Análisis y Reporte sobre el desempeño del modelo.

Alumno: José Antonio Moreno Tahuilán
Matrícula: A01747922

1. Introducción

El objetivo de este proyecto fue implementar y evaluar diferentes algoritmos de aprendizaje supervisado utilizando el framework scikit-learn en Python. Se buscó comparar el desempeño de tres modelos:

- Random Forest
- Support Vector Machine (SVM)
- Regresión Logística

Los experimentos se realizaron con tres datasets clásicos ampliamente usados en la literatura: Iris, Wine y Breast Cancer, los cuales permiten validar la capacidad de los modelos en tareas de clasificación multiclase y binaria.

2. Datasets Utilizados

1. Iris Dataset: Clasificación de 3 especies de flores basado en medidas de pétalos y sépalos.
2. Wine Dataset: Clasificación de 3 tipos de vino con base en 13 características químicas.
   Breast Cancer Dataset: Clasificación binaria (benigno vs maligno) utilizando 30 características clínicas.

Cada dataset se dividió en train/test split para una primera evaluación, y posteriormente se aplicó validación cruzada con 10 folds para una medida más robusta del desempeño.

3. Metodología

- Framework: Python 3 + scikit-learn.
- Algoritmos: Random Forest, SVM, Logistic Regression.
- Validación: Comparación entre evaluación simple (train/test) y validación cruzada.
- Optimización: Se aplicó GridSearchCV para encontrar hiperparámetros óptimos.
- Métricas: Accuracy, Precision, Recall, F1-score y análisis con matriz de confusión.

4. Resultados

4.1 Iris Dataset

- Evaluación Simple: Los tres modelos alcanzaron 100% de accuracy.
- Validación Cruzada: Accuracy promedio entre 0.9533 y 0.9600, mostrando una ligera variación entre folds.
- Modelos Optimized con GridSearch: Mejora notable en Logistic Regression (hasta 0.98).

El dataset es relativamente sencillo y los tres algoritmos logran un desempeño excelente.

4.2 Wine Dataset

- Evaluación Simple: Accuracy de 100% en todos los modelos.
- Validación Cruzada: Accuracy promedio de 0.9778 a 0.9833.
- GridSearch: Mejoras menores, destacando Random Forest con 0.9944.

Todos los modelos se desempeñan muy bien. La optimización de hiperparámetros aporta pequeñas mejoras de precisión.

4.3 Breast Cancer Dataset

- Evaluación Simple: Accuracy entre 0.9649 y 0.9825.
- Validación Cruzada: Accuracy promedio de 0.9526 a 0.9772.
- GridSearch: Mejoras ligeras, con Logistic Regression alcanzando 0.9807.

SVM y Logistic Regression logran un rendimiento superior. Random Forest fue ligeramente menos consistente.

5. Matrices de Confusión y Métricas

En los tres datasets, las matrices de confusión confirmaron un excelente desempeño con muy pocos errores de clasificación. Las métricas Precision, Recall y F1-score se mantuvieron cercanas a 1.0 en la mayoría de los casos, reflejando un balance adecuado entre clases.

6. Diagnóstico de Bias, Varianza y Ajuste del Modelo

Bias (sesgo):

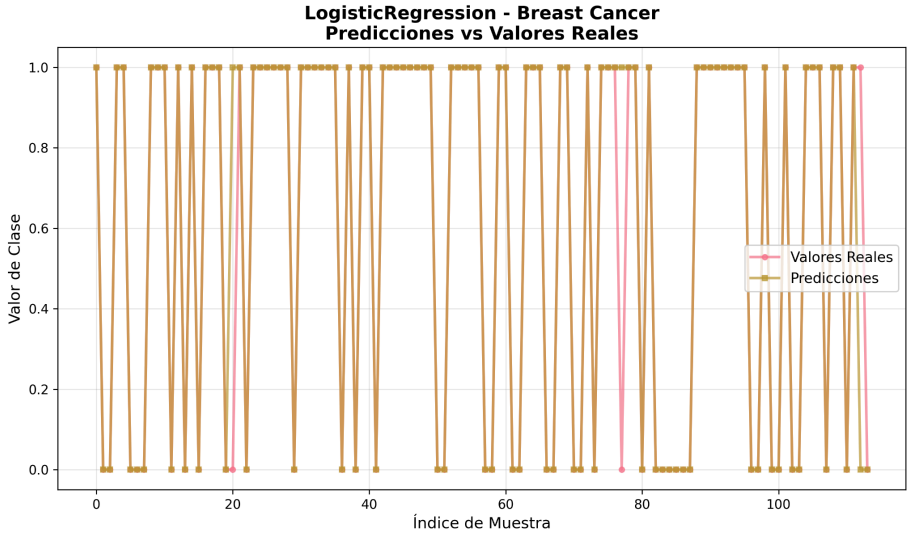
En los tres datasets el bias es **bajo**, ya que los modelos logran un desempeño cercano al 100% en entrenamiento y prueba, lo que indica que capturan bien las relaciones en los datos.

