Universidad Autónoma de Nuevo León

Licenciatura en Actuaría

Minería de datos

AVANCE I PROYECTO INTEGRADOR

Equipo 14

Suarez Martinez Tadeo Alejandro #1806069

Grupo: 002

7to Semestre

Octubre del 2019

Título de la base de datos:

Breast Cancer Wisconsin (Diagnostic) Data Set

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data)

Descripción de los datos:

Los datos vienen en una tabla con columnas, a continuación en la siguiente tabla describo lo que cada columna se refiere.

Columna	Descripción
ID number	Número de identificación del paciente.
Diagnosis	Diagnostico del paciente. M = malingo, B = benigno
Radius_mean	Media de las distancias de los radios
texture_mean	Desviación estándar de los valores de la escala de grises
perimeter_mean	Tamaño medio del tumor central
area_mean	Media del area
smoothness_mean	Media de variación local en longitudes de radio
compactness_mean	media del perímetro
concavity_mean	media de la gravedad de las porciones cóncavas del contorno
concave points_mean	media del número de porciones cóncavas del contorno
symmetry_mean	simetría media
fractal_dimension_mean	media de la dimensión fractal
radius_se	Desviación estándar para la media de las distancias desde el centro hasta los puntos del perimetro
texture_se	Desviación estándar para la desviación estándar de los valores de la escala de grises

perimeter_se	Desviación estándar del perimetro
area_se	Desviación estándar del area
smoothness_se	Desviacion estándar para variación local en longitudes de radio
compactness_se	Desviación estándar para el perímetro
concavity_se	Desviacion estándar para la severidad de las porciones cóncavas del contorno.
concave points_se	Desviacion estándar para el número de porciones cóncavas del contorno
symmetry_se	Desviación estándar de la simetria
fractal_dimension_se	Desviación estándar para la dimensión fractal
radius_worst	Mayor valor medio más grande para la media de las distancias desde el centro hasta los puntos del perímetro
texture_ worst	Mayor valor medio más grande para la desviación estándar de los valores de la escala de grises
perimeter_ worst	Mayor valor medio más grande del perimetro
area_ worst	Mayor valor medio más grande del area
smoothness_ worst	Mayor valor medio más grande para variación local en longitudes de radio
compactness_ worst	Mayor valor medio más grande para el perímetro
concavity_ worst	Mayor valor medio más grande para la severidad de las porciones cóncavas del contorno.
concave points_ worst	Mayor valor medio más grande para el número de porciones cóncavas del contorno
symmetry_ worst	Mayor valor medio más grande de la simetria
fractal_dimension_ worst	Mayor valor medio más grande para la dimensión fractal

Justificación del uso de datos:

Me decidí a trabajar con esta base de datos porque me quería centrar en alguna enfermedad mortal, para ayudar con el diagnostico de esta, ya que se pueden emplear muchas técnicas, busqué en varias bases de datos y esta fue la que vi que tenía información más útil y concreta.

Cuenta con información sólida y los datos están muy completos, confió plenamente en que puedo hacer un buen trabajo de predicción con estos, ya que tiene toda la información necesaria acerca del tumor de cada paciente.

Planteamiento del problema:

El cáncer de mama es el tipo de tumor más frecuente en mujeres, es una enfermedad sumamente grave, pero se puede curar si se detecta a tiempo, es por eso que este modelo quiere ayudar con esta problemática del mundo de la medicina, este programa ayudara a cualquier hospital, clínica o grupo médico.

Objetivo final:

El objetivo final es predecir si el tumor es maligno o benigno en base a sus características anteriormente mencionadas como por ejemplo el tamaño de este, su perímetro entre otros, esto ayudara a la empresa que compre este programa en su diagnóstico para esta enfermedad.

Planeación de la herramienta a utilizar.

El modelo para predecir este problema puede ser varios, ya que es un problema de clasificación, y los modelos útiles para estos problemas pueden ser los siguientes:

1) Regresión logística binaria:

   Este método estadístico sirve para predecir clases binarias, es decir de naturaleza dicotómica ósea solo hay dos clases, en este caso es si el tumor es maligno o benigno, debido a esas características es inevitable no pensar en este modelo para hacer esta predicción, obviamente mi variable dependiente será el diagnostico y mis variables independientes podrían ser el radio, perímetro y concavidad.

2) Árbol de decisión:

   Otro algoritmo que pienso implementar es el árbol de decisión, ya que también es un algoritmo de clasificación muy útil para resolver esta clase de problemas, utilizaremos reglas de decisión en base al radio, perímetro y concavidad (Probablemente).

3) Bosque aleatorio:

   Y no podía faltar el bosque aleatorio también uno de los algoritmos mas usados para la clasificación el cual al igual que el árbol lo implementare y comprobare que variables tienen mayor “exactitud” para usarlas como nodos.

Después de usar estos tres métodos de clasificación comprobare cual tiene un mejor porcentaje de producción.