PART 3

OUTLIERS

In the original data, only a small number of outliers appear as isolated cases of alcohol use, systolic blood pressure and tobacco use attributes.

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 appear to be correlated. We have generated a number of correlation plots that are color-coded for the actual appearance of Coronary Heart Disease (CHD) and these findings are summarized as follows: i) The LDL-Type A graph show that persons that exhibited the Type A behavior tended to have a lower LDL concentration in the blood. This is to be expected since Type A behavior is notably associated with greater daily activity which is documented as being an LDL decreasing factor, ii) The Age-Type A graph shows that regardless of the overwhelming majority of the young participants correlate to a Type A behavior, the CHD incidences distribution seem to be unaffected by this, iii) The Age-Obesity graph describes though that obesity increase with age, while most of the young participants suffered from CHD regardless of the lower obesity score the have gotten. This may indicate that younger persons are more susceptible to the effects of obesity, iv) The above observation also correlates well and it is further elaborated by the Age-Adiposity graph. Again, we see that adiposity increases with age but strikingly it also matches with the CHD affected participants even better. Which could possibly mean, based on this data only, that adiposity is an even more important driving force of the CHD occurrence than obesity in general.

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 more than 40% of this variance, while the ongoing addition of components increases the variance retained. Moreover, these complementary components PC 3 & 5 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.

(Talk about clustering in the future based on trends and scattering)

PCA

VARIATION EXPLAINED

The fraction of the variance explained by the first two components is 43% which is sufficient to move on with the PMLA.

The principal directions of the PCA components are:

;

Stan adding the table of principal directions

;

Thanos creates the 3d plot and write a description

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. 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.