**Introducción**

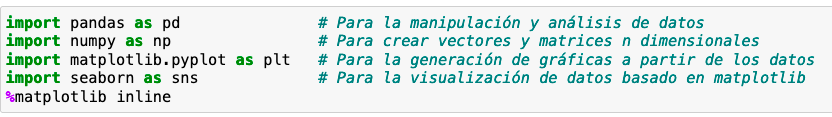
En esta práctica se utilizarán pronósticos de bosques aleatorios para predecir el área del tumor de pacientes con indicios de cáncer de mama. Se compararán los resultados con los obtenidos en prácticas pasadas utilizando pronósticos de un árbol. Además se obtendrá la estructura de un árbol que conforma parte del bosque para saber cómo se conforma.

**Objetivos**

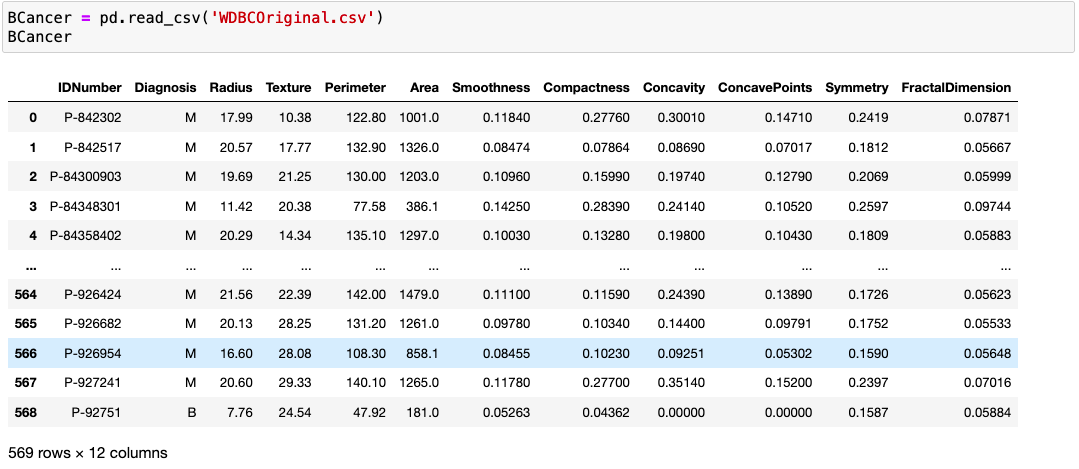
Pronosticar el área del tumor de pacientes con indicios de casos de cáncer de mama a través de bosques aleatorios utilizando estudios clínicos a partir de imágenes digitalizadas de pacientes con cáncer de mama de Wisconsin (WDBC, Wisconsin Diagnostic Breast Cancer).

**Desarrollo**

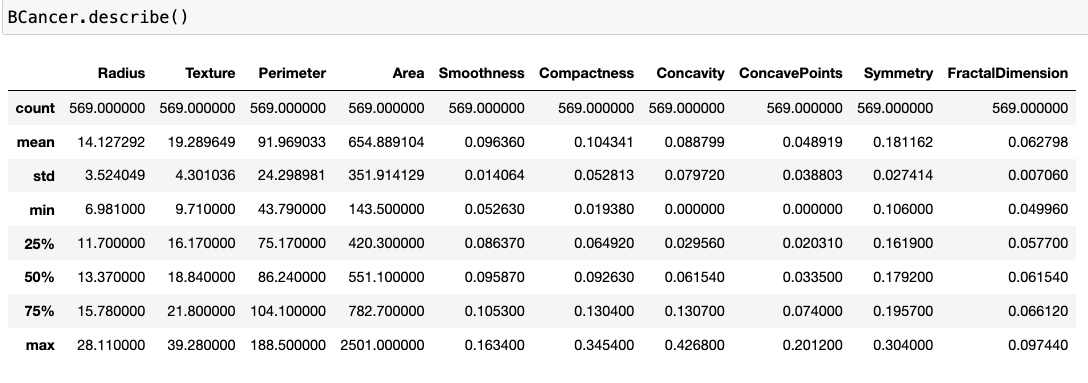
Se importan las bibliotecas necesarias para la manipulación de datos.



