PONTIFICIA UNIVERSIDAD JAVERIANA

JUAN CAMILO SARMIENTO PEÑUELA

ANDRÉS SEBASTIÁN SEGURA RODRÍGUEZ

FACULTAD DE INGENIERÍA

INGENIERÍA ELECTRÓNICA

INTELIGENCIA ARTIFICIAL

PROYECTO FINAL

DETECCIÓN CANCER DE MAMA

• INTRODUCCIÓN

El cáncer de mama es de las patologías más comunes en las mujeres. Este se puede detectar mediante anomalías físicas presentadas en el seno de la mujer mediante el auto examen y confirmados mediante patología después de confirmar mediante rayos x la detección de masas anómalas o de una pequeña partícula de calcio en el pecho de la mujer.
En este documento se presenta la clasificación de tumores entre benignos y malignos por medio de las propiedades obtenidas en el dataset del hospital de la universidad de Wisconsin del doctor William H. Wolberg, a partir de diferentes algoritmos de clasificación como la regresión logística, SVM, KNN y redes neuronales.

• OBJETIVOS

a. Objetivo general.

Predecir mediante las características del tumor caracterizado en pacientes, si hay o no cáncer de mama.

1. Objetivos específicos.

Utilizar los clasificadores de machine learning con el fin de realizar predicciones de situaciones de la vida real

- **DESCRIPCIÓN DEL PROYECTO:**

El cáncer de mama es de las patologías más comunes en las mujeres. Este se puede detectar mediante anomalías físicas presentadas en el seno de la mujer mediante el auto examen y confirmados mediante patología después de confirmar mediante rayos x la detección de masas anómalas o de una pequeña partícula de calcio en el pecho de la mujer.

En este proyecto se va a realizar la predicción de la patología de cáncer de seno mediante características físicas de los tumores detectados en pacientes. Estas características son:

2 Radio
3 Textura
4 Perímetro
5 Área
6 Suavidad

El dataset de estudio fue obtenido del hospital de la universidad de Wisconsin del doctor William H. Wolberg. **[1]**

- **DESARROLLO DEL TRABAJO.**

Se identifica que el problema a desarrollar es un problema de clasificación, por lo cual se va a implementar clasificadores de clasificación, específicamente binaria. Se escoge para este proyecto los clasificadores de regresión logística, máquinas de soporte vectorial (SVM), método de k-vecinos. Finalmente, se implementa una red neuronal de clasificación. El objetivo de realizar los clasificadores es verificar mediante medidas de desempeño de clasificadores como la matriz de confusión, el coeficiente de correlación de matthews, la exactitud y precisión. Lo mismo se realiza para evaluar la implementación en redes neuronales.

1. Preprocesamiento de los datos:

Los datos del dataset, como se mencionó anteriormente, son el radio del tumor, su textura, el perímetro, el área, la suavidad (smothness), y el diagnóstico. Estos datos utilizados tienen la ventaja que están completos. Además, que las etiquetas a considerar en este proyecto contienen datos de tipo numéricos, por lo cual para la implementación de los clasificadores y de la red neuronal no fue necesario realizar un preprocesamiento de los datos.

Implementación de los clasificadores:

Para todos los clasificadores utilizados, previamente se han dividido los datos para que los datos de entrenamiento sean el 70% del total de datos del dataset, y 30% para realizar la evaluación.

Regresión logística

Para la regresión logística, se normalizan los datos usando el standard scaler, y como resultado de la clasificación, se obtiene la matriz de confusión con los siguientes resultados:

	Predicted No	Predicted YES
Actual NO	50	8
Actual YES	5	108

El puntaje de precisión obtenido es del 93.1%, el de accurarcy de 92.4%, y el MCC de 0.8289

Se obtiene la curva de ROC mostrada en la figura 1:

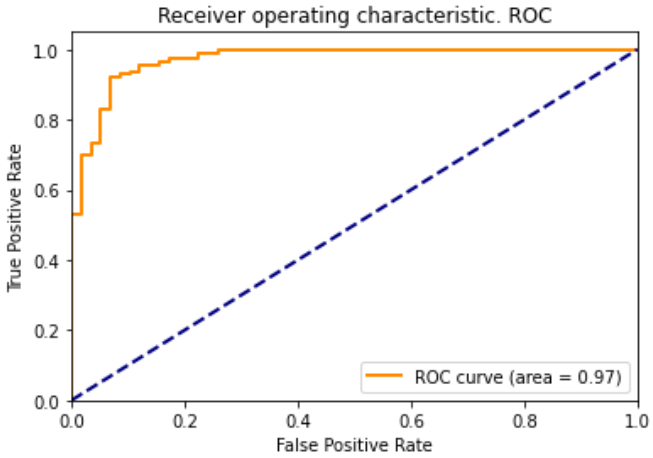


Figura1. ROC para la clasificación con regresión logística

Máquina de soporte vectorial

Para la SVM se implementa con el kernel lineal, con un C de 1, y el gamma='scale', el cual equivale al valor de gamma de 1/(n_features*X.var()). Nosotros en este caso, tendríamos 5 features.

