Forma

Descripción generada automáticamente con confianza baja

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

**Inteligencia artificial avanzada para la ciencia de datos I**

**Módulo 4**

**Práctica 1: Conceptos y componentes básicos**

**Grupo 101**

Ricardo Ramírez Condado - A01379299

**Profesor:**

Jorge Adolfo Ramírez Uresti

**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.

Repositorio de GitHub: [Link del repositorio] (Aquí deberías agregar el enlace al repositorio donde has alojado tu código.)

## **Justificación de uso de dataset:**

Los datos se importan desde la librería ”from sklearn.datasets import load\_breast\_cancer”, sin embargo, para detallarlo de mejor forma, este dataset en realidad son datos de cáncer de mama de Wisconsin, uno de los datasets 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:** Este dataset se enfoca directamente en el problema médico, contiene diferentes variables que permiten hacer un modelo con aprendizaje robusto y preciso.
* **Características fijas:** Las características en este dataset están relacionadas con atributos reales y medibles de las células tumorales, lo que permite una interpretación clínica y científica de los resultados del modelo.
* **Uso común:** Previo a presentar el problema, se buscó un dato de bancos por diferentes fuentes, como UCI Machine Learning, Kaggle, etc, sin embargo, el dataset más completo fue este ya que permitiía una sencilla y precisa estimación de la enfermedad.

**Aplicación de código (hiperparámetros y métricas):**

En la ejecución del modelo SVM se generaliza bien al problema, alcanzando una precisión de más del 97%. Sin embargo, es necesario ponernos en contexto y explicar un poco a detalle sobre el modelo:

Los SVM, son algoritmos de aprendizaje supervisado que buscan encontrar el hiperplano que mejor separa las clases en un espacio de características. Para optimizar el rendimiento de una SVM, es esencial ajustar adecuadamente sus hiperparámetros:

Para este escenario, estos fueron los hiperparámetros:

**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.

También se han presentado múltiples métricas de evaluación, incluyendo precisión, recall, f1-score, y una matriz de confusión. Estas métricas proporcionan una vista completa del desempeño del modelo en los conjuntos de validación y prueba.

**Técnicas de Mejoramiento en rendimiento:**

Texto

Descripción generada automáticamentePara 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 modelo logre un mejor rendimiento, se realizó una optimización de hiperparámetros, y para ello se usó la técnica de Validación Cruzada con Búsqueda Exhaustiva (Grid Search Cross-Validation)

Con base a lo mostrado se llegó a lo siguiente:



**Pruebas:**

La métrica para evaluar fueron:

**Curva ROC (Receiver Operating Characteristic)**

La Curva ROC es una representación gráfica que ilustra la capacidad discriminativa de un modelo de clasificación binaria a medida que varía el umbral de discriminación. El eje Y muestra la Tasa de Verdaderos Positivos, mientras que el eje X muestra la Tasa de Falsos Positivos.

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).

Gráfico

Descripción generada automáticamente

**Matriz de Confusión**

La matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

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.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Conclusión:**

Gracias a la implementación de buenos parámetros y de un buen análisis y entrenamiento logró que el rendimiento fuera óptimo y llegar a un excelente modelo.