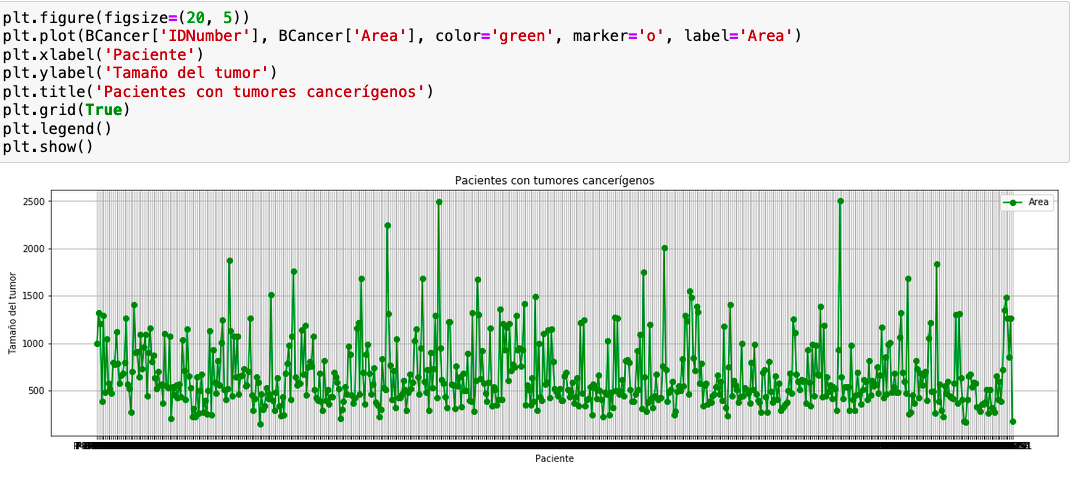
Se importan los datos del cáncer de mama para utilizarlos.



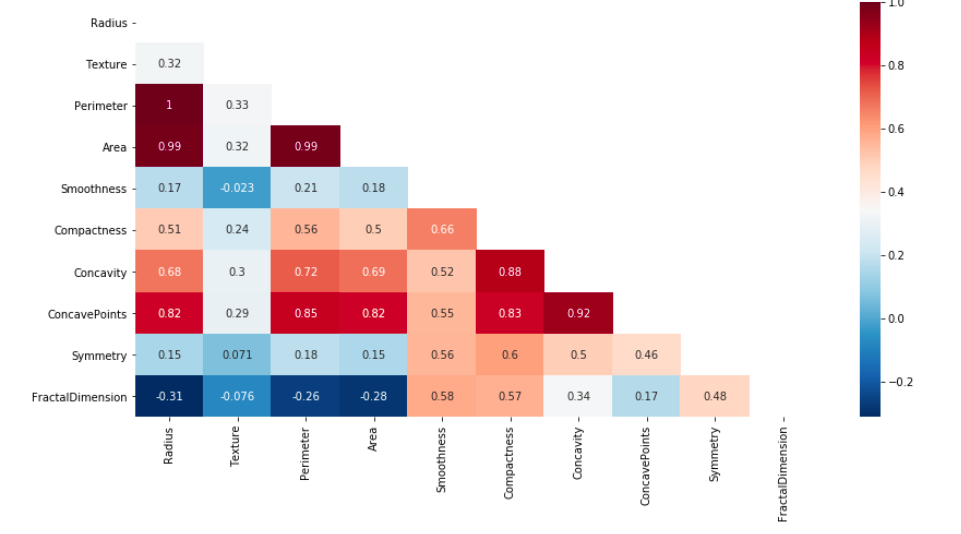
Se describen los datos que se utilizarán en el algoritmo para obtener las estadísticas generales.



Se realizó una gráfica del área de tumor por paciente para visualizar los datos.



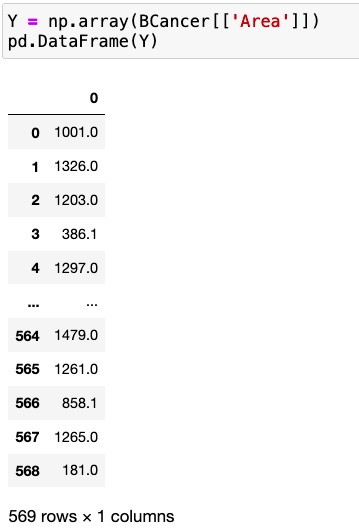
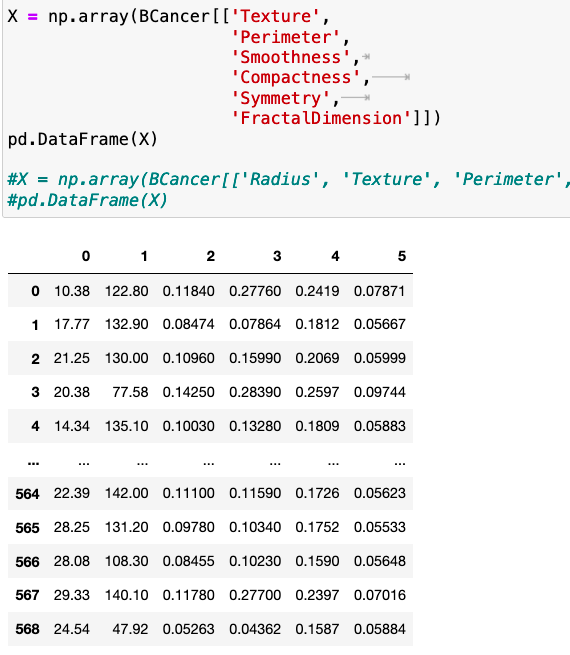
Se realiza la selección de variables utilizando un mapa de calor y las variables seleccionadas son: textura, área, smoothness, compactness, symmetry, FractalDimension y perímetro.



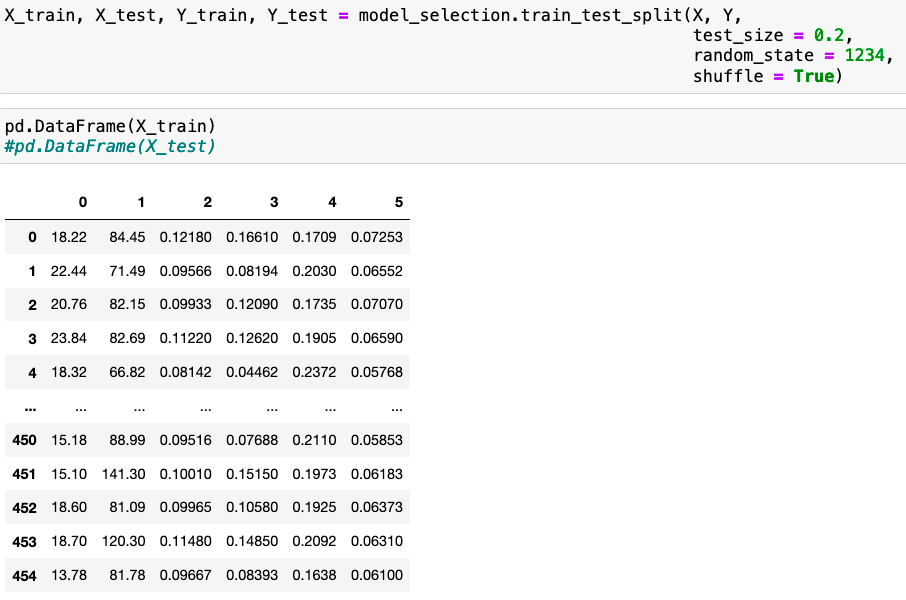
Se importan algunas bibliotecas para estadísticas sobre el modelo y la creación del bosque aleatorio.



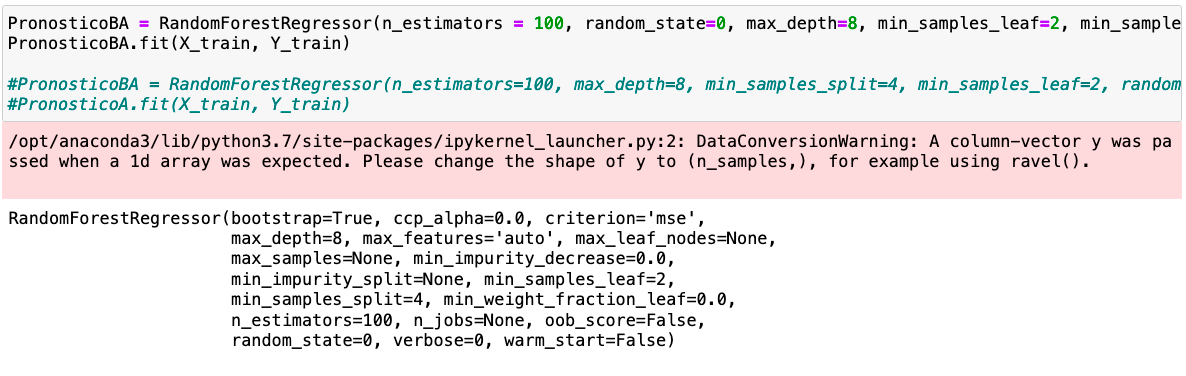
Se seleccionan las variables que serán las predictoras y las variables a predecir para generar *dataFrames* a partir de ellas. Los conjuntos de datos generados son los siguientes.



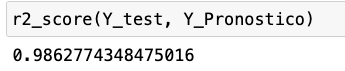
Se generó la división de datos para separar los datos de entrenamiento y de prueba.



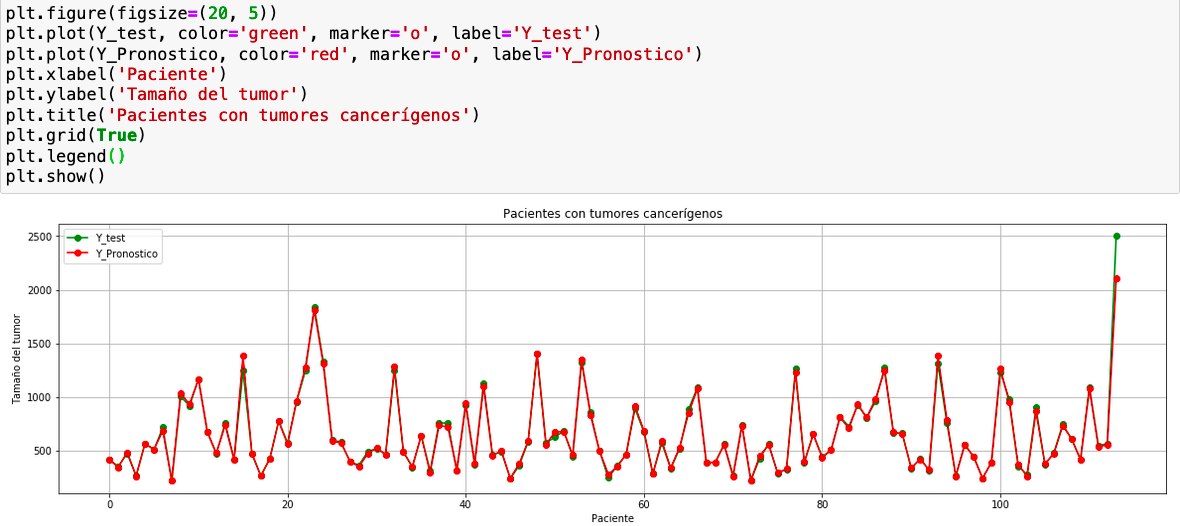
Se crea el bosque aleatorio utilizando la biblioteca de *sklearn*. La salida del modelo creado se ve de la siguiente manera.



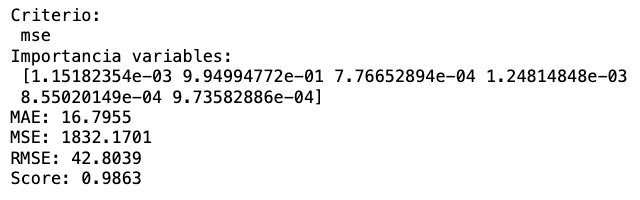
Se generan los pronósticos y los valores reales para poder obtener el *score* de los pronósticos. Con 100 estimadores se tiene un *score* de .986. Este *score* puede cambiar con los estimadores utilizados y con la selección de variables. Con 50 estimadores el *score* baja a .985 y con 200 estimadores cambia a .986 que es muy parecido al primer valor.