Al entrenar y evaluar la clasificación con la máquina de soporte vectorial, se obtienen 2 puntajes de evaluación, el MCC y el accuracy, los cuales son de 88.89%, y de 0.7588 respectivamente.

La ROC obtenida por el clasificador por medio de SVM, se observa en la figura 2.

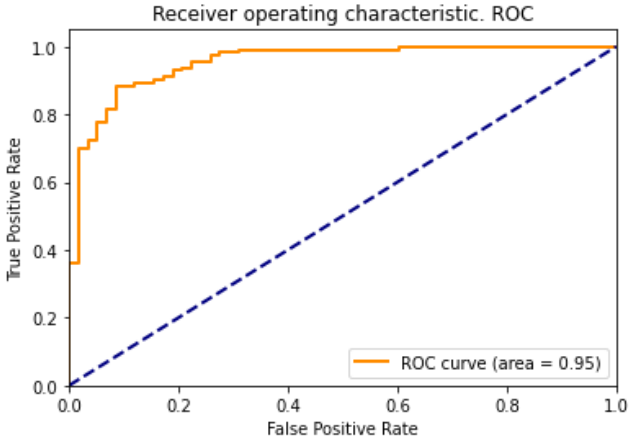


Figura 2. ROC para la clasificación con SVM

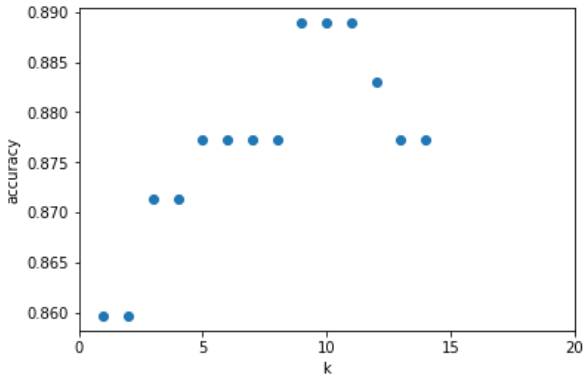
1. KNN o K-vecinos

Para la clasificación con K-vecinos, inicialmente se utiliza un k=15, con una metrica de distancia 'minkowski'. Al entrenar y evaluar, se obtiene un accuracy de 94% para el entrenamiento, y de 91% para el test.

Se obtiene entonces, la siguiente matriz de confusión:

	Predicted No	Predicted YES
Actual NO	51	8
Actual YES	7	105

Evaluando para diferentes k's en un rango de 1 a 15, se obtiene que el con k=9,k=10 y k=11, se obtiene un accuracy cercano a 89%:



Al evaluar con k=10, se obtiene un accuracy menor para el train, de 88%, pero se mantuvo para el test, que es del 91%, y la matriz de confusión sigue siendo parecida.

	Predicted No	Predicted YES
Actual NO	50	9
Actual YES	7	105

ANN

Se implementa una red neuronal de 3 capas internas (un total de 5 capas incluyendo las de entrad y de salida). Estas capas internas son de 10 neuronas cada una. Al entrenar y evaluar, se obtiene un accuracy de 91%

Se obtiene entonces, la siguiente matriz de confusión:

	Predicted No	Predicted YES
Actual NO	49	8
Actual YES	7	107

- **CONCLUSIONES.**

Se encontró que los clasificadores de mejor desempeño fueron el método de k-vecinos, con valores de accuracy del 91 %. Este desempeño fue muy parecido al de la red neuronal, la cual se esperaba la mejor clasificación.

La correlación de los datos del dataset, muestra la alta correlación en área, radio y perímetro, lo que se ve reflejado en el clasificador de maquina de soporte vectorial ya que en la superficie de decisión se iban a ver involucrados datos de otras etiquetas. Justificando el mas bajo desempeño de los clasificadores implementados

Link video: https://youtu.be/vB5DFWApY5E

Link repositorio: https://github.com/asegura1998/proyectoIA_grupo4

Referencias:

[1] https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset