**CMP3751M - Machine Learning**

**Assessment 2 - Written Report**

Shangyuan Liu

University of Lincoln, School of Computer Science

[25344136@students.lincoln.ac.uk](mailto:25344136@students.lincoln.ac.uk)

# Section I: **Data import, summary, pre-processing and visualisation**

## Load the Data

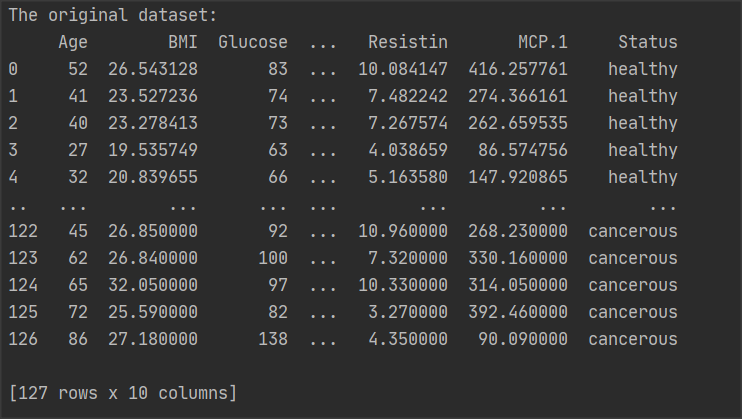
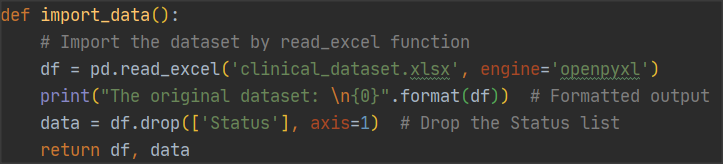
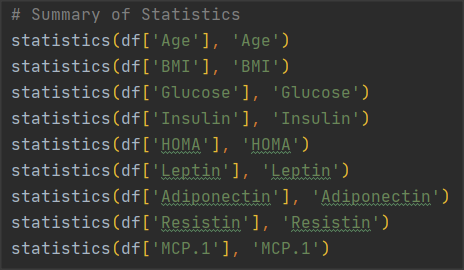
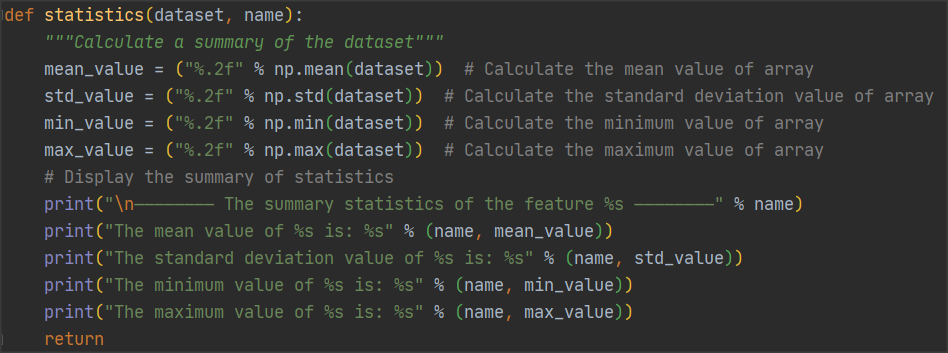


Figure 1: Function import\_data() - Load the dataset into Python IDE

Function import\_data is used to load the data. The approach of the function is importing Pandas, which is a data analysis package for python providing functions and methods to manipulate data quickly and easily. And use read\_excel to an Excel file into a DataFrame, the original dataset size is 127 rows × 10 cols.



## Summary of the Dataset

Figure 2: statistics() - Statistical summary of data set

This function (as shown in figure 2) provides a summary (mean, standard deviations, min/max values) of each feature in the dataset. Using the mean(), max(), and min() methods in the numpy library in the function to provide a summary. And the result of the summary is shown in Figure 3.

