

# Metodología para la aplicación de la ciencia de datos en el diagnóstico del cáncer de mama

Presenta:

Esp. Jorge Armando Millán Gómez

Universidad Distrital "Francisco José de Caldas"  
Maestría en ciencias de la información y las comunicaciones

24 de mayo de 2023



# Elementos principales de la investigación

## Planteamiento del Problema

Según el informe de la organización mundial de la salud del año 2020 los casos detectados de cáncer de mama en Colombia fueron 15.509 de los cuales 4.411 casos terminaron en muerte ocupando el primer puesto de la tasa de letalidad sobre los demás tipos de cáncer[1]. Si no se tiene un diagnóstico a tiempo que detecte los aspectos más significativos que caracterizan el cáncer de mama es posible que la cifra de muertes en Colombia sea mayor en los años posteriores. En consecuencia, es necesario desarrollar una metodología que facilite la aplicación de la ciencia de datos en el diagnóstico de esta enfermedad.



# Elementos principales de la investigación

## Formulación del Problema

- ¿Una metodología aplicada a técnicas en ciencias de datos para el diagnóstico de cáncer de mama mejora y facilita el análisis de patrones característicos en cada individuo para encontrar errores en el diagnóstico?



## Planteamiento de la Hipótesis

- Una metodología para comparar técnicas y grandes cantidades de datos que contienen información de resultados diagnósticos de pacientes particulares con los datos característicos de pacientes que padecen de cáncer de mama, permite hallar la similitud del comportamiento de los datos y predice de manera correcta el padecimiento de este tipo de cáncer de los pacientes particulares e identifica las variables que más influyen para contraer dicha enfermedad.



# Objetivos

## Objetivo General

Desarrollar un sistema con modelos de Machine Learning para diagnosticar el padecimiento de cáncer de mama.

## Resultado

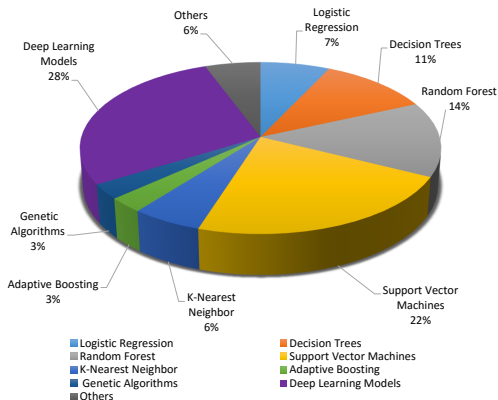
Desarrollo de una aplicación web para el diagnóstico del cáncer de mama donde se utilizaron modelos de Machine Learning seleccionados a partir de la exploración y comparación de los modelos más utilizados por diferentes investigadores y la precisión de dichos modelos en el diagnóstico de este tipo de Cáncer.



# Objetivos

## Objetivos Especificos

Efectuar el entrenamiento de los modelos en Machine Learning que apliquen para el diagnostico de cáncer de mama.



## Resultado

Para realizar el entrenamiento de los modelos de Machine Learning seleccionados, se utilizó el data set de cáncer de mama, elaborado por la Universidad de Wisconsin y se utilizó la librería Scikit-Learn para el procesamiento de datos y el posterior entrenamiento de dichos modelos.



# Objetivos

## Objetivos Especificos

Realizar un análisis comparativo de la precisión de los modelos existentes utilizados en el diagnostico de cáncer de mama.

## Resultado

La determinación de la precisión de los modelos es expresada como la relación entre los diagnósticos realizados correctamente y el número total de diagnósticos.

Modelo	Presición
Decision Trees	1.0
Random Forest	0.9953
Support Vector Machines(SVM)	0.9647
Logistic Regression	0.9600
Gaussian Naive Bayes	0.9507
K-Nearest Neighbor (KNN)	0.9413





# Objetivos

## Objetivos Especificos

Elaborar un aplicativo que sea capaz de diagnosticar el padecimiento de cáncer de mama de un paciente en particular.

## Resultado

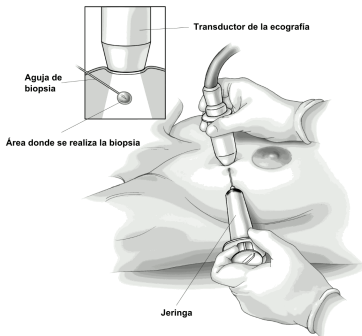
Implementación de la aplicación web llamada BreastApp V1.0 la cual está conformada por dos componentes: El Back-End de la aplicación el cual fue realizado en Python y el Front-End de la aplicación el cual fue realizado en Angular. Esta aplicación brinda un dictamen del padecimiento de cáncer de mama soportado en datos y gráficas proporcionados por los modelos de Machine Learning utilizados.



# Desarrollo de la Investigación

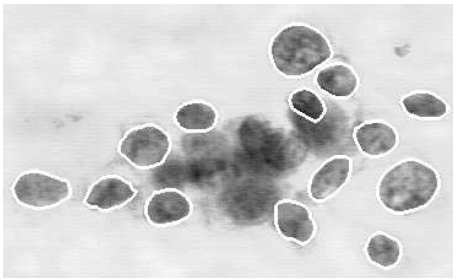
## Recolección de datos

En la investigación se identificaron nueve características evaluadas visualmente de una muestra obtenida por el método de aspiración por aguja fina (FNA) que se consideraron relevantes para el diagnóstico. El conjunto de datos resultante es conocido como el Data-Set de cáncer de mama de Wisconsin[?].



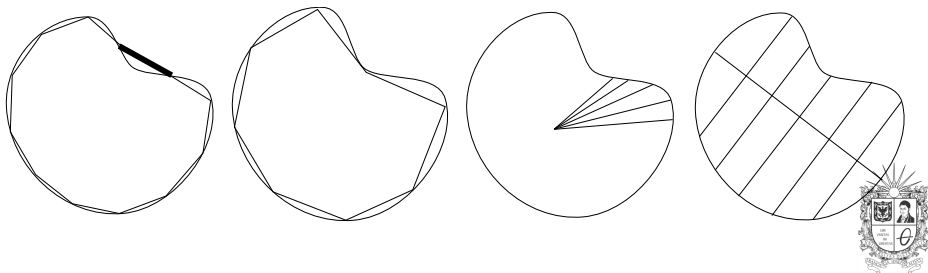
## Verificación de datos

La verificación de los datos extraídos por medio de la muestra FNA y la veracidad del Data-Set de cáncer de mama de Wisconsin se realiza con base a los resultados del proceso realizado por el software conocido como Xcyt. Este sistema ha diagnosticado correctamente 176 nuevos pacientes consecutivos (119 benignos, 57 malignos)[?].



## Clasificación de datos

El Sistema de enfoque de visión por computadora extrae diez características diferentes de los límites (snakes) de los núcleos celulares de la muestra FNA. Las características extraídas son: *Radius*, *Perimeter*, *Area*, *Texture*, *Smoothness*, *Concavity*, *Concave Points*, *Symmetry*, *Compactness* y *Fractal Dimension* [?].



# Desarrollo de la Investigación

## Limpieza de datos

Para el correcto entrenamiento de los modelos con el Data-Set es relevante tener alta calidad en los datos, para ello fue necesario la corrección de los errores de los mismos para el posterior procesamiento con los algoritmos de Machine Learning.

## Formateo de datos

Para la comprobación de los tipos de datos del Data-Set de cáncer de mama de la Universidad de Wisconsin se utilizó la librería *Pandas* la cual cuenta con la propiedad *dtypes* que describe los tipos de datos de las variables de dicho Data-Set.



## Comprobación de datos faltantes

Para la comprobación de datos faltantes se utilizo la librería *Pandas* la cual cuenta con la función *isna()* que valida si las variables contienen valores Nulos(*Nan*). En este caso en especifico el Data-Set de cáncer de mama de la Universidad de Wisconsin no contiene ningún valor nulo.

## Eliminación de datos Innecesarios

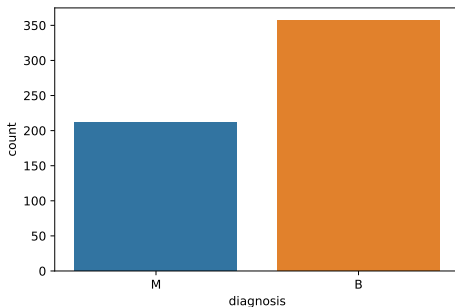
Para la eliminación de datos innecesarios se utilizo a la librería *Pandas* la cual cuenta con la función *dropna()* con la cual se eliminaron datos que no aportaban informacion necesaria para el diagnostico.



# Desarrollo de la Investigación

## Procesamiento de los Datos

El Data-Set se encuentran en formato .csv que es leído y cargado en memoria con la ayuda de las librerías de Python para luego ser asignado a una variable que hace que la información sea manejable. Dentro del Data-Set se encuentran 357 registros de personas con tumores diagnosticados como Benignos y 212 como Malignos, esta información fue utilizada para entrenar y testear el modelo [?].



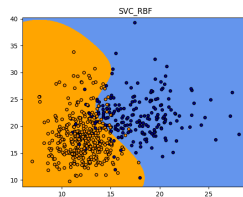
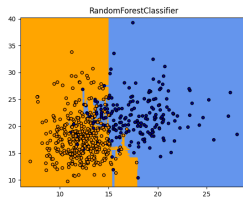
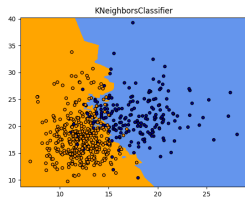
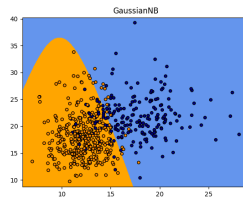
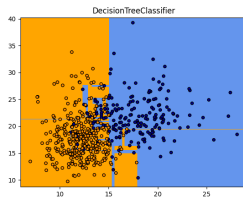
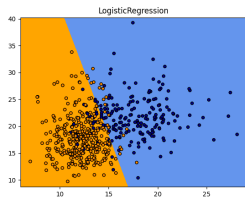
## Modelos de Machine Learning

Para la implementación técnica se utilizó la librería para *Python* llamada *Scikit-Learn* la cual contiene un conjunto de funciones para aplicar dichos modelos. Los modelos seleccionados fueron: *Logistic Regression*, *Decision Trees*, *Gaussian Naive Bayes*, *K-Nearest Neighbors(KNN)*, *Random Forest*, y *Support Vector Machines(SVM)*[?].





# Desarrollo de la Investigación



- Con respecto a los modelos de Machine Learning utilizados por diversos investigadores para predecir el Cáncer de mama se puede evidenciar que el algoritmo más utilizado es el Support Vector Machines (SVM), pero al realizar la comparación con los resultados obtenidos de la precisión de los métodos de Machine Learning observados anteriormente, el algoritmo Decision Trees es el que mejor resultado tiene, con una exactitud del 100 %, esto quiere decir que clasifico correctamente el total de las muestras. Por lo tanto según la investigación realizada se concluye que si se va a diagnosticar el Cáncer de mama los modelos más indicados para hacerlo son el Support Vector Machines(SVM) y Decision Trees.



- Según el análisis realizado con base en el diagrama de calor conformado por la correlación de las variables del Data-Set de la Universidad de Wisconsin se puede evidenciar que las variables *conca-ve\_points\_worst* y *area\_worst* generan información relevante en la realización del diagnóstico de Cáncer de mama debido a que expresan una deformidad mayor de los núcleos celulares encontrados en las masas mamarias extraídas por el método de Aspiración con Aguja Fina(FNA).




- Implementación de una capa de servicios REST basada modelos de Machine Learning para el diagnóstico de Cáncer de mama que podría ser utilizada en diferentes ámbitos en la detección y el diagnóstico de dicho Cáncer.
- Diseño y Arquitectura de un aplicativo web enfocado en el uso de modelos de Machine Learning aplicados en la rama de la Medicina especializada en Oncología.



- Creación de una aplicación web llamada OncoAnalysisApp la cual permita el diagnóstico de cualquier tipo de Cáncer teniendo como entrada Data-Sets obtenidos por diversos métodos médicos.
- Creación de una aplicación que permita el análisis de imágenes y que diagnostique el padecimiento de Cáncer de mama con base a los modelos de Deep-Learning existentes.
- Creación de una aplicación que permita crear nuevos Data-Set dinámicamente según parámetros proporcionados por el usuario.



-  International Agency for Research on Cancer.  
170 Colombia fact sheets.  
*Globocan 2020*, 509:1–2, 2020.