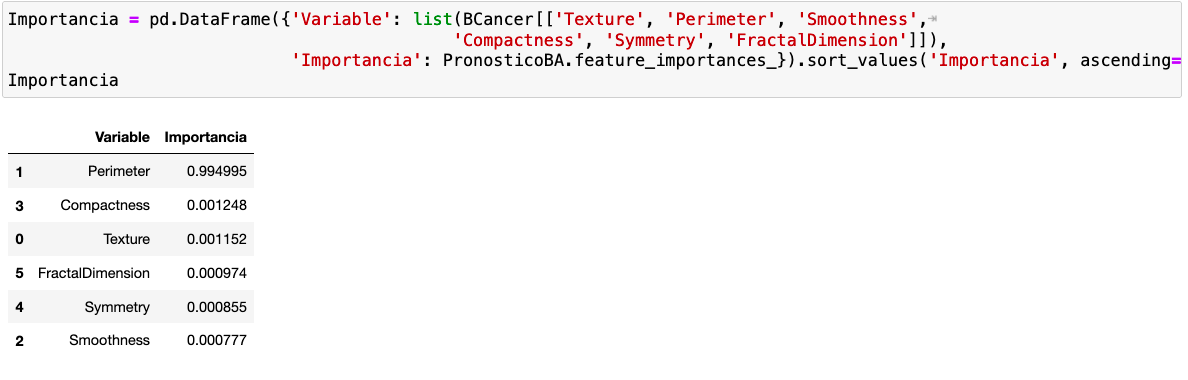
Con la configuración de 100 estimadores se generó una gráfica para observar la diferencia entre los pronósticos generados por el modelo y los datos reales. La gráfica que se obtuvo fue la siguiente.



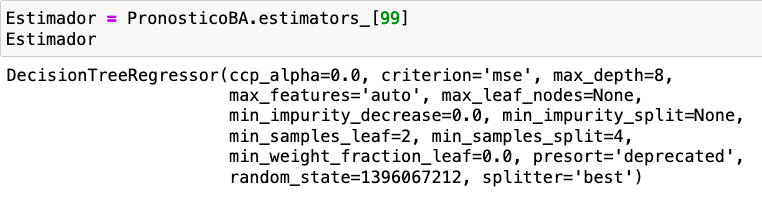
También se obtuvieron los parámetros de este modelo los cuales se enseñan a continuación.



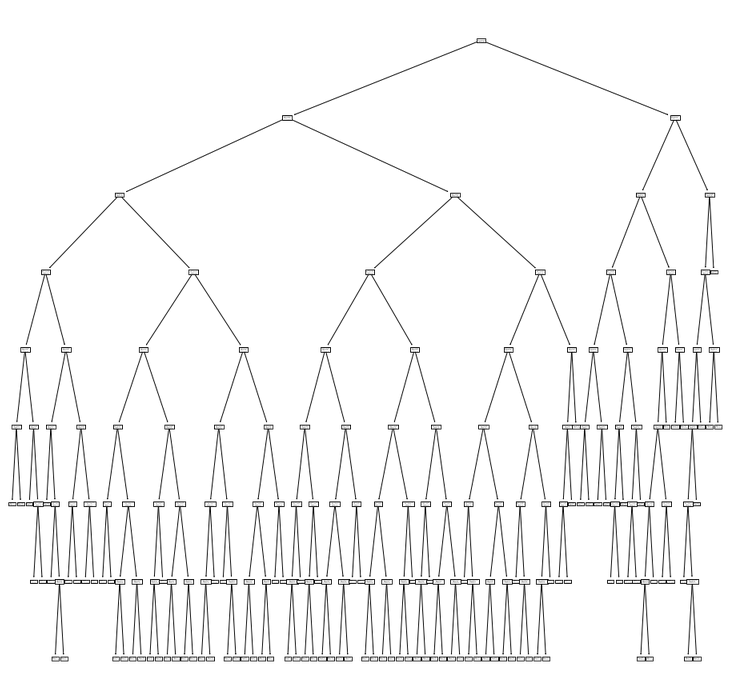
También, se obtuvo la importancia para el modelo de cada una de las variables que se utilizan. De estos datos se pudo obtener que la variable más importante para el modelo generado es el perímetro.



