UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ALESSANDRA ROBLES GOMEZ
321164315

Análisis del mejor modelo de decisión para la base “cancer.scv”.

En el presente documento se comparan distintos modelos de clasificación utilizados para la predicción a partir de patrones y relaciones para las clases y las etiquetas en la base de datos “cancer.scv”, con el objetivo de identificar cuál es el mejor árbol de decisión en función de varias métricas clave de desempeño. Se evaluaron los siguientes modelos: Random Forest, Regresión Logística, Árbol de Decisión, K-Nearest Neighbors (KNN) y Support Vector Machine (SVM). A continuación, se analiza cada modelo, sus métricas y, finalmente, se determina cuál es el modelo más adecuado para el problema.

NOTACIÓN:

Recall: El recall (sensibilidad o tasa de verdaderos positivos) indica qué tan bien el modelo captura todas los objetos que realmente pertenecen a la clase positiva.

F1-Score: El F1-score es una métrica que combina tanto la precisión como el recall en un solo valor.

Matriz de confusión: La matriz de confusión se presenta como una tabla de 2x2, que evalúa visualmente el rendimiento de un modelo como se muestra a continuación:

[[Verdaderos Positivos, Falsos Positivos] [Falsos Negativos, Verdaderos Negativos]]

1. RANDOM FOREST

   F1-Score: 0.9667
   Reporte de Clasificación:

	Precisión	Recall	F1-Score
Benigno	0.96	0.99	0.97
Maligno	0.98	0.93	0.96
Macro avg	0.97	0.96	0.97
Weighted avg	0.97	0.97	0.97

   Matriz de confusión: [[88, 1] [4, 57]]

   Random Forest es un modelo basado en un conjunto de árboles de decisión. En este caso, muestra una excelente capacidad para clasificar correctamente

tanto los casos negativos (Clase 0 Benigno) como los positivos (Clase 1 Maligno). Con una precisión de 0.98 para la Clase 1 y 0.96 para la Clase 0, y un recall superior a 0.90 para ambas clases, este modelo tiene una precisión bastante alta. Además, la matriz de confusión refleja un bajo número de falsos negativos (4) y falsos positivos (1), lo que indica que el modelo es eficiente para identificar tanto los casos de cáncer como los no cancerosos.

2. REGRESIÓN LINEAL

F1-Score: 0.9467
Reporte de Clasificación:

	Precisión	Recall	F1-Score
Benigno	0.94	0.98	0.96
Maligno	0.96	0.90	0.93
Macro avg	0.95	0.94	0.94
Weighted avg	0.95	0.95	0.95

Matriz de confusión: [[87, 2] [6, 55]]

La regresión logística es un modelo lineal basado en iteraciones que también muestra un desempeño sólido, con un F1-Score de 0.9467. Aunque su precisión es muy alta (0.96 para la Clase 1 y 0.94 para la Clase 0), su recall para la Clase 1 es algo más bajo (0.90), lo que significa que identifica correctamente un poco menos de casos positivos en comparación con el modelo Random Forest. Además, tiene una cantidad mayor de falsos negativos (6), lo que puede ser horrible cuando se trata de una enfermedad tan grave como el cáncer.

3. ÁRBOL DE DECISIÓN

F1-Score: 0.9333
Reporte de Clasificación:

	Precisión	Recall	F1-Score
Benigno	0.93	0.96	0.94
Maligno	0.93	0.90	0.92
Macro avg	0.93	0.93	0.93
Weighted avg	0.93	0.93	0.93

Matriz de confusión: [[85, 4] [6, 55]]

El árbol de decisión, es un modelo de clasificación que divide los datos en función de características particulares. Aunque bastante efectivo, tiene un rendimiento ligeramente inferior al de Random Forest y Regresión Logística. Su F1-Score es 0.9333, lo que indica que tiene una menor capacidad para manejar tanto la precisión como el recall de manera equilibrada en comparación con los dos modelos anteriores. A pesar de tener un buen desempeño en términos de precisión, la cantidad de falsos positivos y falsos negativos es un poco más alta que la de Random Forest y Regresión Logística.

4. K-NEAREST NEIGHBOURS (KNN)

F1-Score: 0.7333
Reporte de Clasificación:

	Precisión	Recall	F1-Score
Benigno	0.71	0.94	0.81
Maligno	0.84	0.43	0.57
Macro avg	0.77	0.69	0.69
Weighted avg	0.76	0.73	0.71

Matriz de confusión: [[84, 5] [35, 26]]

KNN muestra el rendimiento más bajo entre los modelos evaluados hasta ahora. Aunque tiene una precisión decente para la Clase 1 (0.84), su recall es considerablemente bajo (0.43), lo que significa que identifica incorrectamente muchos de los casos positivos. Además, su F1-Score es el más bajo hasta ahora (0.7333), lo que indica que no logra un buen equilibrio entre la precisión y el recall, sin mencionar su número alarmante de falsos negativos (35). Esto sugiere que KNN no es el modelo ideal para esta tarea, a mí no me gustaría que un modelo se equivocase y me dijera que no tengo cáncer cuando si lo poseo.

5. SUPPORT VECTOR MACHINE (SVM)

F1-Score: 0.5733
Reporte de Clasificación:

	Precisión	Recall	F1-Score
Benigno	0.60	0.85	0.70
Maligno	0.43	0.16	0.24
Macro avg	0.52	0.51	0.47
Weighted avg	0.53	0.57	0.51

Matriz de confusión: [[76, 13] [51, 10]]

SVM tiene un rendimiento significativamente inferior en comparación con los otros modelos. Con un F1-Score de 0.5733, la precisión y recall son considerablemente más bajas, especialmente para la Clase 1 (con un recall de solo 0.16). Esto significa que SVM tiene una tasa extremadamente alta de falsos negativos, además del alto numero de falsos positivos (13) y falsos negativos (51), lo cual es muy preocupante, ya que podría dejar pasar muchos casos de cáncer sin ser detectados.

SVM es conocido por ser efectivo en problemas de clasificación complejos. Sin embargo, también puede ser sensible a la elección de los parámetros y la escala de los datos, lo que puede afectar su rendimiento. (que fue, muy posiblemente lo que pasó aquí. La cantidad muy limitada de datos no ayuda.)

Al analizar todos estos modelos, **Random Forest** emerge como el mejor modelo debido a su alto F1-score de 0.97, (al menos en cancer.scv) lo que indica un equilibrio excelente entre precisión y recall, especialmente en la clase positiva. Además, el recall de la clase negativa (0.99) es sobresaliente, lo que implica que el modelo tiene una capacidad excepcional para identificar instancias negativas sin dejar escapar muchas.

Si bien otros modelos, como la regresión logística y el árbol de decisión, tienen desempeños bastante buenos, el Random Forest sobresale por su capacidad de manejar tanto la detección de negativos como positivos de manera equilibrada y precisa. Por otro lado, el modelo SVM y el K-Nearest Neighbors muestran deficiencias graves en la detección de la clase positiva, lo que los hace menos adecuados para este tipo de tarea.