Informe de Resultados y Configuraciones de Experimentación

Proyecto: Implementación de Algoritmos de Clasificación

Nombre: Ricardo Coronado Mera
Docente: Cristian Olivares
Fecha: 21 de Junio, 2022

Introducción
En este informe, se presentan los resultados obtenidos para diferentes algoritmos de clasificación implementados en el proyecto. Se utilizaron diferentes técnicas de validación y se variaron las configuraciones para evaluar el rendimiento de cada algoritmo. A continuación, se detallan los algoritmos implementados y las métricas utilizadas para evaluar su desempeño:

Resultados:

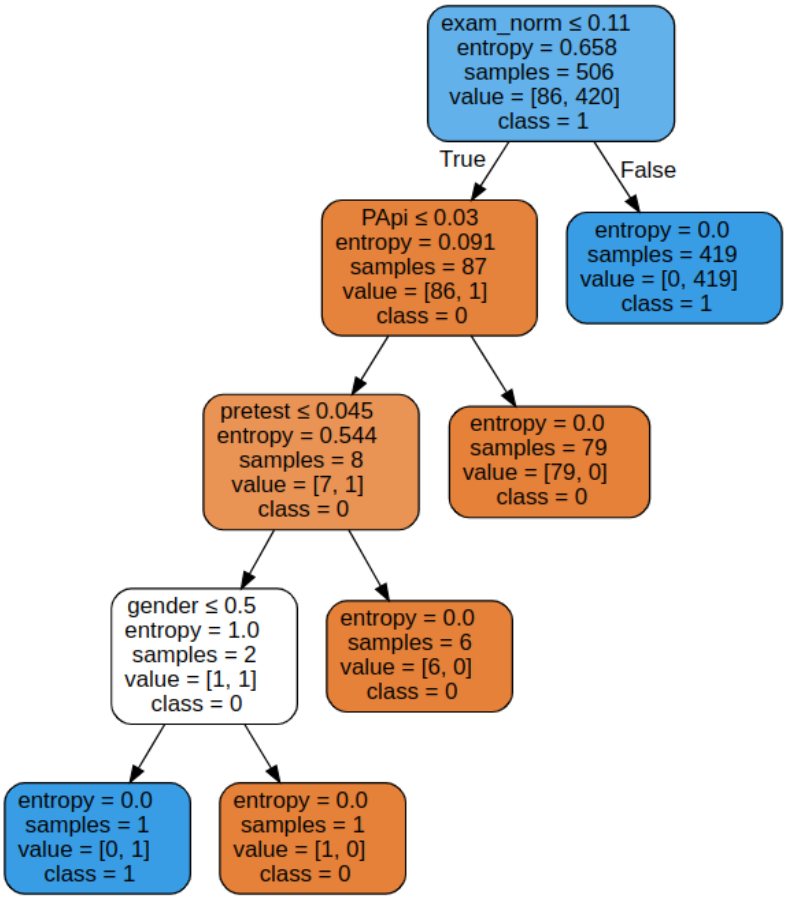
1. Árbol de decisión con entropía:

    - Configuración: Se utilizó la medida de entropía para la selección de atributos.
    - Métricas evaluadas:

validación	precisión cross-validation	precisión leave-one-out
0.1	0.99	0.99
0.2	0.99	0.99
0.3	0.99	0.99

Para test size 0.2	precision	recall	f1-score	support
no da examen	1.00	1.00	1.00	17
da examen	1.00	1.00	1.00	110
accuracy			1.00	127
macro avg	1.00	1.00	1.00	127
weighted avg	1.00	1.00	1.00	127

Árbol:

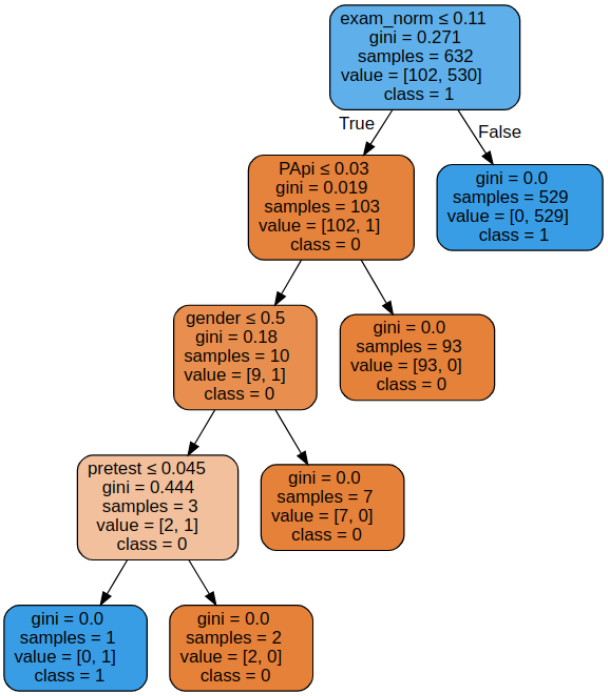


2. <u>Árbol de decisión con ganancia de información (Gini):</u>

   - Configuración: Se utilizó la medida de ganancia de información (Gini) para la selección de atributos.
   - Métricas evaluadas:

validación	precisión cross-validation	precisión leave-one-out
0.1	1.00	0.99
0.2	1.00	0.99
0.3	0.99	0.99

Árbol:

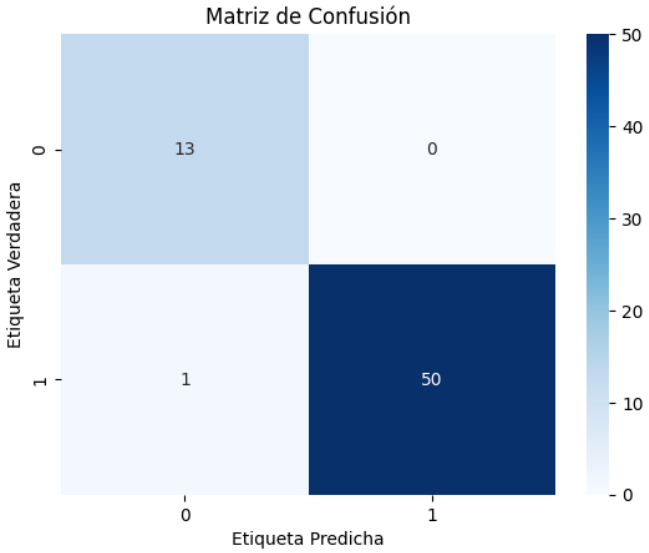


3. Naive Bayes:

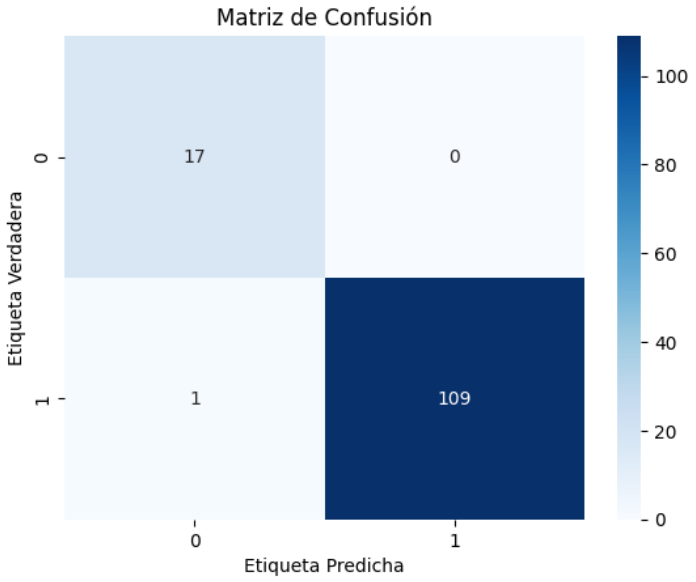
- Configuración: Implementación del algoritmo Naive Bayes para clasificación.
- Métricas evaluadas: Precisión, recall y F1-score.

Test size	Accuracy
0.1	0.984375
0.2	0.9921259842519685
0.3	0.9947368421052631

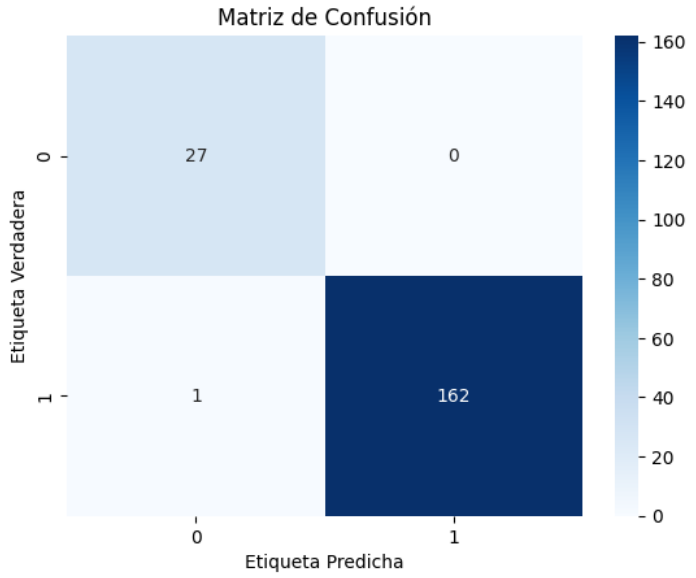
Matriz de confusión para test_size: 0.1



Matriz de confusión para test_size: 0.2



Matriz de confusión para test_size: 0.3



4. <u>K-Nearest Neighbors (KNN):</u>

   - Configuración: Implementación del algoritmo K-Nearest Neighbors para clasificación con 3, 4 y 5 vecinos.
   - Métricas evaluadas: Precisión, recall y F1-score.

test_size: 0.1

k: 3	precision	recall	f1-score	support
no da examen	0.75	0.69	0.72	13
da examen	0.92	0.94	0.93	51
accuracy			0.89	64

k: 4	precision	recall	f1-score	support
no da examen	0.75	0.69	0.72	13
da examen	0.92	0.94	0.93	51
accuracy			0.89	64

k: 5	precision	recall	f1-score	support
no da examen	0.89	0.62	0.73	13
da examen	0.91	0.98	0.94	51
accuracy			0.91	64

test_size: 0.2

k: 3	precision	recall	f1-score	support
no da examen	0.86	0.71	0.77	17
da examen	0.96	0.98	0.97	110
accuracy			0.94	127

k: 4	precision	recall	f1-score	support
no da examen	0.80	0.71	0.75	17
da examen	0.96	0.97	0.96	110
accuracy			0.94	127

k: 5	precision	recall	f1-score	support
no da examen	0.92	0.65	0.76	17
da examen	0.95	0.99	0.97	110
accuracy			0.94	127

test_size: 0.3

k: 3	precision	recall	f1-score	support
no da examen	0.91	0.78	0.84	27
da examen	0.96	0.99	0.98	163
accuracy			0.96	190

k: 4	precision	recall	f1-score	support
no da examen	0.84	0.78	0.81	27
da examen	0.96	0.98	0.97	163
accuracy			0.96	190

k: 5	precision	recall	f1-score	support
no da examen	0.95	0.74	0.83	27
da examen	0.96	0.99	0.98	163
accuracy			0.96	190

5. Support Vector Machine (SVM) con estrategia One-vs-All:

    - Configuración: Utilización de SVM con la estrategia One-vs-All para problemas de clasificación multiclase.
    - Métricas evaluadas: Precisión, recall y F1-score.

    Para test_size: 0.1

	precision	recall	f1-score	support
no da examen	1.00	0.00	0.00	13
da examen	0.80	1.00	0.89	51
accuracy			0.80	64

    Para test_size: 0.2

	precision	recall	f1-score	support
no da examen	1.00	0.00	0.00	17
da examen	0.87	1.00	0.93	110
accuracy			0.87	190

    Para test_size: 0.3

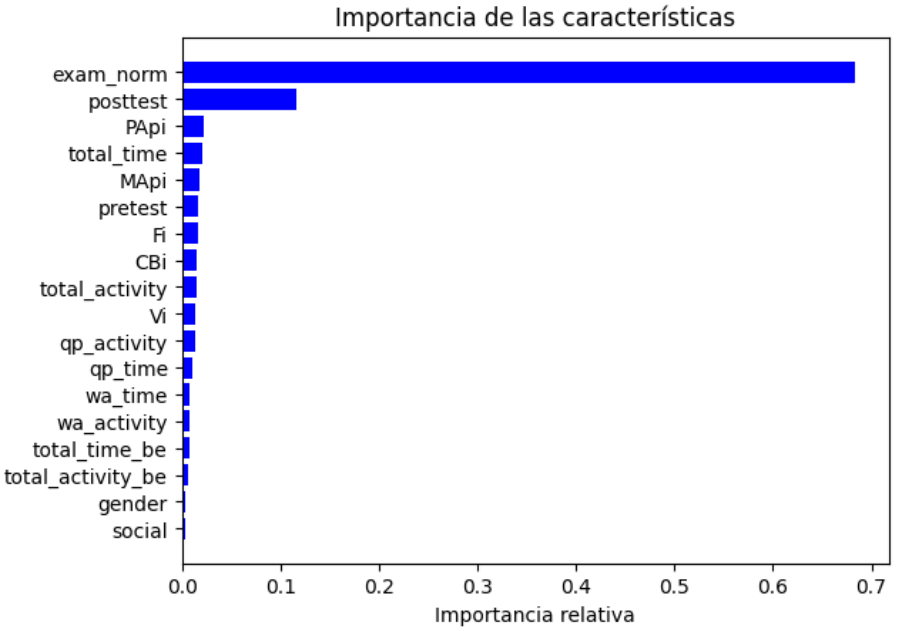
	precision	recall	f1-score	support
no da examen	1.00	0.00	0.00	27
da examen	0.86	1.00	0.92	163
accuracy			0.86	190

