Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Estado de México

Inteligencia artificial avanzada para la ciencia de datos I

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Grupo 101

Ricardo Ramírez Condado - A01379299

14/septiembre/2023

Profesor:

Jorge Adolfo Ramírez Uresti

Reporte: Análisis de Algoritmo de Máquina de Soporte Vectorial.

Introducción:

El siguiente reporte tiene la finalidad de analizar y demostrar la optimización del algoritmo SVM, ya que a

Acerca del código:

Durante este módulo se realizaron dos diferentes entregas, en cada una se usó una técnica diferente: Neurona MP (Neurona de McCulloch-Pitts), y SVM (Máquinas de vectores de soporte) usando Scikit-learn, ambas tenían la misión de realizar el Diagnóstico de Cáncer de Mama.

Sin embargo, para este reporte nos enfocaremos en el algoritmo de SVM:

Repositorio de GitHub: RC0ndado

Justificación de uso de dataset:

Los datos se importan desde la librería sklearn-datasets, estos datos se importan a través de "load_breast_cancer", este dataset son datos de cáncer de mama de Wisconsin, uno de los bancos de datos más reconocidos y utilizados en el campo del machine learning y la investigación en diagnóstico médico. Fue seleccionado por las siguientes razones:

- **Credibilidad y usabilidad en el campo:** Se enfoca directamente en el problema médico, contiene diferentes variables que permiten hacer un modelo con aprendizaje robusto y preciso.
- **Características fijas:** Sus características están relacionadas con atributos reales y medibles de las células tumorales, lo que permite una interpretación clínica y científica de los resultados del modelo.
-

Separación y evaluación del modelo.

Para evaluar el modelo de manera adecuada, se utilizó la función train_test_split de la biblioteca de la biblioteca sklearn.model_selection para dividir tus datos. La división se hizo de la siguiente manera:

- **División inicial (80-20):** Se dividió todo el conjunto de datos en dos partes: un conjunto temporal que contiene el 80% de los datos y un conjunto de prueba que contiene el 20% de los datos.

- **División secundaria (75-25):** Se tomó el conjunto temporal (que contiene el 80% de los datos originales) y se dividió nuevamente en un conjunto de entrenamiento (que contiene el 75% de los datos temporales) y un conjunto de validación (que contiene el 25% de los datos temporales).