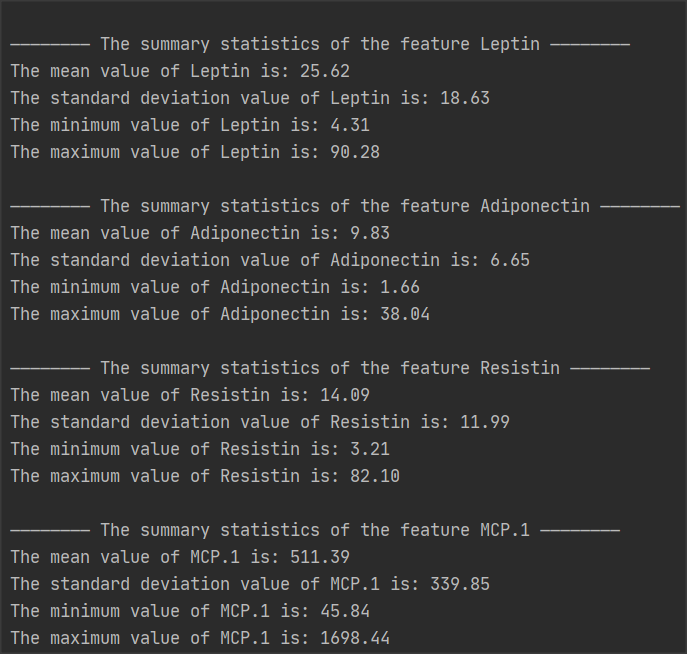
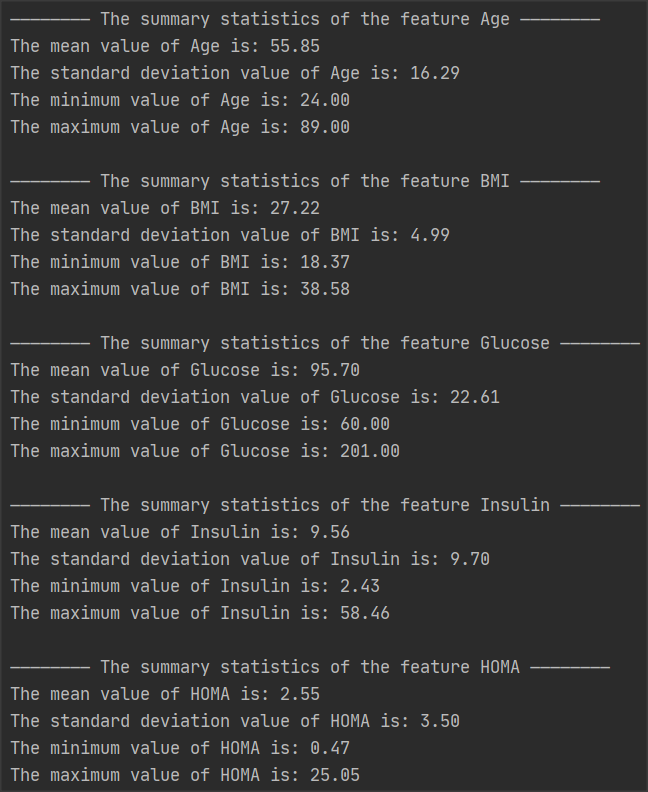


Figure 3: The result of statistical summary

**Report the Data Size and Features**

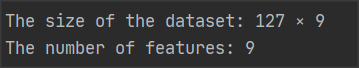
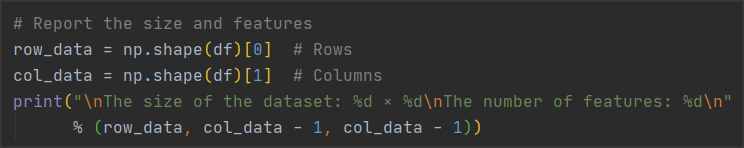


Figure 4: Report the size of the data and number of features

The np.shape method is used in this step, which is to report the dimension of a matrix or array. Shape()[0] is for counting the number of rows, and Shape()[1] is for counting the number of cols. As can be seen from figure 4 that the size of data is 127 rows × 9 cols, and there are 9 clinical features that can classify patients as healthy or cancerous.

**Find Missing Values**

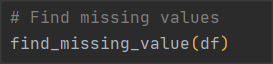
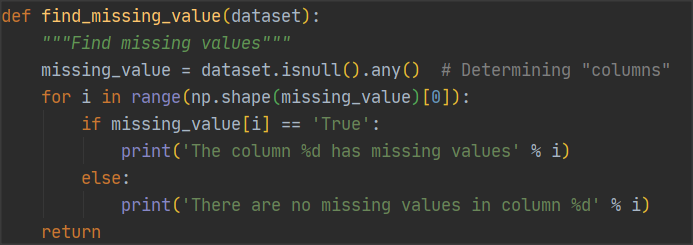
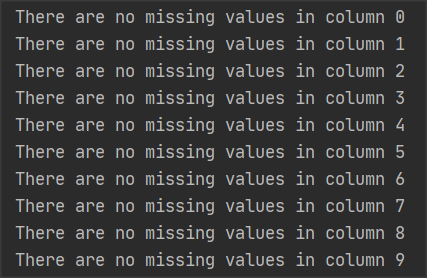


Figure 5: Function find\_missing\_value to find the missing value in dataset



In the pandas library, there is an effective function isnull() , which can be used to identify missing values. And isnull().any() will determine which columns contain missing values, and return True if there are missing values in the column, otherwise False. Therefore, as shown in the figure on the right, there is no missing value in this dataset.

**Find Categorical Variables**

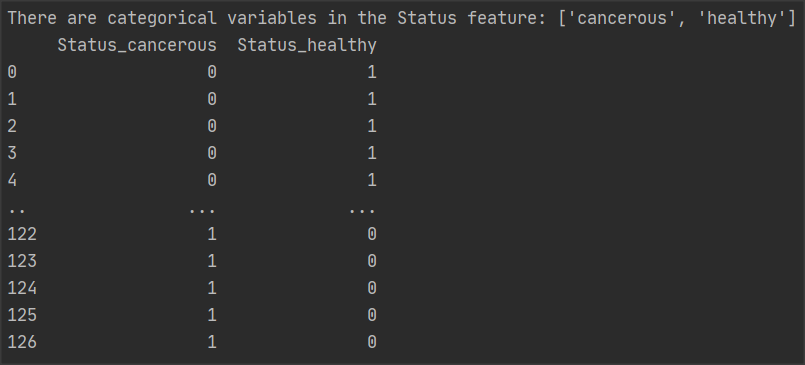
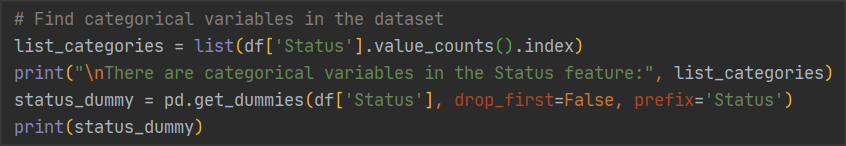


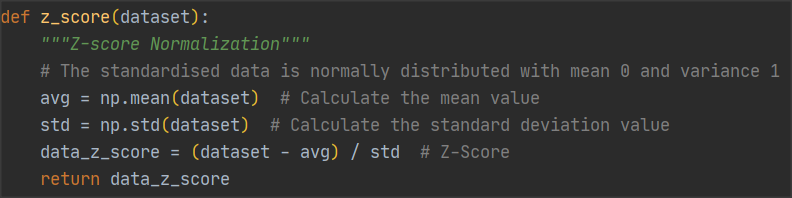
Figure 6: Find categorical variables

There are any categorical variables in the dataset, the categorical variables are 'cancerous' and 'healthy' respectively in the 'Status' list It is clear from figure 6 that value\_counts() is a method to check how many different variables are in a column of the table, and then use the list method to store them in the list.

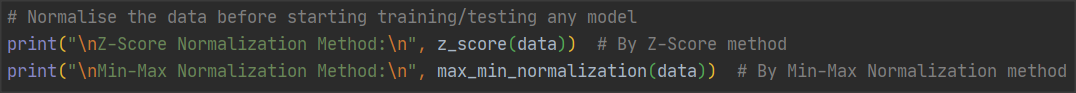
get\_dummies

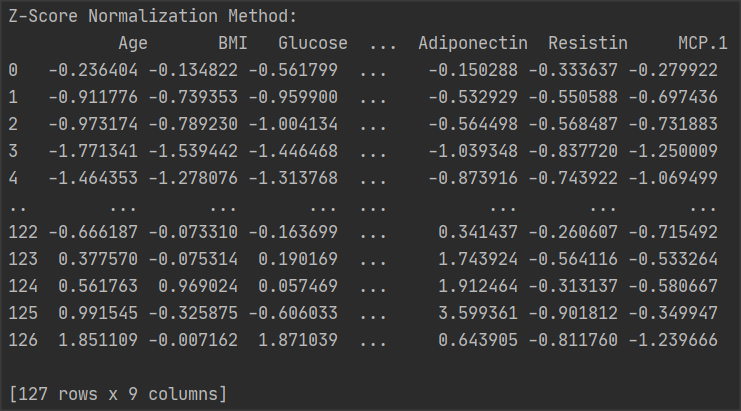
Convert categorical variable into dummy / indicator variables.

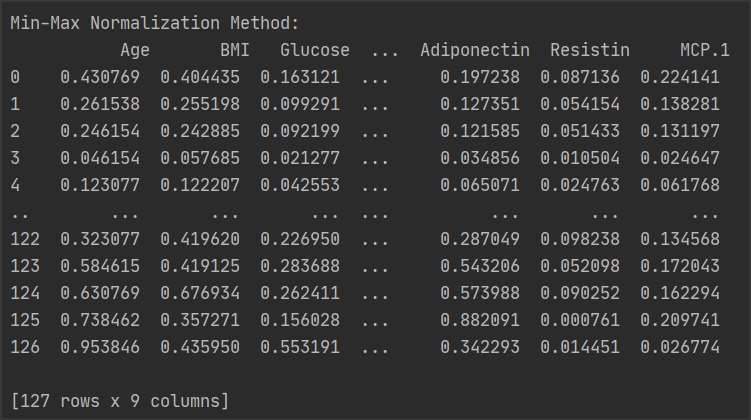
**Data Normalization**

****



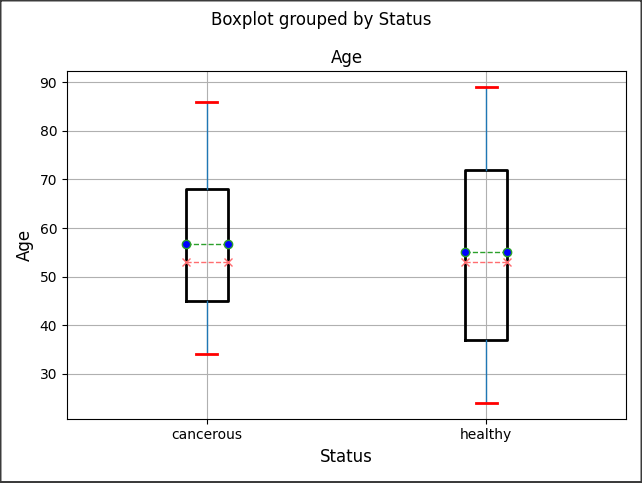
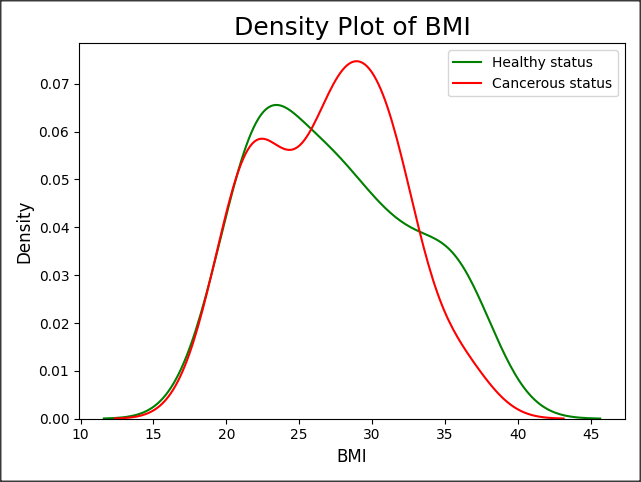






**Data Visualisation**





**Section II: Discussion on selecting an algorithm**