Identificación de los caracteristicas mas influyentes en la predicción de cáncer de mama a partir de mediciones a muestras obtenidas mediante la técnica FNA

Leonardo Hernando Dallos Martínez
22971 – Inteligencia Artificial I – H1
Escuela de Ingeniería de Sistemas e Informática

Resumen

En este trabajo se encontrara un estudio hecho con el uso de herramientas de machine learning (ML) , el objetivo es identificar los factores mas influyentes en la predicción de cáncer de mama a partir de mediciones hechas a muestras de tejido sospechoso obtenido con la técnica FNA (biopsia por aspiración fina con aguja.
Los datos para el estudio se obtienen a partir de un dataset de Wisconsin en el cual se detallan características de mediciones para cada muestra como lo son el radio, perímetro, área, suavidad... Y una clase discriminadora para saber si es benigna o maligna.
Con los datos se prueban diferente modelos matemáticos para aprendizaje y predicción. El objetivo es seleccionar el mejor de los modelos y en base a el identificar los factores mas influyentes y así realizar una predicción mas acertada.

Introducción

El cáncer de mama es el segundo tipo de cáncer que mas afecta a mujeres en el mundo. Su temprana identificación es clave en la lucha por la mortandad que este genera, pues se ha visto que la tasa de supervivencia a 5 años de mujeres con cáncer de mama invasivo en fase 1 (masa solo en la mama) es del 90%, si este se ha diseminado a los ganglios linfáticos la tasa es del 85% y si se ha diseminado a otra parte mas lejana es del 27%; solo al 62% de las personas que tienen cáncer de mama se les diagnostica en fase 1. Para tener un temprano y confiable diagnostico se suelen hacer pequeñas biopsias como la FNA (biopsia por aspiración fina con aguja). El interpretar y establecer patrones correctamente en estas pruebas es de gran relevancia pues nos aportara información de algunas características que toma la aparición de este cáncer, para ello se debe hacer una relación entre los diversos y mas importantes aspectos medidos a partir de la prueba, y el uso de modelos estadísticos y de ML son de gran ayuda con este objetivo.

Proceso y método

Se quiere realizar un análisis predictivo a partir de un data set clasificado del diagnostico de cáncer de mama en Wisconsin. Con el uso de algunos modelos matemáticos de clasificación y también por aprendizaje por redes neuronales, los cuales me sirven para crear modelos predictivos a partir de unos datos mediante aprendizaje y testeo, se probaron los datos de estudio para luego seleccionar el mejor de estos modelos en estos datos y luego definir las características mas relevantes en el.

- Modelos usados: - Modelos de clasificación: GaussianNB, DecisionTreeClassifier, RandomForestClasifier, SVC.
  - Redes neuronales profundas (DNN).

- Análisis estadístico :
  - Preprocesamiento de datos.
  - Testeos con los modelos de clasificación con el uso de herramientas tales como: PCA, cross-validation
  - Testeos con DNN.
  - Selección del modelo con el mejor comportamiento.
  - Determinación de las características mas influyentes para el modelo seleccionado.

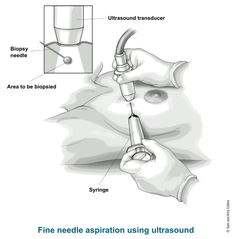

Figura 1. biopsia **FNA**

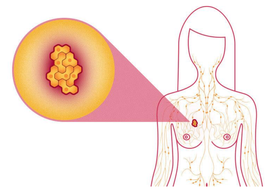

Figura 2. masa sospechosa cancer de mama

Resultados

El método con el que mejor se pueden modelar los datos para la creación de un modelo predictivo resulto ser el modelo clasificatorio RandomForest, este modelo mostro un mejor resultado en puntuación y un menor error en predicciones en comparación con los demás.

De acuerdo con el modelo que ha sido seleccionado se identificaron los factores mas influyentes en el y por lo tanto para nuestro estudio, los cuales fueron : 1. putos cóncavos, 2. perímetro, 3. área. 4. radio. 5. compactividad.

- En el preprocesamiento de datos se eliminaron algunos parámetros irrelevantes para el estudio.
- Para la realización de las pruebas con los distintos modelos se uso un 70 % de los datos para aprendizaje y el restante para el testeo.
- Para los casos en que se uso PCA se tomo la totalidad de los parámetros con el fin de tener una reducción de componentes a partir de todos los iniciales.

Modelo	Score		
	crosval	PCA	PCA y crossval
GaussianNB	0,905	0,918	0,896
DecisionTreeClasifier	0,896	0,877	0,901
RandomForest	0,914	0,953	0,912
SVC	0,898	0,959	0,877

Tabla 1. Modelos mátematicos - puntaje

Modelo	Score
DNN	0,906

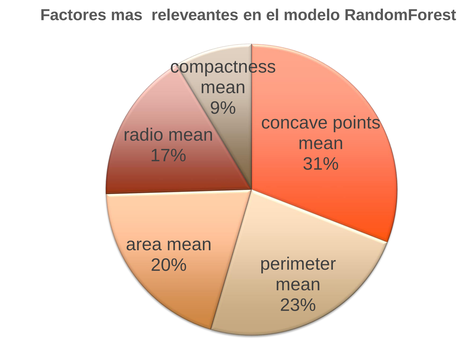

Gráfico 1. Relevancia de los factores.

Conclusiones

1. Para el testeo de los distintos modelos matemáticos se debe modificar correctamente el dataset para quitarle información que no aporte y no corresponda para el estudio y así tener un resultados mas acertados.

2. El algoritmo seleccionado RandomForest no es un algoritmo perfecto pero bien muestra un error bajo en sus predicciones.

3. El modelo de redes neuronales profundas (DNN) mostro un comportamiento similar al modelo de clasificación seleccionado (RandomForest) y bien podría también ser seleccionado como modelo para el análisis con un mayor estudio de su configuración para su uso.

4. Escalar la información en el modelo por red neuronal profunda permite generar una mejor versión del modelo, resultando en un menor error y mejor puntuación.

5. Las características mas relevantes identificadas (perímetro, área, radio, compactividad) según el modelo seleccionado vemos que son las mediciones mas notorias en una muestra FNA.

Trabajo Futuro

Realizar la comparación con mas modelos matemáticos de clasificación.

Realizar un estudio de regresión para identificar cuales serian las mediciones de los principales factores hallados en el estudio para cuando una muestra sea clasificada como maligna.

Información de contacto

Leonardo Hernando Dallos Martínez, Email: leo3,14@hotmail.com

Fabio Martínez Carrillo, Email: farmacar@saber.uis.edu.co

Referencias Bibliográficas

1. Breiman, L. Random Forest. Machie Learning 45, 5-32 (2001), recuperado de : https://doi.org/10.1023/A:1010933404324
2. Instituto Nacional de cancerologia, ESE., Grupo de vigilancia Epidemiologica del Cáncer. Cancer de mam, Colombia 2002-2006.
3. Dataset Recuperado en 30 de Julio de 2020, públicamente de https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
4. Arboles de decision y Random Forest classification. Recuperado de https://uc-r.github.io/random_forests
5. Caruana, Rich; Karampatziakis, Nikos; Yessenalina, Ainur (2008). *An empirical evaluation of supervised learning in high dimensions*