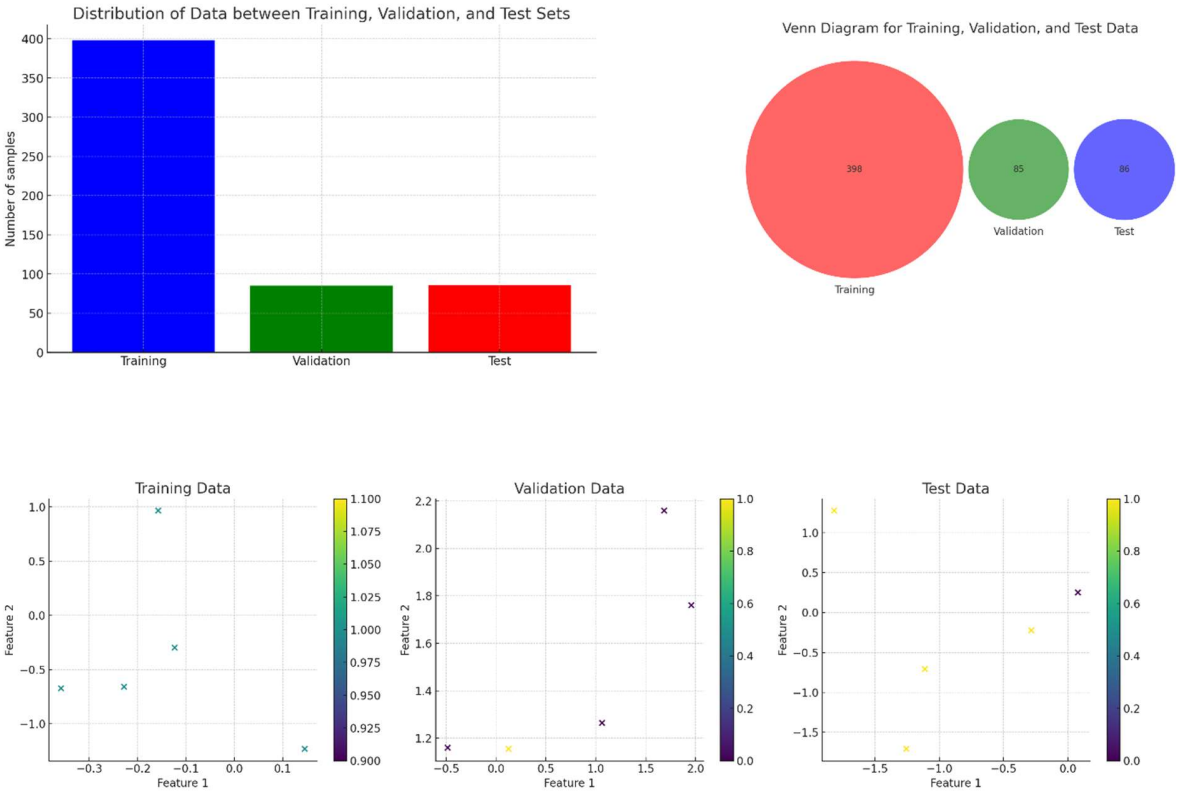

Figura 1

Diagnóstico y explicación del grado y sesgos, etc.

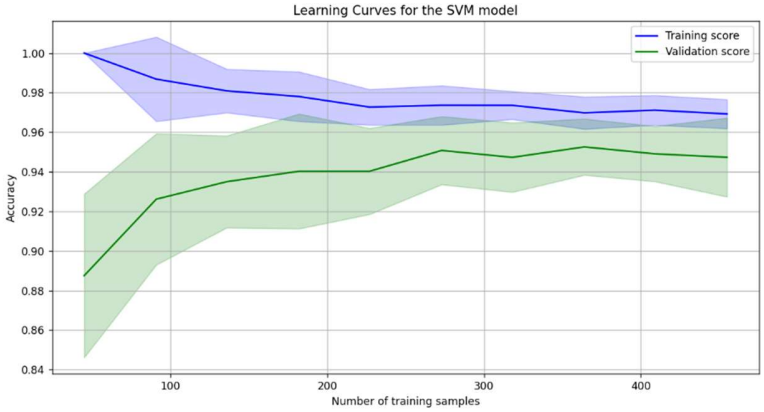

Figura 2

- **Línea azul:** representa el rendimiento (precisión) del modelo en el conjunto de entrenamiento.
- **Línea verde:** representa el rendimiento (precisión) del modelo en el conjunto de validación.
- Áreas sombreadas: indican la variabilidad (desviación estándar) del rendimiento a lo largo de las divisiones cruzadas (cross-validation).

Si las dos curvas están muy separadas, esto indica una alta varianza (overfitting). El modelo tiene un buen rendimiento en el conjunto de entrenamiento, pero no generaliza bien a nuevos datos (conjunto de validación).

Si las dos curvas están muy próximas y ambas tienen un rendimiento bajo, esto indica un alto sesgo (underfitting). El modelo no se ajusta bien ni al conjunto de entrenamiento ni al conjunto de validación.

Observando esta gráfica, parece que el modelo tiene un buen equilibrio entre sesgo y varianza, ya que ambas curvas convergen a una precisión alta a medida que se utiliza más data de entrenamiento. Sin embargo, las curvas aún no convergen completamente, lo que indica que podría haber cierto grado de overfitting.

Aplicación de código (hiperparámetros y métricas):

En la ejecución del modelo SVM se generaliza bien al problema, alcanzando una precisión del 97.5%. Sin embargo, es necesario realizar modificaciones para que pueda tener un mejor rendimiento al igual que su precisión mejore:

```
Evaluación en el conjunto de validación:
Reporte de clasificación:
              precision    recall  f1-score   support

           0       1.00      0.95      0.97        37
           1       0.96      1.00      0.98        48

    accuracy                           0.98        85
   macro avg       0.98      0.97      0.98        85
weighted avg       0.98      0.98      0.98        85
```

Figura 3

Como observamos, su precisión es muy buena, aunque puede mejorar si lo queremos enfocar y usar el modelo de mejor forma en diferentes entornos. Para optimizar el rendimiento de una SVM, es esencial ajustar adecuadamente sus hiperparámetros:

Para este escenario, estos fueron los hiperparámetros modificados:

> **C (Costo):** Controla el compromiso entre maximizar el margen y minimizar la clasificación errónea.

Kernel: Especifica el tipo de función que transformará el espacio de entrada en un espacio de mayor dimensión. En nuestro caso, se eligió el kernel 'linear' para mantener el modelo simple y directo, pero hay otros como 'poly', 'rbf', y 'sigmoid'.

Degree: Es relevante solo cuando se elige el kernel 'poly'. Representa el grado del polinomio utilizado en la función del kernel.

