PART 3

OUTLIERS

In order to identify erroneously existing values in the original data we have generated two boxplots (fig.boxplots) where we have plotted the raw data distribution(left) and their standardized counterparts (right). Upon first inspection, there has been a considerable amount of values for a number of attributes that, at first, appeared to be outliers. As a result, we have searched the web for every attribute in our dataset. This way, we were able to identify the expected, upper tolerable and compatible to life physiological values that in retrospect invalidated that the outreaching ones in our data, are either part of the naturally occurring human diversity or simply feasible.

DISTRIBUTION

What is though a common feature of most attributes of clinical interest, is that the data appear to be normally distributed. Such features, characterized by differential skewness in each attribute, including alcohol use, systolic blood pressure, adiposity, LDL blood level, obesity and the type A behavior.

ORIGINAL DATA CORRELATION

What’s more, the variables in the original dataset appear to be correlated. We have generated a correlation heatmap (fig.heatmap) that demonstrates this and we have included a plot (fig.plot) supporting clearly existing trends. In our heatmap we were able to identify strong correlation between the age of the participants and the level of adiposity and obesity. This trend is also evident by the slope in the provided graph. In our heatmap, CHD becomes more frequent as age progresses as well, as is the case with systolic blood pressure. A reassuring factor of the validity of the data is the substantial correlation between obesity, adiposity and the LDL levels, since their co-existence is well supported in the literature as risk disease factors.

PRIMARY ML AIM FEASIBILITY

In the PCA analysis we have conducted, we were able to identify that there are attributes that appear to be drivers of variance. Specifically, in the fig.(PCA accumulation) the first two components are contributing 45% of this variance, while the ongoing addition of components increases the variance retained. Moreover, these complementary components, PC 3 & 4, allow for a clearer separation of the data points, as it is evident in the fig (separation). Consequently, based on the decomposition of the principal components(fig.coefficients), some attributes can be highlighted as the main drivers of variance, such as the family history and the type A behavior. After extensive analysis, we believe they could hold predictive power and in the view of fig. (separation)clear clustering that the participants’ family history demonstrably achieves, we are confident that the Primary Machine Learning Aim (PMLA) is feasible and supposedly it will be marked by satisfactory predictive accuracy. As our PMLA we have chosen to investigate which possible sets of characteristics arise as having the greatest predictive power for a patient to be diagnosed with CHD, using clustering algorithms.

PCA

VARIATION EXPLAINED

The fraction of the variance explained by the first two components is 45% which is sufficient to move on with the PMLA.

SUMMARY

The goal of this analysis was to determine how the various risk factors of the Coronary Heart Disease dataset correlate to each other and to examine the feasibility of any further manipulation of data towards the PMLA that could elucidate uncovered relationships. We were able to scatter the data into groups and perform dimensionality reduction in order to identify the major drivers of variance. Our dataset had no missing values, presented zero corruption elements or noise and it contained consistent information on the matter in question. What appeared at first to be outliers, after a thorough internet search, it was clear that these values were part of the natural diversity of the attributes. Most importantly though, the dataset is characterized by the presence of attributes that correlate well and in an epidemiologically sound manner. For example, there seems to be a connection between the Coronary Heart Disease (CHD) incidences and physiological attributes such as obesity and adiposity, where older participants exhibit greater possibility of being diagnosed with CHD. Family history appears to be a diagnosis factor of consideration, as it could provide valuable insights as a component of our clustering algorithm analysis in the future, the feasibility of which is supported by the scattering analysis we have performed in this project.

Type A:

<https://www.ncbi.nlm.nih.gov/pubmed/7352459>

Obesity

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5463032/>