**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

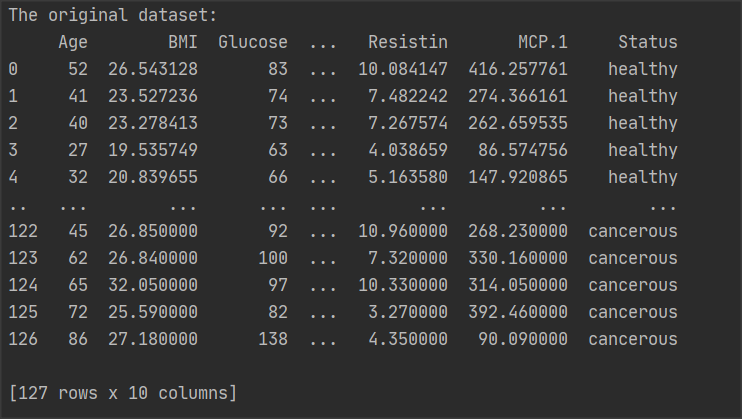
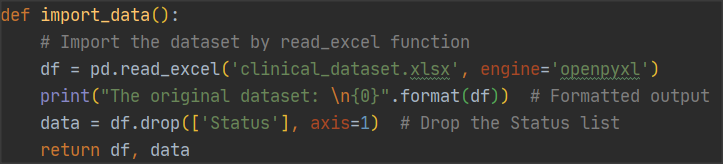
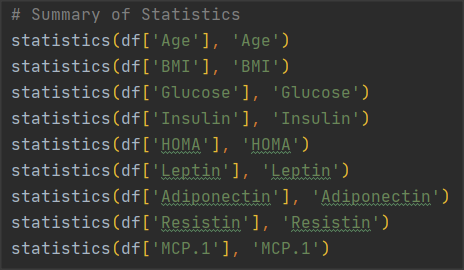
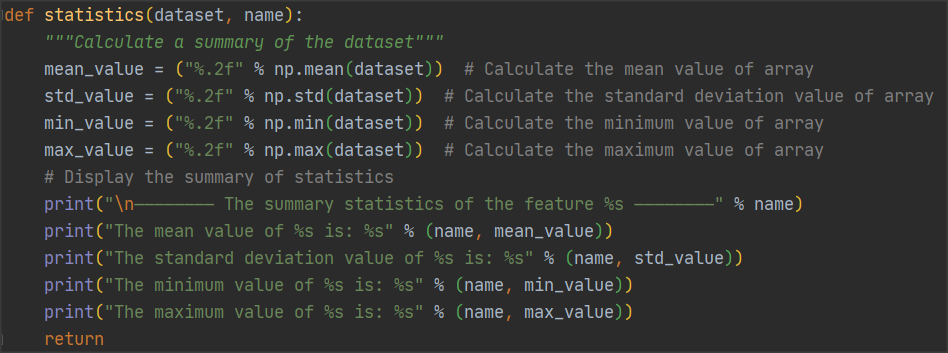


Figure 1: Function import\_data() - Load the dataset into Python IDE

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.