6. Random Forest:
    - Configuración: Implementación del algoritmo Random Forest para clasificación.
    - Métricas evaluadas: Precisión, recall y F1-score.

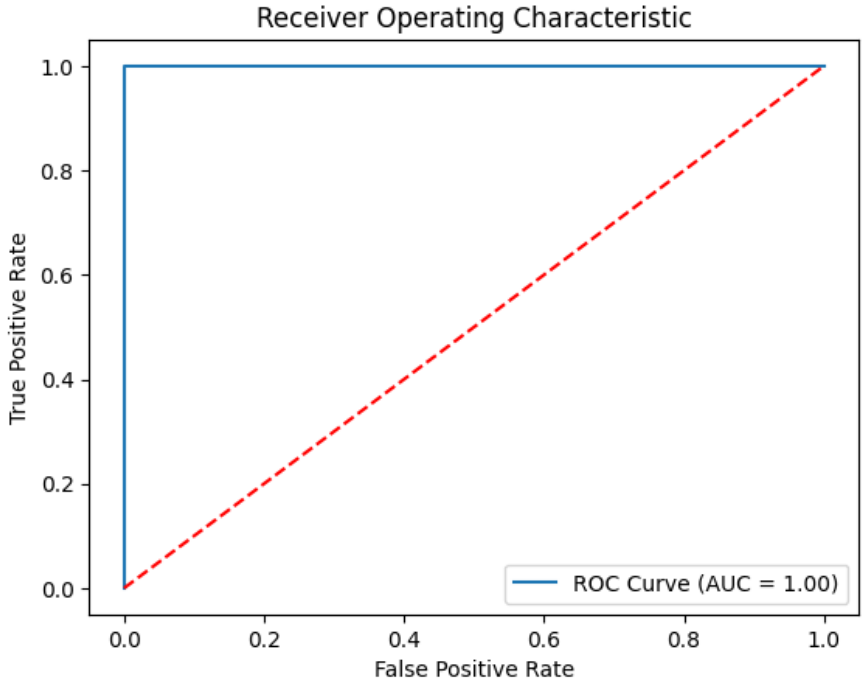
    Para test_size: 0.1, 0.2 y 0.3 el resultado fue:

	precision	recall	f1-score
no da examen	1.00	1.00	1.00
da examen	1.00	1.00	1.00
accuracy			1.00

Importancia de características en un bosque de árboles:
La importancia de las características se determina al observar cómo cada característica contribuye al rendimiento predictivo del modelo.



Curva ROC:



Conclusiones:

Los experimentos llevados a cabo revelaron que todos los algoritmos evaluados mostraron un desempeño aceptable en términos de precisión, recall y F1-score. Los árboles de decisión se destacaron como herramientas útiles para comprender los procesos de decisión, brindando explicabilidad y visualización de la jerarquía de decisiones. Sin embargo, se observó que la efectividad de los criterios de decisión, como Gini o Entropía, puede variar dependiendo de los datos utilizados.

Además, se observó que los modelos de aprendizaje supervisado obtuvieron mejores resultados a medida que aumentaba el tamaño del conjunto de prueba. No obstante, es importante tener en cuenta que el tamaño del conjunto de prueba debe mantenerse dentro de ciertos umbrales establecidos en la literatura.

En particular, el algoritmo K-Nearest Neighbors (KNN) fue efectivo en este caso, mostrando mejores resultados a medida que aumentaba el número de vecinos. Además, no presentó problemas significativos de tiempo de inferencia debido al pequeño tamaño del conjunto de datos.

Por otro lado, el modelo Random Forest se destacó como el que obtuvo los mejores resultados, como se evidencia en su curva ROC y en la importancia asignada a las características.

En conclusión, la elección del modelo de aprendizaje supervisado dependerá de la interpretabilidad deseada, las características de los datos y la disponibilidad de tiempo y recursos. Es importante considerar estos factores al seleccionar el modelo más adecuado para un problema específico.