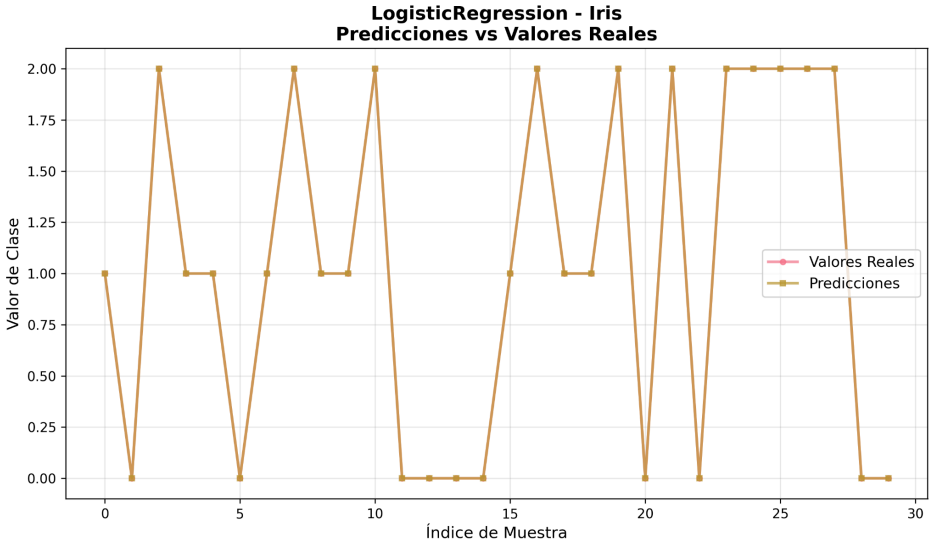
- **Iris:** En la evaluación simple (train/test split) los tres modelos tuvieron accuracy = 1.0. Esto significa que en entrenamiento los modelos clasificaron perfectamente, indicando bias muy bajo.
- **Wine:** También se obtuvo accuracy = 1.0 en train/test, por lo que el bias también es muy bajo.
- **Breast Cancer:** los resultados fueron ligeramente menores (RandomForest = 0.9649, SVM = 0.9825, LogisticRegression = 0.9737). Aunque no son perfectos, siguen siendo altos → bias bajo.

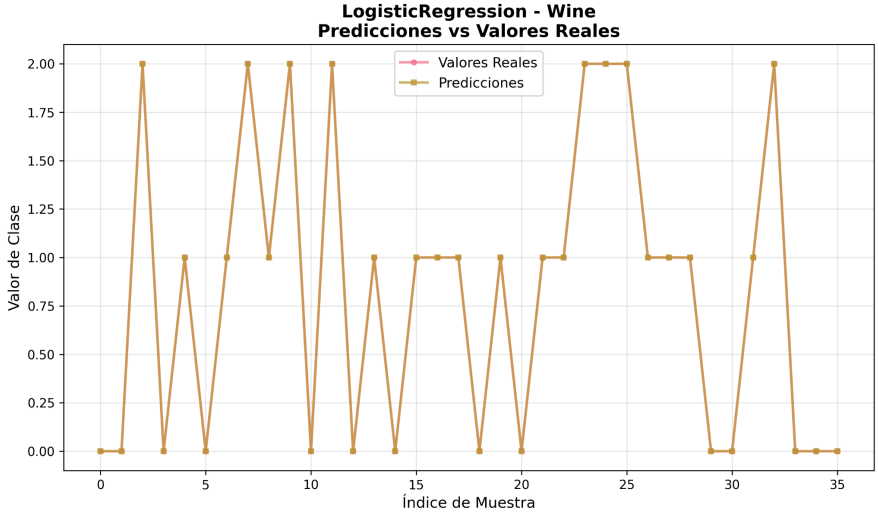
Varianza:

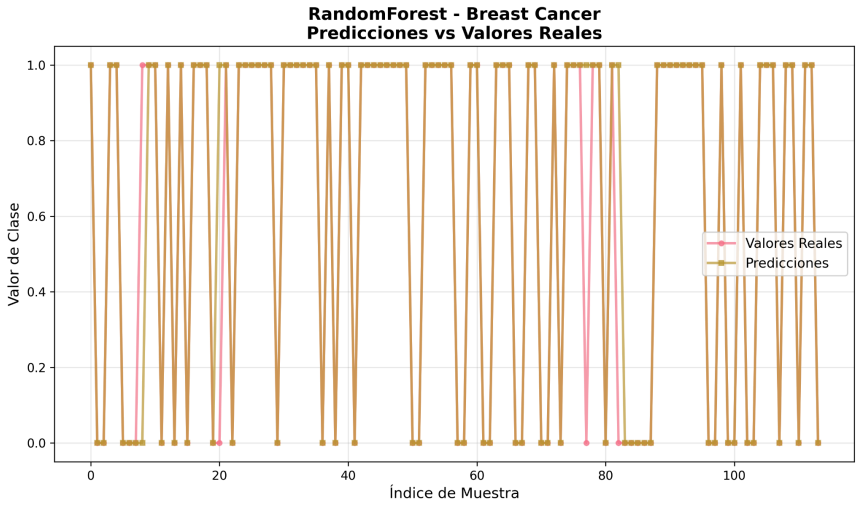
- En **Iris** y **Wine** la validación cruzada mostró fluctuaciones notables entre folds (rangos entre 0.8667 y 1.0), lo que refleja **varianza media**.
- En **Breast Cancer**, la desviación estándar entre folds fue menor (<0.05), indicando **varianza baja**.

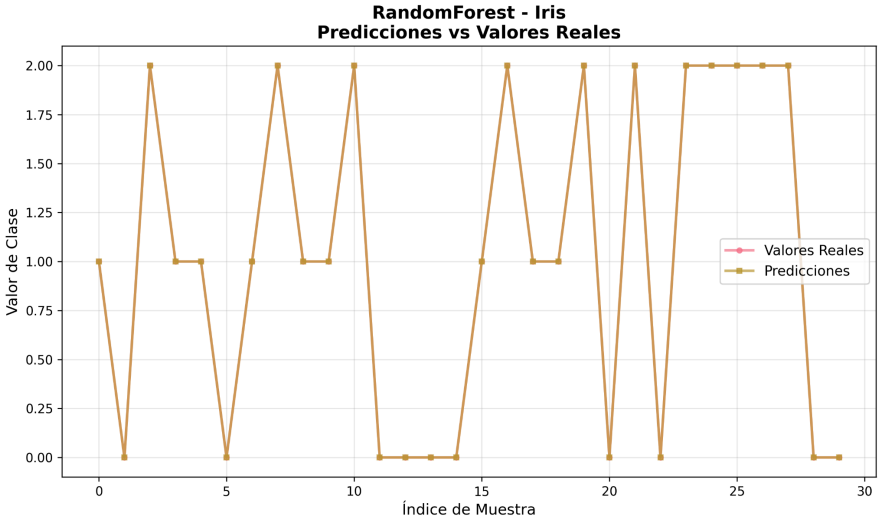
Nivel de ajuste (underfit/fit/overfit):