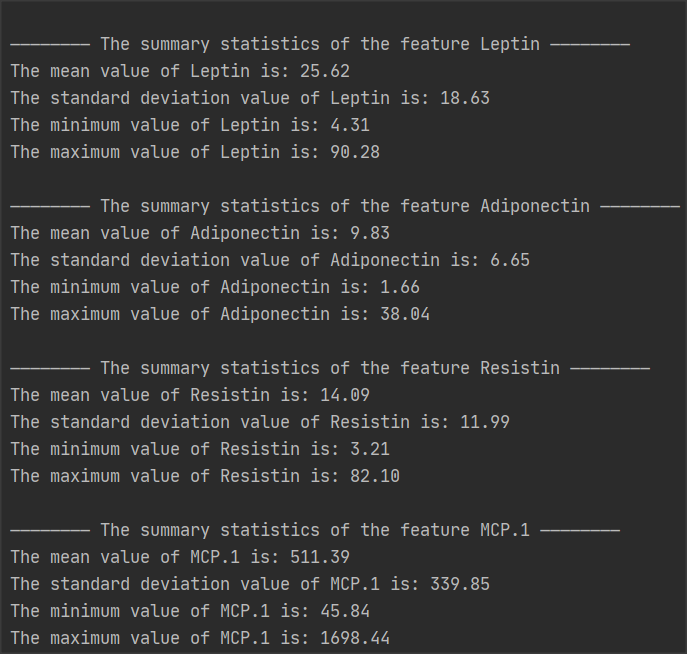
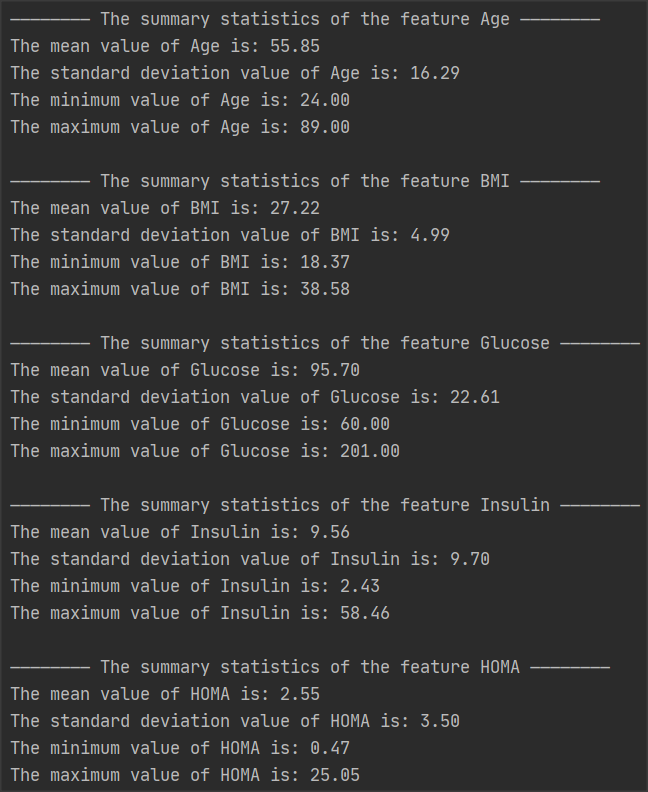


Figure 3: The result of statistical summary

**Report the Data Size and Features**

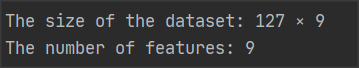
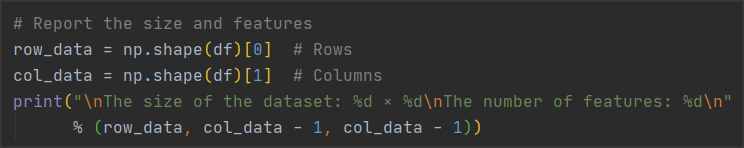


Figure 4: Report the size of the data and number of features

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**

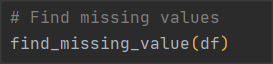
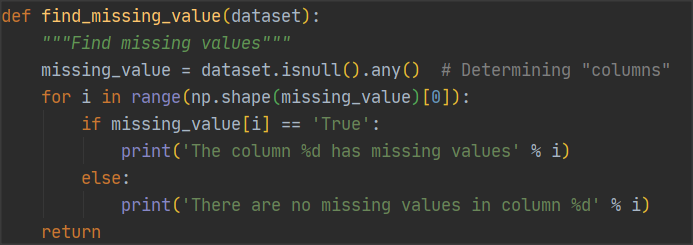
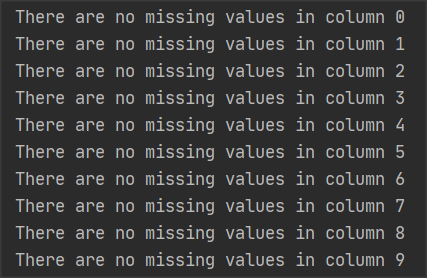


Figure 5: Function find\_missing\_value to find the missing value in dataset



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**

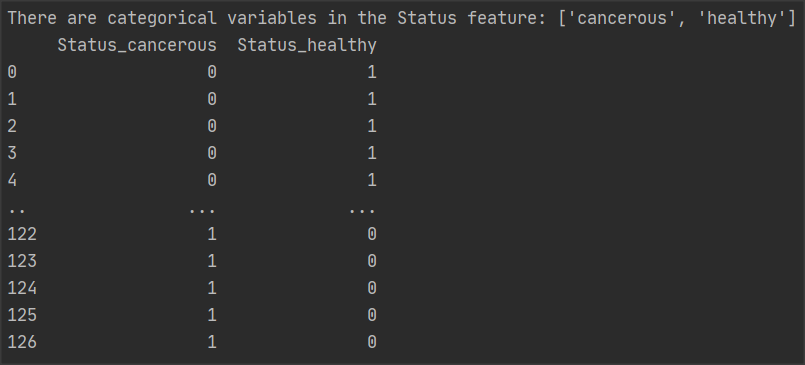
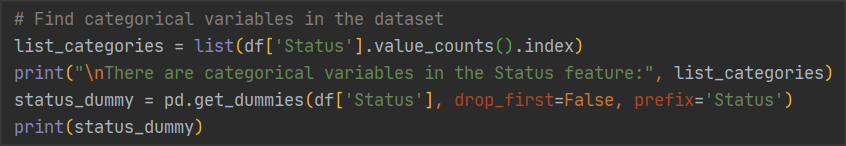


Figure 6: Find categorical variables

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.

