Data Exploration

Mingjun Xie 30971586 mx4n19@soton.ac.uk

Jie Zhou 31384579 [jz5n19@soton.ac.uk](mailto:jz5n19@soton.ac.uk)

Zixuan Cai 31303528 [zc1g19@soton.ac.uk](mailto:zc1g19@soton.ac.uk)

Chenguang Zhu 31369251 cz1g19@soton.ac.uk

ABSTRACT

The dataset we are going to explore is cardiovascular disease dataset. We will first develop the exploratory data analysis (EDA) to get an intuition and important information about the data sets. Then the main task for our project is to use these 70000+ data to predict whether testee faced with cardiovascular disease.

1 Browse the Data Structure

The first thing we need to do is to look at the data structure, we use head(), describe() and info() to print the profile of the data, The output is shown below.

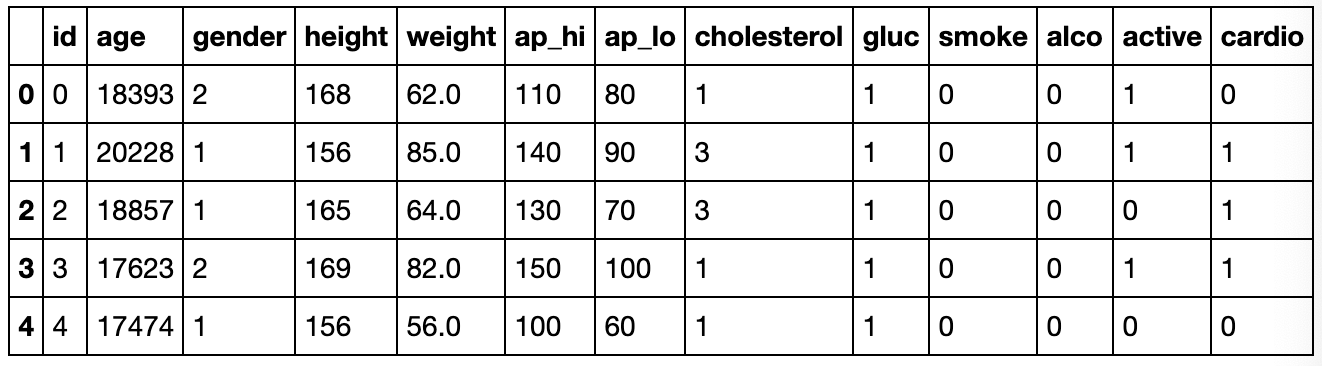


Figure 1: head()

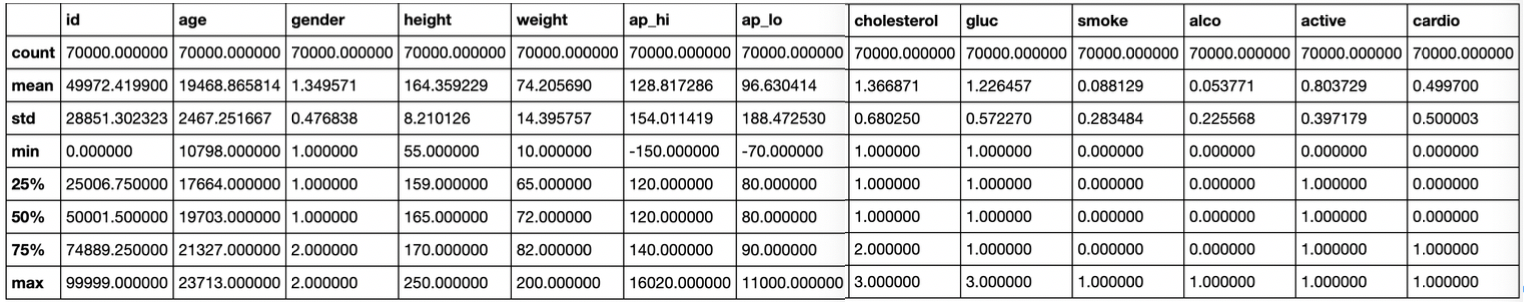


Figure 2: describe()

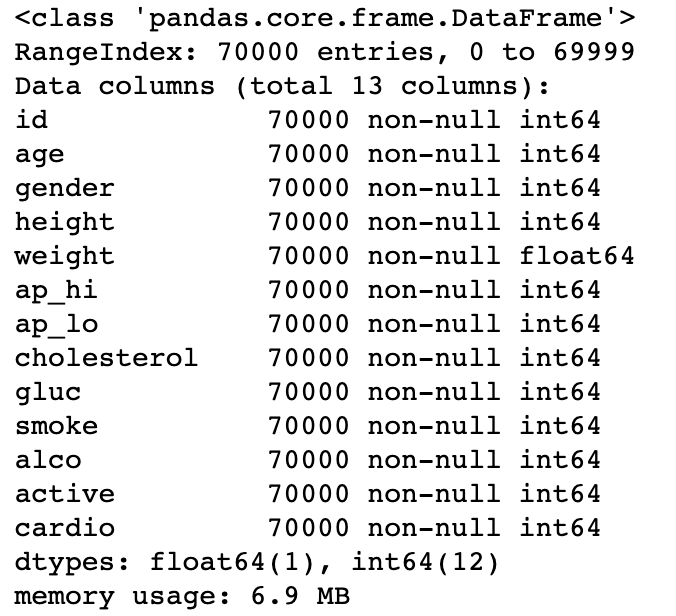


Figure 3: info()

From the Figure 1 and 3, we can see that the dataset consists of 70 000 records of patients data, 11 features + a target and there is no missing data and text data. As the description shown in Figure 2, there are some outliers in the data. The systolic blood pressure (ap\_hi) and diastolic blood pressure (ap\_lo) have negative number, their maximum numbers are also abnormal since the blood pressure can’t go such high like this. Additionally, the minimum number in weight is too small to be an adult’s weight. We will check these outliers and handle it later.

Then we make the histograms about each attribute, we find that the attributes have very different scales and some are tail heavy.

2 Split the training and testing dataset

C

3  Correlations of data

Computing the standard correlation coefficient between each pair of attributes can shows the correlations of data and the approximate contributions of a specific feature to predict the targets. We build a correlation heat map.

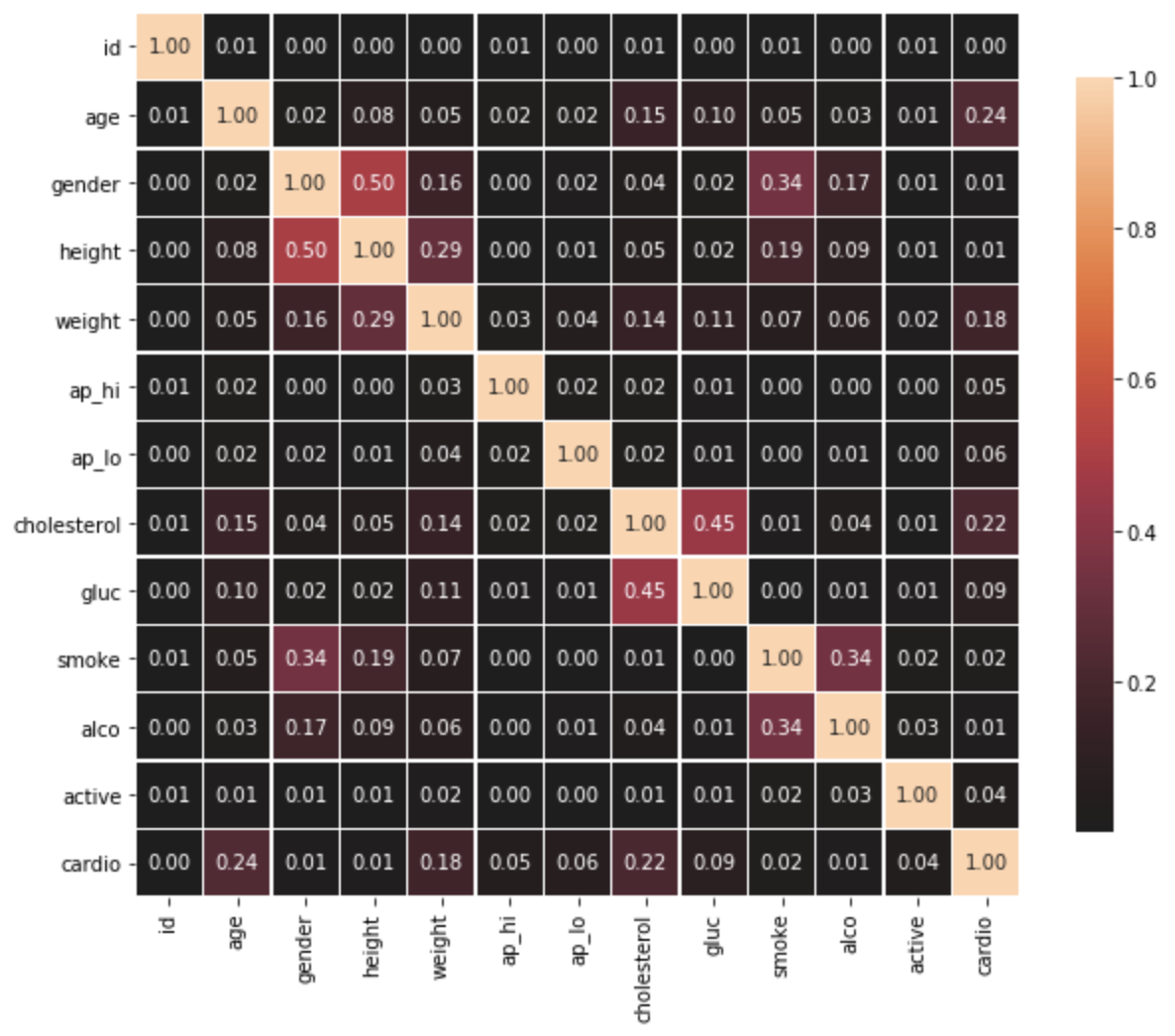


Figure 4: correlation heat map

In terms of the correlations of the data, we can find that the height, smoke and weight have a high correlation with gender, so we may need to consider to average the gender when splitting the training and testing dataset. The height also has high correlation with weight, so we can combine height and weight to a new attribute called BMI (Body Mass Index) which is weight / height2 kg/(m2). We can also combine the systolic blood pressure (ap\_hi) and diastolic blood pressure (ap\_lo) to a new attribute called MAP (Mean Arterial Pressure) which is 1/3 ap\_hi + 2/3 ap\_lo. After we add these two attributes to the data, we get a new correlation heat map like the figure shown below.

