Regresión Logística	Prueba	0.9708	0.9298	1	0.9636

Análisis de Resultados:

1. Comparación de Modelos:

En el conjunto de validación, **Naïve Bayes** obtuvo un F1-Score de 0.9184, mientras que **Regresión Logística** obtuvo un F1-Score de 0.9149. Esto indica que la Regresión Logística tiene un mejor equilibrio entre precisión y recall. En cuanto a la precisión, Regresión Logística obtuvo 0.9149, mientras que Naive Bayes obtuvo 0.8824.
Sin embargo, el F1-Score sugiere que Naive Bayes es más adecuado para este conjunto de datos.

Además, observando las matrices de confusión, Naive Bayes mostró 45 verdaderos positivos, 6 falsos positivos, 2 falsos negativos y 84 verdaderos negativos en validación. En comparación, la Regresión Logística tuvo 43 verdaderos positivos, 4 falsos positivos, 4 falsos negativos y 86 verdaderos negativos.

2. Desbalance de Clases:

La distribución de clases en el conjunto de datos es la siguiente:

- Clase 0 (Benigno): 444 casos
- Clase 1 (Maligno): 239 casos

En el conjunto de prueba, Naive Bayes tiene un recall de 1, lo que indica que identifica correctamente la mayoría de los casos positivos.

Además, Regresión Logística también tiene un reacall de 1, lo que sugiere que ambos modelos son efectivos para identificar casos positivos.

Al observar las matrices de confusión, podemos ver que ambos modelos tienen un buen desempeño en términos de recall, pero la cantidad de falsos positivos y falsos negativos es clave para entender el rendimiento.

3. Valores Faltantes:

Filas originales: 699, filas después de limpieza: 683.

4. Recomendación de Modelos:

En el conjunto de prueba, **Naïve Bayes** obtuvo un F1-Score de 0.9550, mientras que **Regresión Logística** obtuvo un F1-Score de 0.9636. Esto sugiere que Naive Bayes es más adecuado para este conjunto de datos debido a su mejor equilibrio entre precisión y recall.

Naive Bayes es recomendable cuando se trabaja con conjuntos de datos pequeños o cuando las características son independientes.

En este caso, Naive Bayes demostró un mejor rendimiento general en términos de F1-Score.

Conclusión:

En resumen, se recomienda usar Naïve Bayes para este conjunto de datos debido a su mejor rendimiento general en términos de F1-Score, y su capacidad para manejar el desbalance de clases.

Bibliografía (APA):

- Dua, D. y Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- Actividad 3: `https://actividad3-aejhkdnuskdjlknmspe7q9.streamlit.app/`