**Introducción**

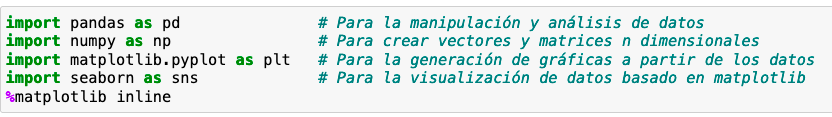
Se utilizarán árboles de decisión para generar un modelo que permita clasificar tumores malignos y benignos a partir de estudios clínicos. Además, se probarán diferentes configuraciones para la creación de árboles de regresión cambiando algunos parámetros para generar diferentes árboles de decisiones. Se analizará cada uno de los árboles para saber cuál es mejor. Se utilizarán conceptos vistos en prácticas pasadas para la selección de variables y la generación de un conjunto de datos de entrenamiento y de prueba.

**Objetivos**

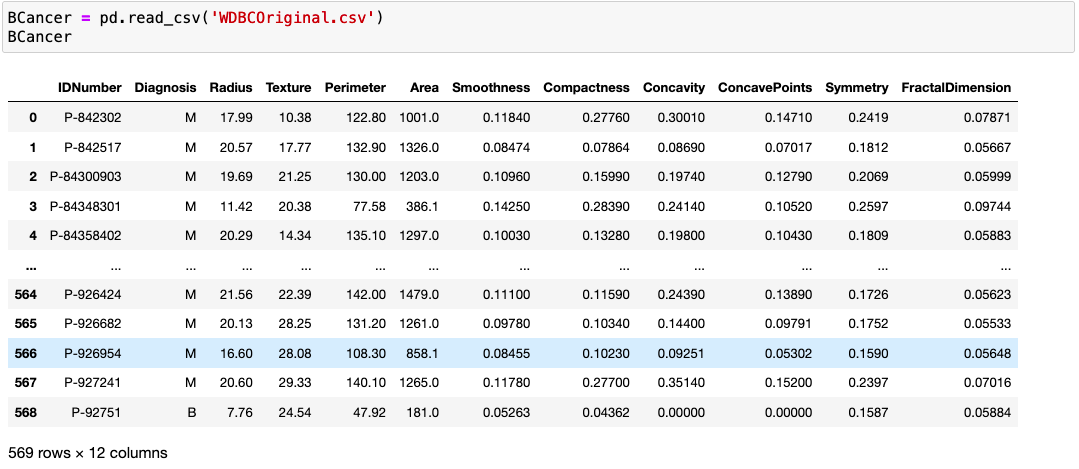
Pronosticar el área del tumor de pacientes con cancer de mama a través de un árbol de decisión utilizando estudios clínicos a partir de imágenes digitalizadas de pacientes con cáncer de mama de Wisconsin (WDBC, Wisconsin Diagnostic Breast Cancer).

**Desarrollo**

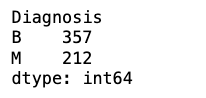
Se importan las bibliotecas necesarias para la manipulación de datos.



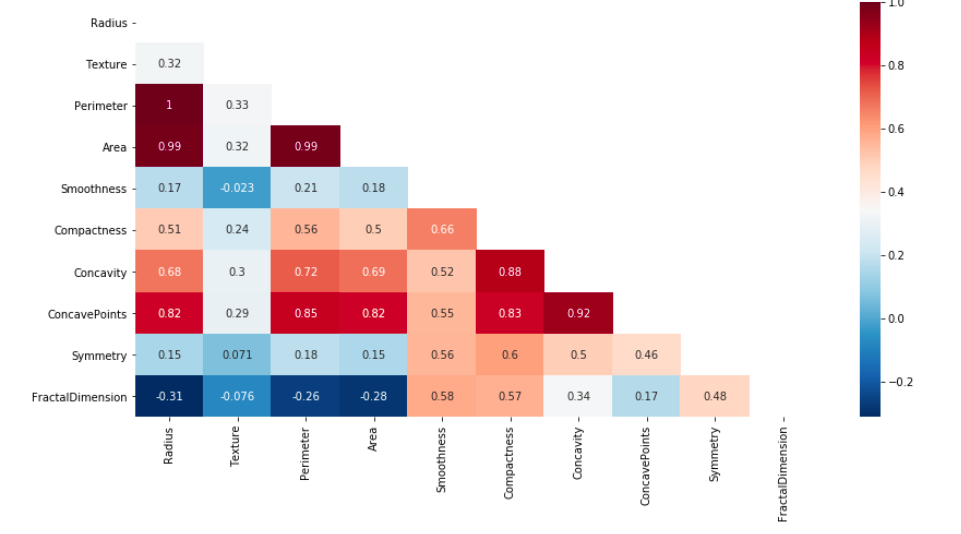
Se importan los datos del cáncer de mama para utilizarlos.



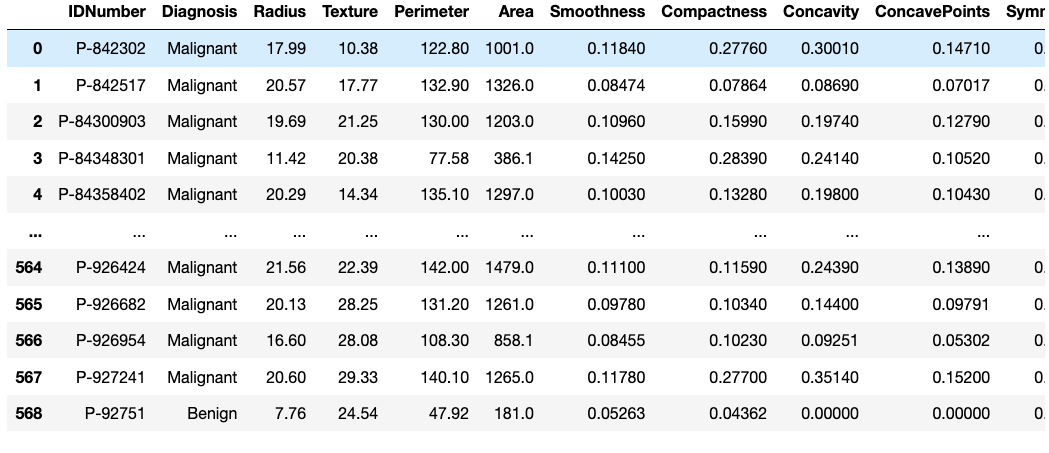
Se agruparon los datos para saber cuántos tumores eran malignos y benignos.



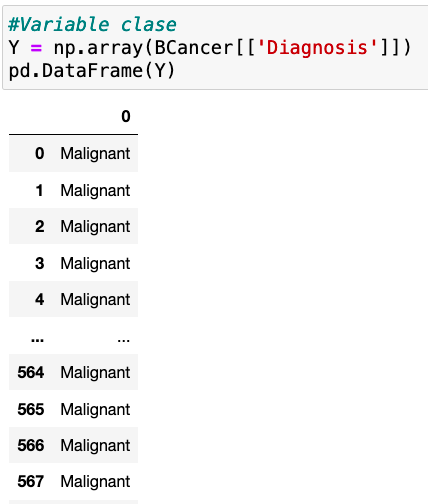
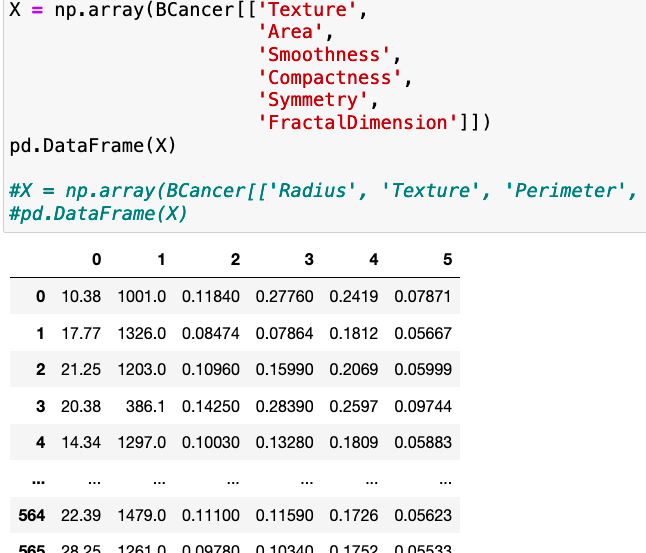
Se realiza la selección de variables utilizando un mapa de calor y las variables seleccionadas son: textura, área, smoothness, compactness, symmetry, FractalDimension y perímetro.



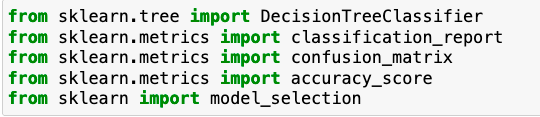
Se reemplaza el valor de maligno por M y de benigno por B para los datos que se utilizarán.



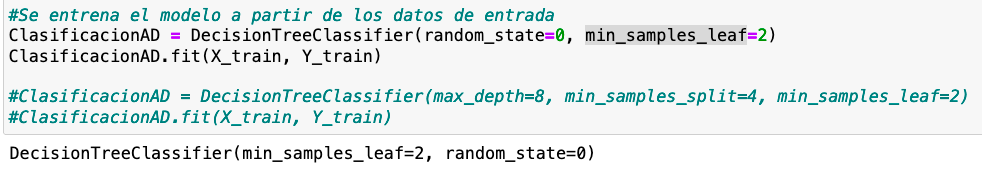
Para esta práctica se generará una predicción de clase utilizando las variables predictoras que se seleccionaron y la variable a predecir será el diagnosis.



Se importan nuevas bibliotecas necesarias para las actividades a realizar en esta práctica.



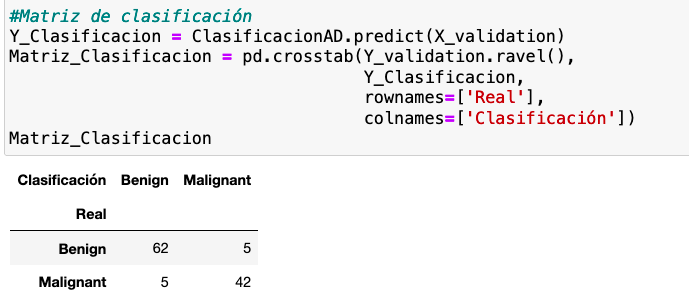
Se entrena el modelo a partir de los datos de entrada utilizando la biblioteca de sklearn para generar árboles de decisión clasificadores. Se genera una semilla para los datos aleatorios y el mínimo de los samples de la hoja son 2.



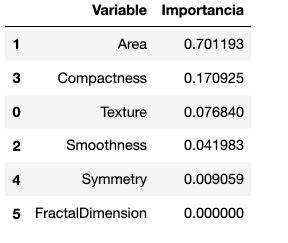
A partir del modelo se generó la predicción a partir de los datos de prueba. Además, se calculó el score entre los datos verdaderos y los datos pronosticados.



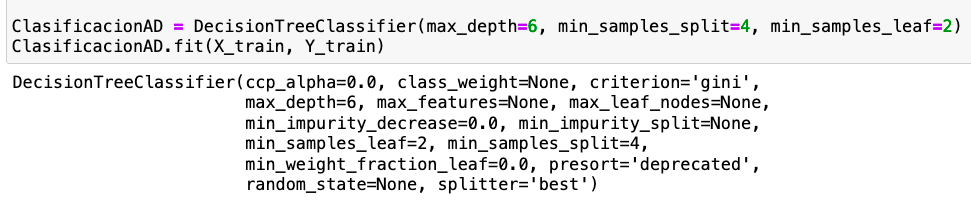
Se generó una tabla de los falsos positivos y negativos para observar cómo funciona el modelo.

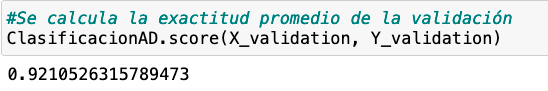


También se genera una tabla para visualizar las variables más significativas para el modelo.

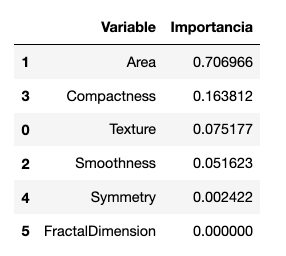
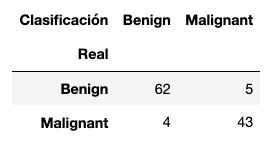


Al igual que en la práctica pasada se busca evitar el sobreajuste del modelo ya que existen hojas que tienen mucha especificidad. Se busca tener un modelo más general. Se generó un modelo con máximo de profundidad de 6 y el *score* mejoró a .921.

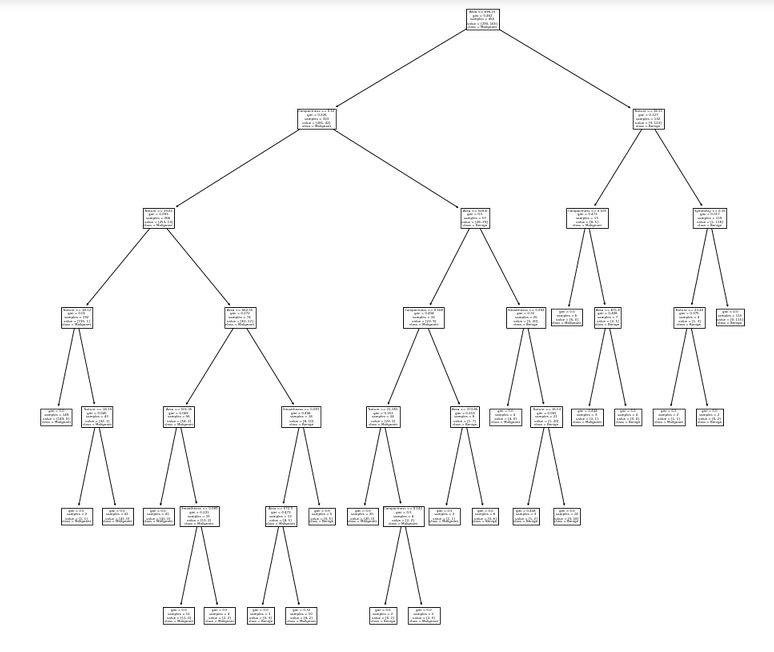




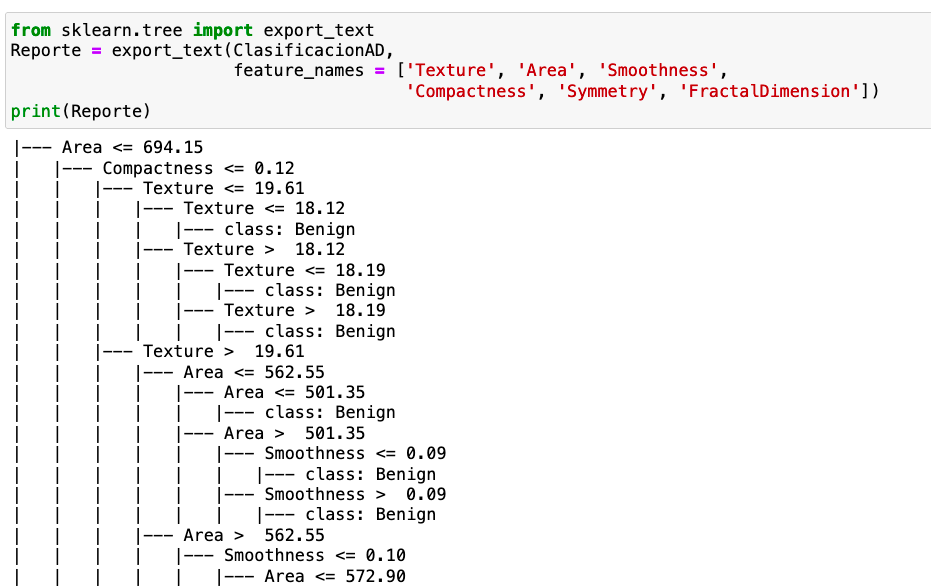
Al igual que en el modelo pasado se obtuvo una tabla de falsos positivos y falsos negativos. Además se generó una tabla donde se muestra la importancia de cada una de las variables que utiliza el árbol.



A partir de la tabla se tomó la decisión de quitar la variable de fractalDimension, pero el *score* empeoró por lo que se mantuvo el modelo que se generó con profundidad máxima de seis y todos las variables. Gráficamente el modelo que fue generado se puede ver cómo:



Se pueden generar múltiples configuraciones seleccionando diferentes variables y múltiples niveles máximos de profundidad. Como en la práctica pasada se generaron reglas para entender cómo funciona el modelo.



Además, se realizaron predicciones para dos pacientes y se clasificaron para saber si el modelo predice que es benigno o maligno. Para cada uno de los pacientes se tuvieron las siguientes salidas.



**Conclusiones**

Se utilizaron árboles de decisión para generar modelos que clasifican un tumor en maligno o benigno utilizando las variables predictoras. Se generaron diferentes configuraciones que permitieron hacer modelos generalizados. Al igual que en la práctica pasada, se generó un modelo utilizando la biblioteca del árbol de decisión y se tuvo un sobreajuste en este caso. No se tuvo un score de .91 lo cual no es tan bueno. Generando un modelo con profundidad máxima de 8 niveles se pudo obtener un árbol con mejor score (.92) y que evita el sobreajuste. Este modelo no obtuvo tan buenos scores por lo que se puede concluir que el modelo anterior es mejor en este ámbito. Sin embargo, no realizan lo mismo, por lo que si se quiere una clasificación convendría utilizar alguno de los modelos antes vistos de clasificación o los bosques aleatorios de clasificación. Se logró analizar los resultados que produce este tipo de modelos y concluir cuál es el mejor y en qué momentos se debe utilizar.