Cardiovascular Disease

Data descriptive analysis

The training data has 69310 rows and 13 columns. The attributes or features is descripted In table 1. There are 5 binary attributes(smoking, alchohol, glucose, and active, gender). 5 continuos attributes. Categrical attributes in the data are gender, cholestrol, Glucose, smoking, alchohol and active. Out of which only gender is nominal and all others are ordinal data.

#	Attribute	Description	Range
1	Age	Age in days	0 to inf
2	Gender	1 for male and 0 for female	0 or 1
3	Height	height in cm	o to sensible
4	Weihgt	weight in kgs	0 to sensible
5	Systoloc blood pressure	High pressure	between 100 to 150 normally
6	Diastoloc blood pressure	Low blood pressure	80 to 130 normally
7	Cholestrol	cholestrol level on a scale of 0 to 3	0 to 3
8	Glucose	GLucose level from 0 to 3	0 to 3
9	Smoking	smooing binary	0 or 1
10	Alchohol	consumption of alchohol binary	0 or 1
11	active	physically active binary	0 or 1

Table 1

Heatmap of null values are shown in the fig 1

The heat map on right shows that training data is clean and doesnot have any null values included. Data doesnot need any cleaning and deletion of rows.

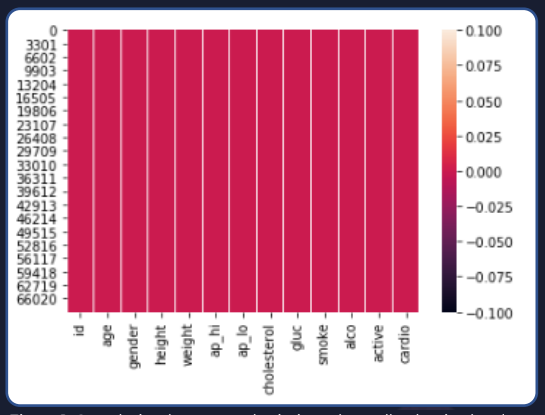
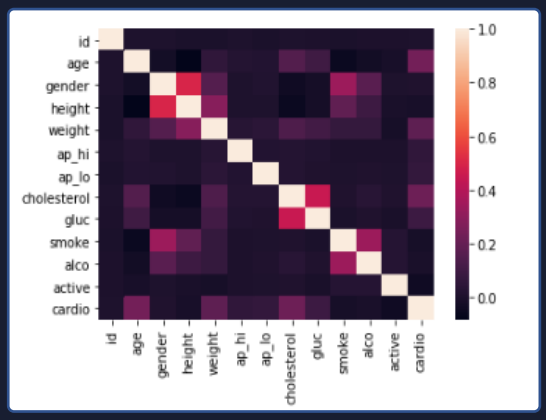


Figure1. Correlation heatmap depicting the null value in the data matrix

Data has column cardio which is the target column to be predicted. Correlation heat map is of the data is shown in the fig 2.



From the correlation map, we can see that there is no substantial correlation between the data. And therefore simple regression models wont fit the data. An ensemble model of fitting the data is being used to build the model. Features that show noticeable correlation with cardio attribute are “age”, “weight”, “cholesterol”, and “glucose” to some level, which make sense.

An xgboost Classifier is being used to train the model. Owing to low correlation between the data we classifier will be non-linear and we need to train the model with high number of estimators or trees. 200 estimators or trees are used in the model. Booster used for model is gradient boosting tree model. The data found the optimal loss value at considerably large learning rate 0.2. The objective function used to “binary:logistic” model of xgbclasssifer model since it is a binary classifier.

Trained model has an accuracy of 82.7%. and the confusion matrix for the validation matrix is shown in the table 2

3117	1332
1332	8356

Confusion matrix