**Proyecto de Machine Learning**

Comparación de Algoritmos de Clasificación: Regresión Logística, SVM, Naive Bayes y Random Forest utilizando Breast Cancer Wisconsin (Diagnostic) Dataset.

# 1. Introducción

El análisis y la clasificación de datos médicos, como los datos de cáncer, son fundamentales para el desarrollo de herramientas que puedan asistir en la toma de decisiones clínicas. En este contexto, los algoritmos de clasificación como la regresión logística, el Support Vector Machine (SVM), Naive Bayes y Random Forest han sido ampliamente utilizados por su capacidad de predecir la categoría a la que pertenece una observación, basándose en características conocidas.

Este proyecto se orienta a investigar y comparar el desempeño de estos cuatro algoritmos de clasificación utilizando un conjunto de datos de cáncer. La finalidad es determinar cuál de estos métodos ofrece la mejor precisión y robustez en la clasificación de pacientes con cáncer.

# 2. Objetivos

Objetivo General: Comparar el desempeño de los algoritmos de Regresión Logística, SVM, Naive Bayes y Random Forest en la clasificación de datos de cáncer.

Objetivos Específicos:

Implementar los cuatro algoritmos en el conjunto de datos proporcionado

Evaluar la precisión, recall, F1-score, AUC y ROC, entre otras métricas relevantes para cada algoritmo.

Realizar un análisis de los resultados obtenidos para determinar las fortalezas y debilidades de cada método en este contexto específico.

# 3. Metodología

## 3.1 Conjunto de Datos

El conjunto de datos utilizado para este estudio es el Breast Cancer Wisconsin (Diagnostic) dataset, disponible en la librería scikit-learn, se recomienda descargarlo de https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. Este dataset contiene características computadas de imágenes digitalizadas de células cancerosas, que se clasifican en dos categorías: malignas y benignas. Para mas detalle ver la descripción del dataset en la url proporcionada.

## 3.2 Implementación de Algoritmos

**Regresión Logística:** Se implementará utilizando la librería scikit-learn, configurando el solver adecuado y ajustando hiperparámetros como el parámetro de regularización $C$.

**Naive Bayes:** Se utilizará el clasificador Gaussian Naive Bayes, adecuado para este tipo de datos continuos.

**Random Forest:** Se implementará un modelo Random Forest, ajustando hiperparámetros como el número de árboles en el bosque (n_estimators) y la profundidad máxima de los árboles (max_depth).

**Support Vector Machine (SVM):** Se implementará un modelo SVM utilizando un kernel adecuado (lineal, RBF, etc.) dependiendo de la distribución de los datos, y se realizará una búsqueda de hiperparámetros para optimizar el modelo. Se debe implementar el algoritmo probando con 4 kernels.

**(Bonus)** Se otorgará puntaje adicional si incluye en su estudio un modelo basado en arboles de decisión en todo el informe.

3.3 Evaluación de Modelos

Cada modelo será evaluado utilizando técnicas de validación cruzada y las siguientes métricas de desempeño:

Precisión: Proporción de verdaderos positivos sobre el total de predicciones positivas.

Recall: Proporción de verdaderos positivos sobre el total de verdaderos positivos y falsos negativos.

F1-Score: Media armónica de precisión y recall.

Matriz de confusión: Para analizar los errores de clasificación.

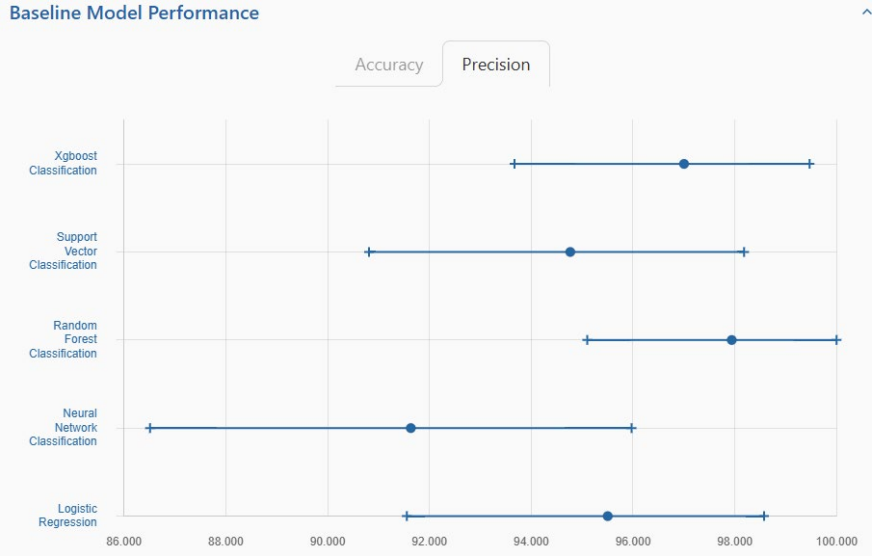
Curva ROC y AUC: Para evaluar la capacidad de discriminación de cada modelo.

Por cada modelo se debe indicar la justificación de cada uno de los hiperparametros utilizados.

3.4 Análisis Comparativo

Los resultados obtenidos de cada modelo se compararán para determinar cuál ofrece el mejor desempeño en términos de precisión y generalización en el conjunto de datos de cáncer. Se discutirá cómo las características del dataset pueden influir en la eficacia de cada algoritmo y se sugerirán posibles mejoras o adaptaciones.

En la url: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic podemos encontrar un reporte de algoritmos utilizados y pueden ser utilizados como línea base para poder comparar sus propuestas. Se presenta copia de esos resultados:




## 4. Resultados Esperados

Se espera encontrar diferencias significativas en el desempeño de los cuatro algoritmos en la clasificación del conjunto de datos de cáncer. Específicamente, se anticipa que:

Regresión Logística podría ofrecer un buen equilibrio entre precisión y facilidad de interpretación.

SVM podría sobresalir en términos de precisión, especialmente si se utiliza un kernel adecuado.

Naive Bayes podría mostrar un desempeño rápido pero menos preciso en comparación con los otros métodos, dependiendo de las características de los datos.

Random Forest podría destacar por su capacidad de manejo de datos con características complejas y correlacionadas, proporcionando un alto nivel de precisión y robustez.

## 5. Conclusiones

Este proyecto proporcionará una visión comparativa del desempeño de cuatro algoritmos de clasificación ampliamente utilizados en machine learning. Los hallazgos podrán guiar la selección del modelo más adecuado para la clasificación de datos médicos similares y podrían sugerir áreas de mejora en los modelos existentes.

## 6. Entregables

- Informe en pdf detallando los resultados obtenidos, contrastando con los resultados esperados, asimismo debe incluir las conclusiones a las que llegaron.

- Jupyter notebook con lo desarrollado, cada modelo debe estar debidamente documentado.