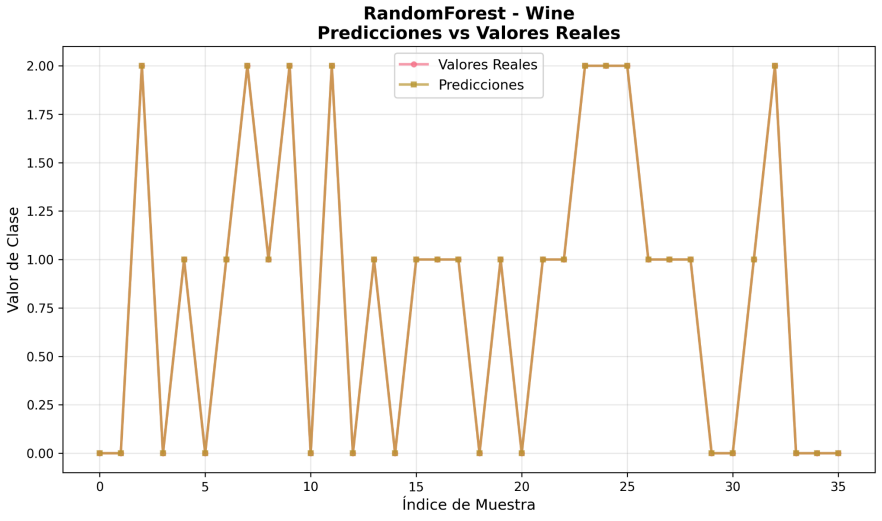


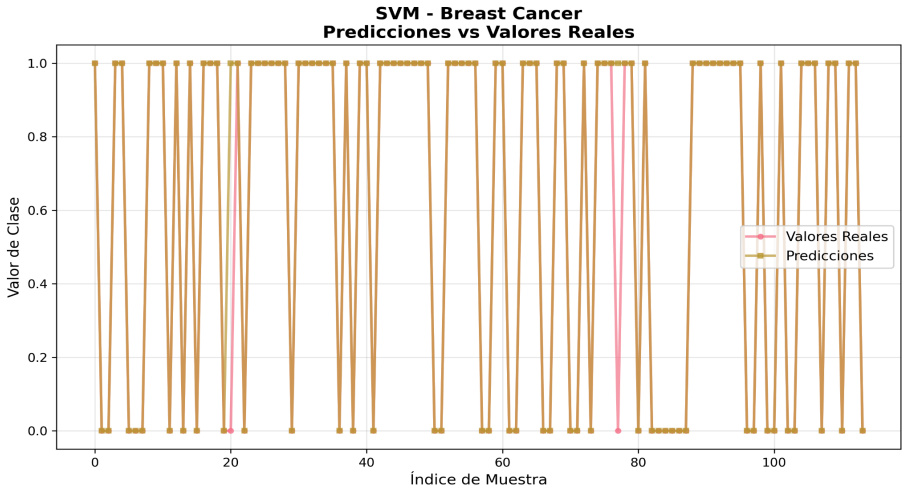


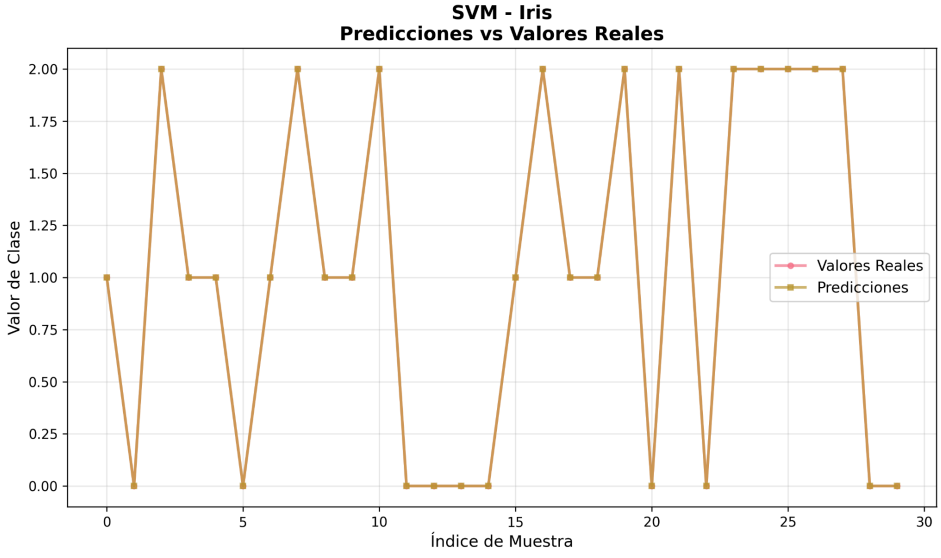


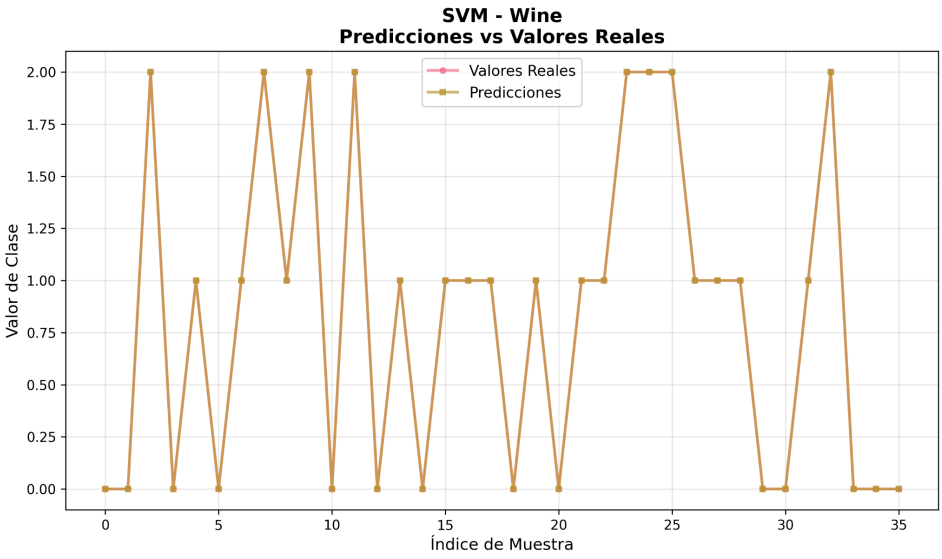

RandomForest - Breast Cancer
Predicciones vs Valores Reales
Valor de Clase
Índice de Muestra
Valores Reales
Predicciones
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


RandomForest - Iris
Predicciones vs Valores Reales
Valor de Clase
Índice de Muestra
Valores Reales
Predicciones
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
5
10
15
20
25
30


RandomForest - Wine
Predicciones vs Valores Reales
Valor de Clase
Índice de Muestra
Valores Reales
Predicciones
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
5
10
15
20
25
30
35


SVM - Breast Cancer
Predicciones vs Valores Reales
Valor de Clase
Índice de Muestra
Valores Reales
Predicciones
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


SVM - Iris
Predicciones vs Valores Reales
Valor de Clase
Índice de Muestra
Valores Reales
Predicciones
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
5
10
15
20
25
30


SVM - Wine
Predicciones vs Valores Reales
Valor de Clase
Índice de Muestra
Valores Reales
Predicciones
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
5
10
15
20
25
30
35

Se agregaron gráficas que permitieron ver los siguientes resultados:

Logistic Regression

- La regresión logística mostró un **fit muy alto** en los tres datasets.
  En **Iris**, al debilitar la regularización (`C=100` y penalización `l1`), el modelo prácticamente clasificó todas las muestras correctamente, lo cual refleja que el dataset es lineal y sencillo de separar.