Finally, using function get\_dummies() for one-hot encoding to convert categorical variables into dummy/indicator variables, and the result of the change in the status column after the one-hot code is applied is as shown in figure 6.

**Data Normalisation**

Figure 7: Functions z\_score() and max\_min\_normalization() to data normalization

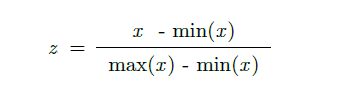
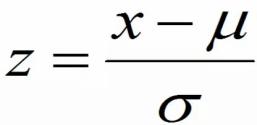
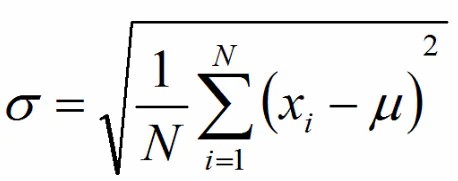
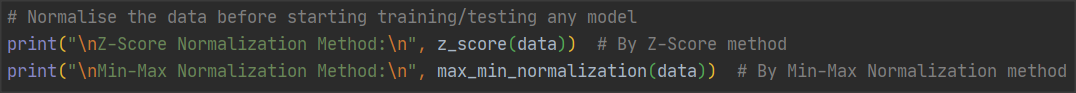
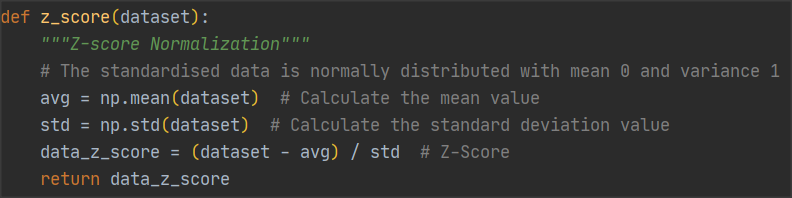
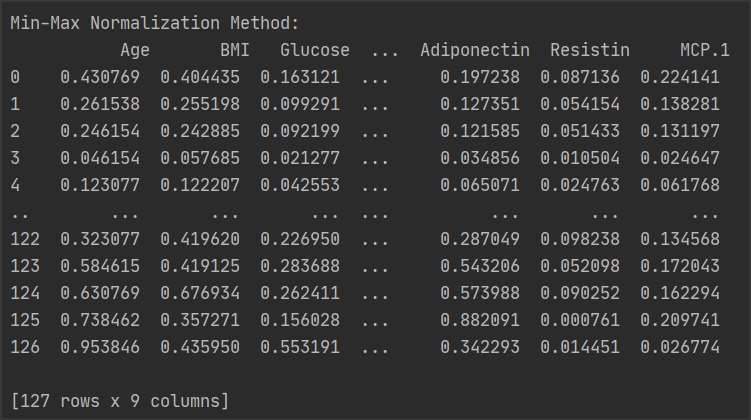
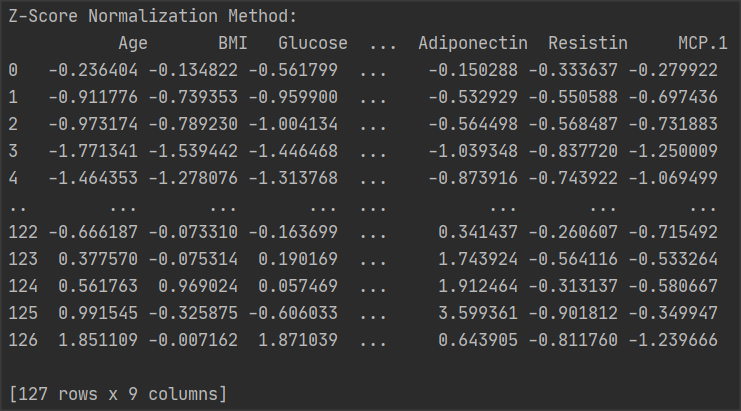


Figure 8: z-score and max\_min\_normalization formulas

In my opinion, it is really necessary to have a data normalisation processing for machine learning models. Sola and Sevilla (1997) indicated that “An adequate normalization, not only for the network output variables but also for the input ones, previous to the training process is very important to obtain good results and to reduce significantly calculation time”. Therefore, it is significant to normalize the data in order to overcome the problem of model learning. Moreover, the model ensures that a similar range of values is used for different features so that gradient descent can converge more quickly.