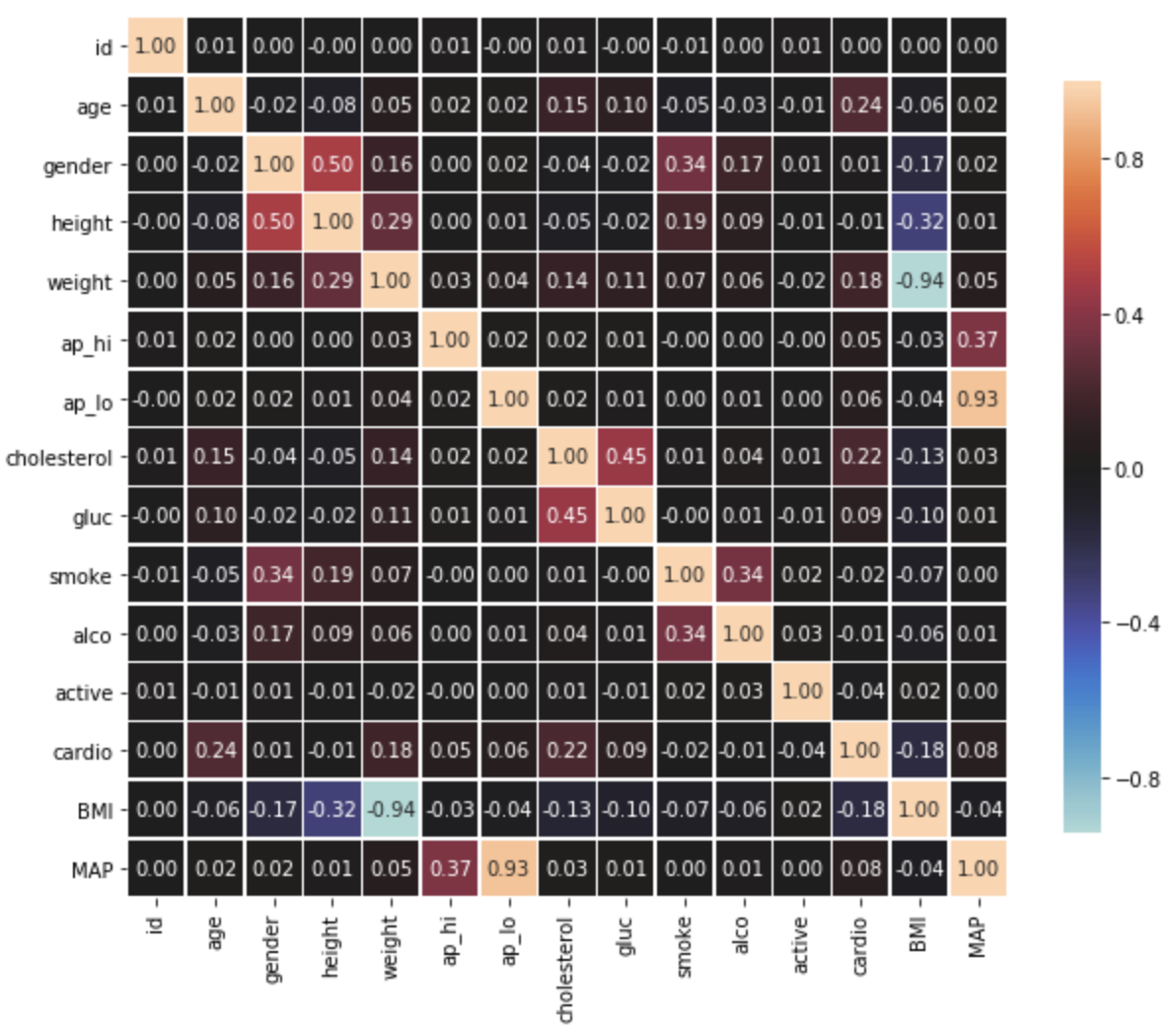


Figure 5: correlation heat map with BMI and MAP

From Figure 5, We can see that the correlations between BMI and cardio, MAP and cardio are not less than using height-weight and ap\_hi-ap\_lo combinations which means that using BMI and MAP will have at least equivalent performance on predicting the cardiovascular disease.

4  Handling Categorical Attributes

Since the dataset doesn’t have missing data and text data, so we just need to handle the categorical attribute. Luckily, we just have gender attributes to handle because male and female are originally represented by 1 and 2 which have magnitude difference while there is no magnitude difference between male and female. So, we use one-hot encoder to represent gender.

5  Data Cleaning

REFERENCES

[1] Osman Yucel, Jon Hoffman and Sandip Sen, Jonny black: A mediating approach to multilateral negotiations. Studies in Computational Intelligence, ISSN: 1860-949X, Vol: 674, 2017, Page: 231-238

[2] Dimitrios Tsimpoukis, Tim Baarslag, Michael Kaisers, Nikolaos G. Paterakis, Automated Negotiations Under User Preference Uncertainty: A Linear Programming Approach

<https://link.springer.com/chapter/10.1007/978-3-030-17294-7_9>

[3] Shogo Kawaguchi, Katsuhide Fujita, and Takayuki Ito, AgentK: Compromising Strategy Based on Estimated Maximum Utility for Automated Negotiating Agents

<https://link.springer.com/content/pdf/10.1007%2F978-3-642-24696-8_8>

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00