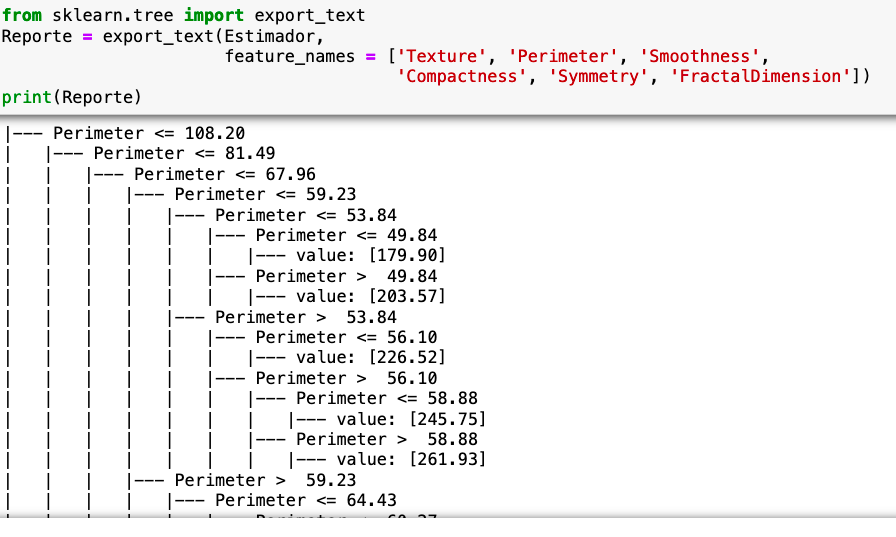
Después se grafica uno de los estimadores de los bosques para graficar y poder observar los datos que contiene y la manera en que trabaja.



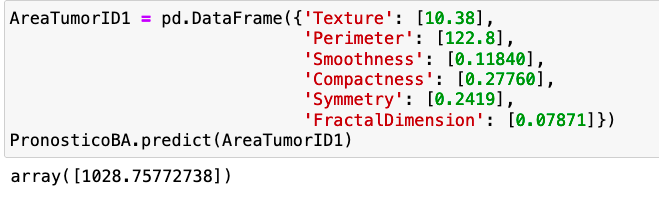
Se grafica el árbol de decisión como en prácticas pasadas y se tiene una versión donde se puede observar cada uno de los campos.



Además se generó una versión del modelo donde se describe dicho estimador utilizando texto. Este texto es bastante extenso, pero puede ser más claro que el diagrama.



Por último, como en prácticas anteriores, se genera una predicción para uno de los datos que se utilizan y predecir el área del tumor. La predicción que se generó es la siguiente.



Se puede observar cómo se tiene menos sobreajuste ya que se tienen varios árboles en conjunto que permiten generar un modelo global. Cuando solo se tiene un árbol esto puede desencadenar en sobreajuste. Se recomienda utilizar bosques aleatorios en vez de un solo árbol.

**Conclusiones**

En esta práctica se logró utilizar bosques aleatorios para generar pronósticos para predecir el área de un tumor de pacientes de cáncer de mamá. Se pudo observar cómo generar bosques aleatorios utilizando Python y *sklearn*. Se ajustaron los diferentes parámetros para generar un bosque aleatorio de 100 árboles. Se compararon los resultados obtenidos con aquellos de un solo árbol y se concluyó que se tiene mejor *score* y que se tiene un modelo que no cae tan fácilmente en sobreajuste como los árboles. El bosque utiliza a los árboles como componentes y permite que se tengan modelos más generales. También se pudo obtener la estructura de uno de los árboles que conforma el bosque. La estructura de este árbol es similar a la vista en prácticas pasadas y la forma en que se obtiene es similar. Con esta práctica se pudo observar la utilidad y eficiencia de utilizar los bosques aleatorios para generar pronósticos y cómo estos son preferibles a solo utilizar un árbol.