# Trabajo Práctico N° 3

**Análisis de Datos para la Predicción de Cáncer de Mama**

Este trabajo tiene como objetivo desarrollar un modelo de clasificación para un conjunto de datos de cáncer de mama, con el fin de predecir si una mujer puede tener cáncer. A continuación, se describen los pasos seguidos y los resultados obtenidos.

## Análisis

El conjunto de datos está compuesto por tres grupos de *features* y una variable objetivo (*target*):

* **Medias**
* **Errores**
* **Peores valores**
* **Target**: benigno o maligno

## Correlación

Se creó un modelo utilizando las *features* que mostraron mejor correlación con el *target*. Al analizar la correlación de cada una, se identificó que las más relevantes son:

* worst concave points
* worst perimeter
* worst area
* mean concave points
* mean perimeter
* mean area

## Test de Normalidad

Se realizó un test de normalidad utilizando *qqplot* para las *features* seleccionadas. Los resultados indicaron que ninguna de las variables se distribuye de manera normal, lo que descarta la posibilidad de utilizar modelos que suponen normalidad en los datos, como Naive Bayes o LDA.

## Outliers

Se creó una función llamada count\_outliers para identificar la cantidad de valores atípicos en cada variable. Al aplicar esta función, se detectaron *outliers* en todas las *features*. Para abordar esto, se utilizó un *RobustScaler* en el preprocesamiento de los datos.

## Modelos

Se consideraron los siguientes cinco modelos de clasificación:

* Regresión Logística
* KNN
* SVM
* Árbol de Decisión
* Random Forest

Los datos fueron preparados dividiéndolos en conjuntos de entrenamiento, prueba y validación, utilizando una partición *hold-out*, dejando un 20% de los datos para prueba. Después de entrenar los cinco modelos, se comparó su precisión:



El modelo con mejor desempeño fue la **Regresión Logística**.

## Ajuste de Hiperparámetros

Dado que la Regresión Logística fue el mejor modelo, se llevó a cabo una búsqueda de la mejor combinación de hiperparámetros utilizando *GridSearch*. La mejor combinación encontrada fue:

* **C: 10**
* **max\_iter: 100**
* **solver: newton-cg**

## Conclusión

En resumen, los principales hallazgos de este análisis son:

* Las *features* que mejor clasifican los datos son mean concave points, mean perimeter, mean area, worst concave points, worst perimeter y worst area.
* Ninguna de las variables sigue una distribución normal, lo que impide el uso de modelos como Naive Bayes y LDA.
* Los datos contienen *outliers*, por lo que se utilizó *RobustScaler* para escalarlos adecuadamente.
* El mejor modelo para la clasificación de los datos es la **Regresión Logística**.
* La mejor combinación de hiperparámetros para la Regresión Logística fue C: 10, max\_iter: 100 y solver: newton-cg.