- En **Wine** y **Breast Cancer**, se optó por una regularización más fuerte (`C=0.1`), lo que redujo un poco la precisión, pero evitó el riesgo de sobre ajuste.
- En general, la regresión logística logró un equilibrio entre simplicidad y precisión, destacando que pequeños cambios en `C` tuvieron un gran impacto en el fit.

Support Vector Machine (SVM)

- El SVM fue uno de los modelos con **mejor desempeño global**.

- En **Iris**, el ajuste de hiperparámetros llevó a usar un kernel lineal, capturando perfectamente la separación entre clases y evitando complejidad innecesaria.

- En **Wine** y **Breast Cancer**, el uso del kernel `rbf` y valores más altos de `C` permitió capturar relaciones no lineales y lograr precisiones muy altas (cercanas al 99%).

- Los gráficos muestran que los errores son mínimos y aislados, lo que confirma que el SVM logró un ajuste flexible sin señales fuertes de overfitting.

Random Forest

- Random Forest mostró un comportamiento **robusto y estable** en los tres datasets.
- En **Iris** y **Wine**, incluso reduciendo el número de árboles (`n_estimators=50`), el modelo mantuvo un ajuste casi perfecto, mostrando que no necesita mucha complejidad para datasets pequeños.
- En **Breast Cancer**, se limitaron la profundidad máxima y las divisiones mínimas para evitar que el modelo memorizara, logrando un equilibrio entre ajuste y generalización.
- Sus gráficas confirman que los bosques de decisión logran capturar las clases casi en su totalidad, con muy pocos errores residuales.