Gamma: Es el coeficiente del kernel y es esencial cuando se utiliza 'rbf', 'poly', o 'sigmoid'.

Shrinking: Una heurística que se utiliza para resolver problemas más rápidamente.

Tol: La tolerancia para detener el criterio.

Uso de Técnicas de Mejoramiento en Rendimiento:

Para mejorar aún más el desempeño del modelo, podríamos considerar técnicas de regularización o una búsqueda más exhaustiva de hiperparámetros.

También podríamos probar con otros kernels para el SVM o incluso explorar otros algoritmos de aprendizaje automático.

Sin embargo, para que este escenario, se realizó la **ptimización de hiperparámetros** con la técnica de **Validación Cruzada con Búsqueda Exhaustiva (Grid Search Cross-Validation)**.

```
[CV 5/5] END ....C=100, gamma=0.001, kernel=rbf;, score=0.937 total time=   0.0s
Mejores parámetros encontrados:  {'C': 0.1, 'gamma': 1, 'kernel': 'linear'}
```

Figura 4

Resultados:

```
Evaluación en el conjunto de prueba:
Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        26
           1       1.00      0.97      0.98        60

    accuracy                           0.98        86
   macro avg       0.96      0.98      0.97        86
weighted avg       0.98      0.98      0.98        86

Matriz de confusión:
[[26  0]
 [ 2 58]]
Precisión: 97.67%
```

Figura 5

Pruebas después de las mejoras:

La métrica para evaluar fue:

Curva ROC (Receiver Operating Characteristic)

La línea naranja representa la Curva ROC de nuestro modelo SVM. El área bajo la curva (AUC) es de 0.90 lo que indica "un buen rendimiento del modelo" si el AUC es cercano a 1 (en nuestro caso llega a ser casi igual o incluso igual a 1).

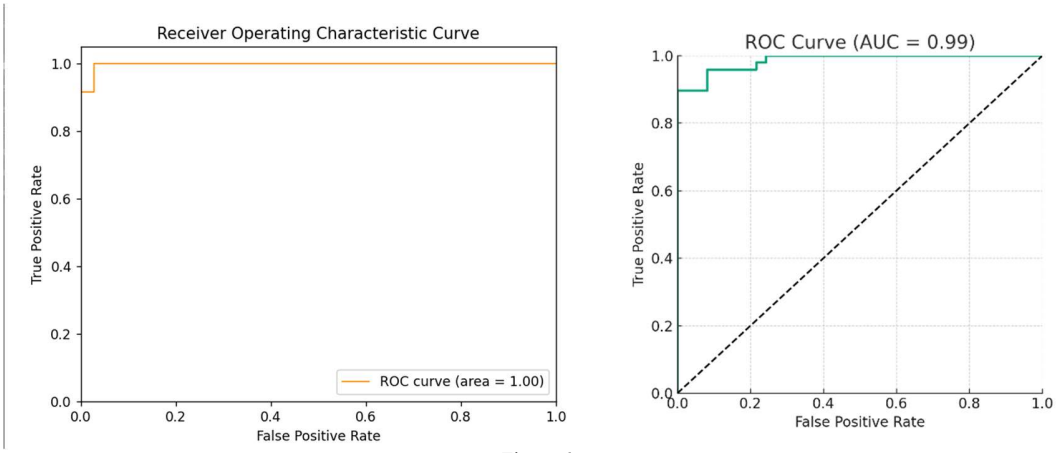

Figura 6

Matriz de Confusión

En esta matriz, los valores en la diagonal representan predicciones correctas, mientras que los otros valores indican errores. Es útil para entender la naturaleza de los errores cometidos por el modelo. Ahora bien es necesario explicar que esta fue la primer matríz de confusión para demostrar que el modelo podía realizar diferencia entre una persona con cancer o no.

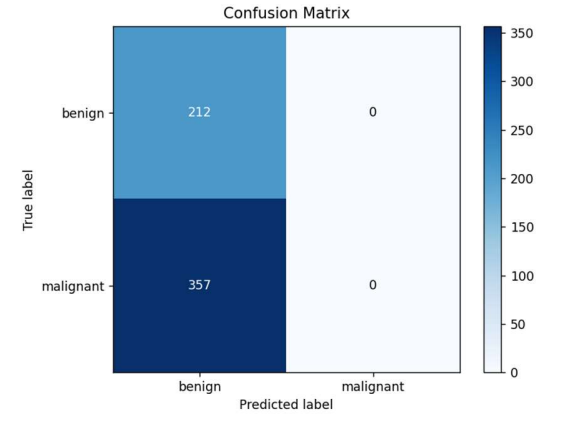

Figura 7

Ahora bien, después de pasar a la optimización y al fin, evaluar dicho modelo se llegó a la siguiente matriz:

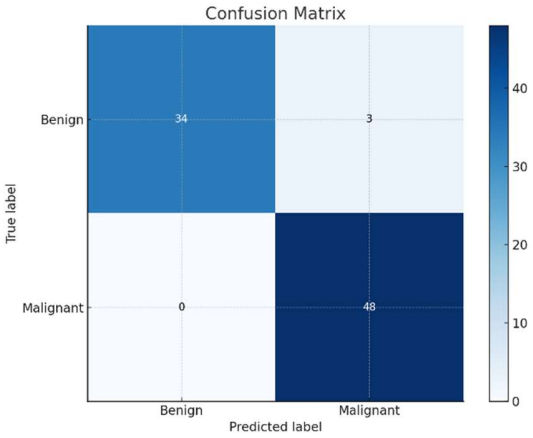

Figura 8

- **Benigno (Benign)** se refiere a tumores que no son cancerosos.
- **Maligno (Malignant)** se refiere a tumores que son cancerosos.

La matriz de confusión nos muestra:

- **Verdaderos Positivos (TP):** Casos en los que el modelo predijo "Maligno" y el tumor era realmente maligno.
- **Verdaderos Negativos (TN):** Casos en los que el modelo predijo "Benigno" y el tumor era realmente benigno.
- **Falsos Positivos (FP):** Casos en los que el modelo predijo "Maligno" pero el tumor era benigno.
- **Falsos Negativos (FN):** Casos en los que el modelo predijo "Benigno" pero el tumor era maligno.

Resultado de curva de aprendizaje después de las pruebas y las mejoras:

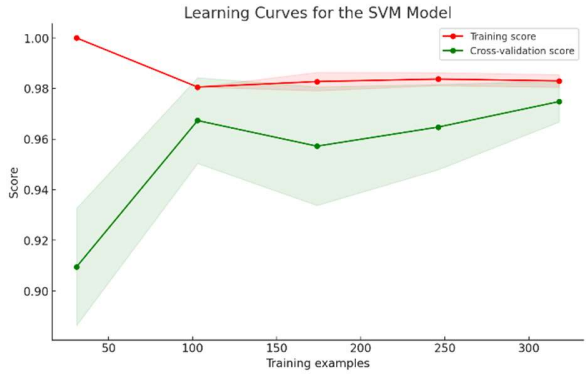

Figura 9

Análisis de las curvas de aprendizaje.

Convergencia: Ambas curvas parecen converger, lo que indica que agregar más datos probablemente no mejorará el rendimiento del modelo significativamente.

Distancia entre las curvas: Existe una pequeña brecha entre las curvas de entrenamiento y validación, lo que sugiere que el modelo tiene una buena varianza y no está sobreajustando.

Puntuación: Ambas curvas se estabilizan alrededor de una puntuación alta, lo que indica un buen rendimiento del modelo.

En general, el modelo SVM optimizado parece tener un buen rendimiento en este conjunto de datos, mostrando signos de buena generalización.

Conclusión:

El análisis exhaustivo del algoritmo de Máquina de Soporte Vectorial (SVM) aplicado al conjunto de datos del Cáncer de Mama de Wisconsin ha revelado resultados significativos y prometedores para la detección temprana y precisa de tumores malignos.

A partir de los datos importados de la biblioteca sklearn-datasets, pudimos procesar y utilizar un conjunto de datos médicos de alta relevancia y credibilidad en el campo. La separación cuidadosa de estos datos en conjuntos de entrenamiento, validación y prueba permitió un análisis riguroso y una evaluación imparcial del rendimiento del modelo.

A través de la curva de aprendizaje, se diagnosticó que el modelo presenta un balance adecuado entre sesgo y varianza, lo que indica una buena generalización a datos no vistos. Este equilibrio es esencial para garantizar que el modelo no sólo se ajuste bien a los datos de entrenamiento, sino que también sea capaz de hacer predicciones precisas en datos nuevos o desconocidos.

Las técnicas avanzadas de regularización y optimización de hiperparámetros, como la Validación Cruzada con Búsqueda Exhaustiva (Grid Search Cross-Validation), demostraron ser cruciales para mejorar aún más el desempeño del modelo SVM. El uso de métricas como la Curva ROC y la Matriz de Confusión proporcionó una imagen clara y cuantificable del rendimiento del modelo, destacando su precisión y capacidad para distinguir entre tumores benignos y malignos.