There are two approaches used for data normalization, which are Z-Score and [Min-Max normalisation respectively. The formulas are displayed in figure 8 and implemented in the functions z\_score and max\_min\_normalisation respectively (as shown in figure 7). The Min-Max normalisation is a linear transformation of the original data to map the values to between [0, 1]. However, the Z-Score is based on the mean and standard deviation of the original data to standardise the data, the main purpose is to standardise data of different magnitudes to the same magnitude, using the calculated Z-Score value to ensure comparability between data.](https://www.codecademy.com/articles/normalization) The results of the data normalisation are displayed in figure 9.

Figure 9: The result of two methods for data normalisation



**Data Visualisation**

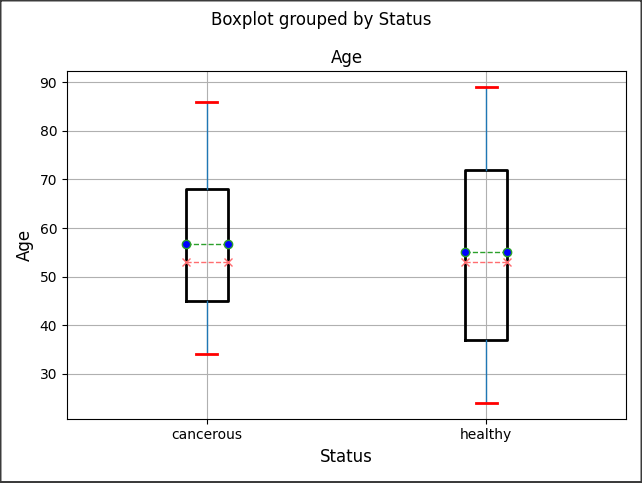
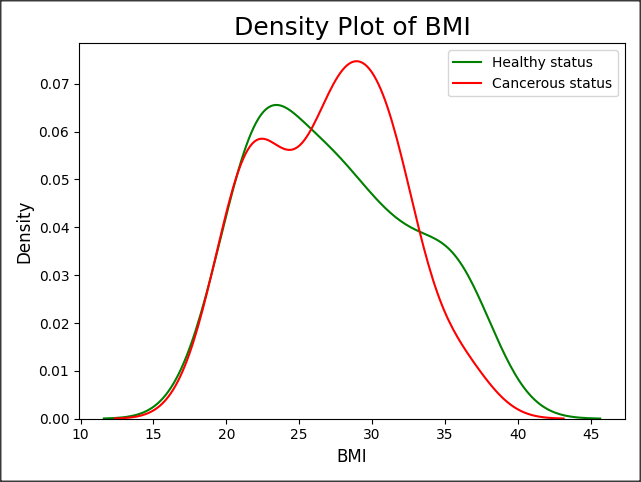
Seaborn is used in this function (shown in figure 10), which is a Python data visualization library based on Matplotlib. The boxplot and density plot are grouped by the categorical variables of ‘Status’, and these plots can be seen in the figure 11.

The boxplot provides significant information about the location and dispersion of the data, especially Figure 10: Function box\_density\_plots

when comparing different parent data. As can be seen in figure 11 in the x-axis contains the two classes (Cancerous and Healthy). Each boxplot contains six main data nodes. The Age data are arranged from largest to smallest and its upper edge, upper quartile, mean value, median, lower quartile, and lower edge are calculated.

The density diagram reflects the proportion of cancerous and healthy patients in the different BMI ranges. As can be seen from the graph below, the green curve represents Healthy Status and the red curve represents Cancerous Status, with the highest proportion of healthy patients in the BMI range of 20 to 25 and the highest proportion of cancerous patients between 25 and 33, which accounts for the highest proportion is over 7%.

Figure 11: The boxplot and density plot



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

I disagree with this view, the model is only as good as the scenario in which it is applied. Although in most research scenarios high precision is pursued, in the field of disease diagnosis a test set achieving 90% precision is not a good enough measure of how well a model works in a scenario. I will provide a detailed analysis and explanation of why I disagree with this view in the following points.

**Reference List**

Sola, J. and Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, 44(3), pp.1464–1468.