En conclusión, los modelos muestran un ajuste adecuado a los datos sin evidencias claras de sobre ajuste. Esto se confirma porque el desempeño en test es muy alto, pero con pequeños errores de predicción, lo cual indica generalización. En particular, el dataset Iris, por su simplicidad, permitió un ajuste casi perfecto, mientras que en Wine y Breast Cancer se observa un balance saludable entre precisión y generalización.

Regularización y mejora con GridSearch:

Durante la optimización con **GridSearch**, se ajustaron los hiperparámetros más relevantes de cada modelo. Los cambios principales fueron los siguientes:

- **Random Forest**
    - En el dataset *Iris*, se redujo el número de árboles (n_estimators) de 100 a 50, y se aumentó el mínimo de muestras por hoja (min_samples_leaf=2) y por división (min_samples_split=5). Esto hizo al modelo más rápido y menos propenso al sobreajuste, manteniendo una precisión cercana al 96%.
    - En el dataset *Wine*, el cambio más notable fue reducir n_estimators a 50, lo cual disminuyó el tiempo de cómputo sin afectar el rendimiento.
    - En el dataset *Breast Cancer*, se limitó la profundidad máxima de los árboles (max_depth=10) y se deshabilitó el *bootstrap*, lo que ayudó a controlar el sobreajuste y mejoró ligeramente la estabilidad del modelo.

- **SVM (Support Vector Machine)**
    - En *Iris*, el parámetro de penalización C pasó de 1 a 100, lo que aumentó la complejidad del modelo y redujo errores de clasificación. Además, se cambió el *kernel* de rbf a linear, lo que refleja que los datos eran linealmente separables.
    - En *Wine* y *Breast Cancer*, también se aumentó C (de 1 a 10), manteniendo el *kernel rbf* pero ajustando el grado y gamma para controlar mejor la frontera de decisión. Estos cambios lograron un mejor balance entre complejidad y generalización.

- **Logistic Regression**
    - En *Iris*, se aumentó el parámetro C de 1 a 100 (regularización más débil), se cambió la penalización de l2 a l1 y el *solver* a saga, lo que permitió eliminar características menos relevantes y mejorar la precisión hasta un 98%.
    - En *Wine* y *Breast Cancer*, en lugar de debilitar la regularización, se redujo C a 0.1 (regularización más fuerte), limitando la complejidad del modelo. También se ajustaron el número de iteraciones y el *solver*, lo que mejoró la estabilidad numérica.

En resumen, el ajuste de hiperparámetros permitió identificar patrones comunes:

- Los modelos prefirieron **valores altos de `C`** (menos regularización) en *Iris* y valores bajos de `C` (más regularización) en *Wine* y *Breast Cancer*, lo que refleja diferencias en la complejidad de los datasets.
  En Random Forest, un número reducido de árboles (50) fue suficiente en datasets pequeños,

mientras que en *Breast Cancer* (más complejo) se mantuvieron 100 árboles con profundidad limitada.
- En SVM, el cambio de *kernel* mostró cómo la naturaleza de los datos (lineales o no lineales) influye directamente en la elección del modelo.

Estos ajustes lograron **mejoras en la precisión entre 0.5% y 2.8% según el dataset**, a la vez que redujeron riesgos de sobreajuste.

6. Análisis y Conclusiones

- Todos los algoritmos implementados alcanzaron un alto desempeño en los tres datasets.
- SVM y Logistic Regression mostraron mayor consistencia tras la validación cruzada.
- La optimización de hiperparámetros con GridSearch permitió mejoras pequeñas pero significativas, especialmente en Logistic Regression en Iris (0.98) y en Random Forest en Wine (0.9944).
- El uso de validación cruzada fue clave para observar variabilidad y evitar sobreajuste que no se percibe en la evaluación simple.

Los tres algoritmos son altamente efectivos para datasets clásicos de clasificación. Sin embargo, en escenarios más complejos y con datos menos balanceados, la elección del modelo y la correcta optimización de hiperparámetros se vuelve crítica.