CMPE 257 - LAB1

Suriya Palanikumar

Introduction

The main focus of CMPE 257 LAB1 is to predict a classification problem. The dataset has several independent features like Smoking habits, Systolic Pressure etc., Using those independent variables, a classification has to be made if a person is likely to receive cardiac Arrest.

About the tasks and Dataset

CMPE 257 LAB1 has 3 tasks

1. Task 1 - Classification problems with Visualizations as a Kaggle Competition
2. Task 2 - Testing the prediction in Cardio_complete file
3. Task 3 - Testing the polynomial features

1) Task 1:

Datasets used in Task 1 are Cardio_train.csv, Cardio_validation.csv and Cardio_test.csv

Training & Validation Dataset : Cardio_train.csv and Cardio_validation.csv
Test dataset : Cardio_validation.csv

Training and Validation Dataset:(cardio_train.csv & cardio_validation.csv)

Attribute Name	DataType	Attribute Type
ID	Int	-
age(in days)	float	Independent
gender	object	Independent
height	float	Independent
weight	float	Independent
ap_hi	float	Independent
ap_lo	float	Independent
cholesterol	object	Independent
gluc	object	Independent
smoke	float	Independent
alco	float	Independent
active	float	Independent
cardio	int	Dependent

Cardio_test has same variables and dataset as training set except the predictive variable ie., cardio

2) Task 2

Task 2 uses cardio_complete.csv which has the same variables and dataset as the training set of task 1.

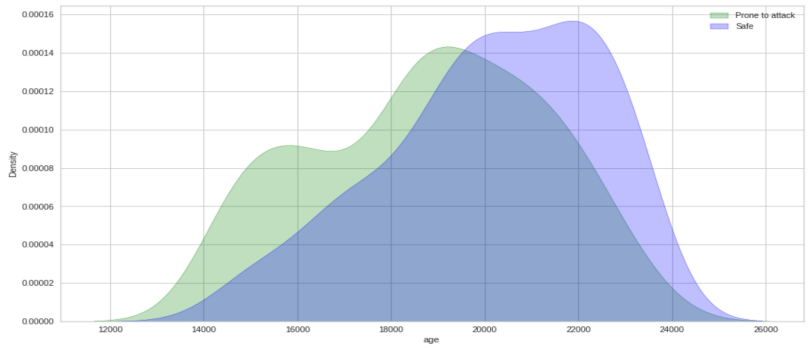
3) Task 3

Task 3 uses the datasets same as task 1.

Inference about predictive variable and other attributes

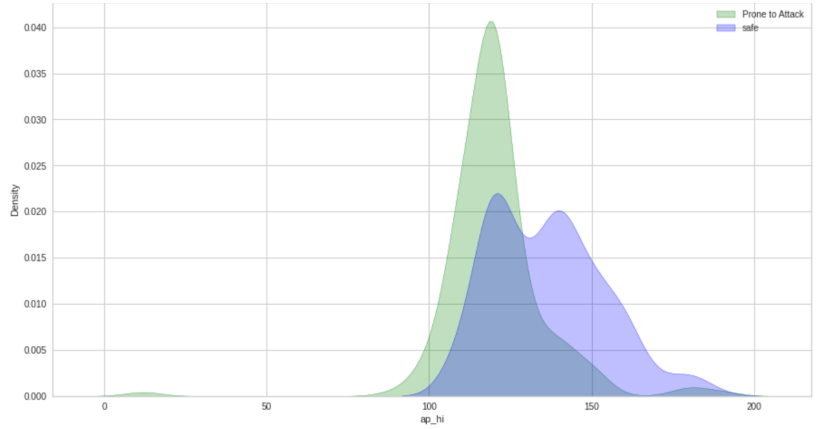
There are few attributes in the dataset which are related to the predicted variable in terms of distribution. We could visualize it with kdeplots.

a) Relationship between cardio and Age



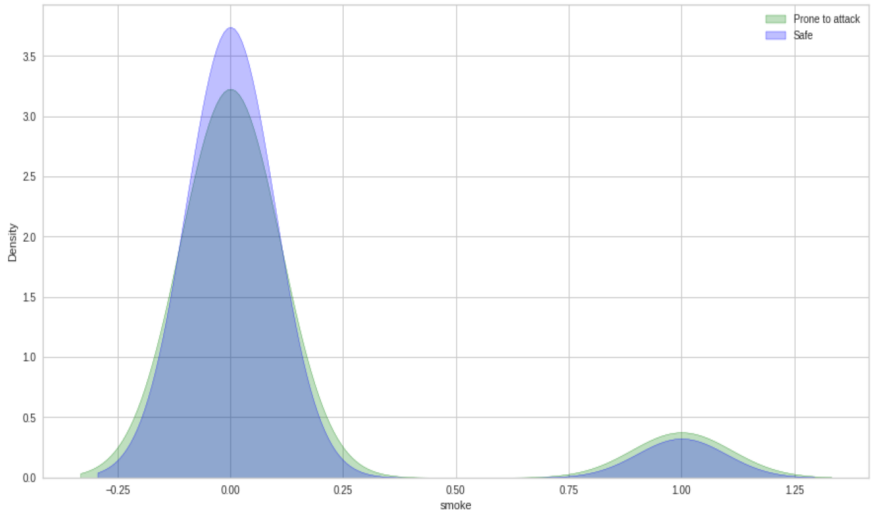
From the Correlation heatmap and the kdeplot, we could see that people "prone to cardiac attack" is higher around the age of 19000 days.
Irrespective of the age, cardio attack may or may not be possible, which indicates other attributes along with age would influence the
predictive variable. ie., cardiac

b) Relationship between cardio and ap_hi(Systolic BP)



From the Correlation heatmap and the kdeplot, we could infer that ap_hi has a huge impact of a person getting cardio attack or not.
From the plot, it is evident that ap_hi is definitely one of the best feature to predict cardio.

c) Relationship between cardio and smoke



From the Correlation heatmap and the kdeplot, we could see that the relationship between smoke and cardio is uncertain.

"Safe" and "Prone to Cardio" are common irrespective of smoking rates here.

One inference is that both "Prone to Cardio" and "Safe" is high when smoke=0

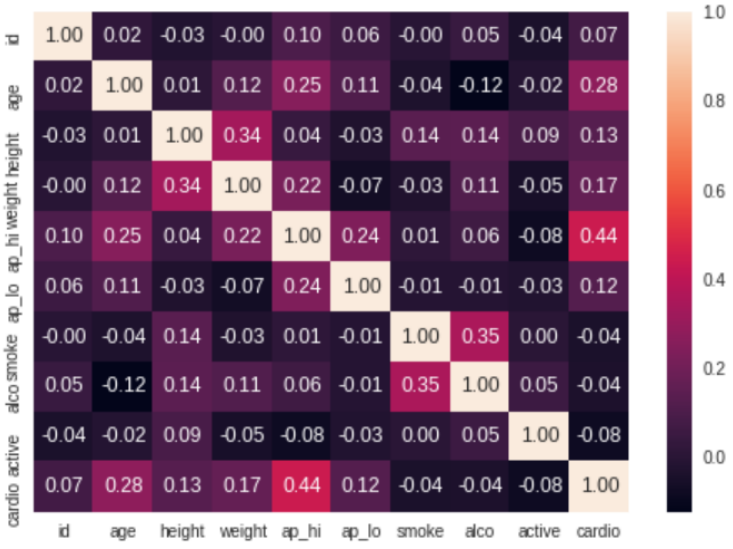
Data Preprocessing

1) Categorical attributes are replaced with mode and used LabelEncoder
2) NaN values are replaced with column mean.

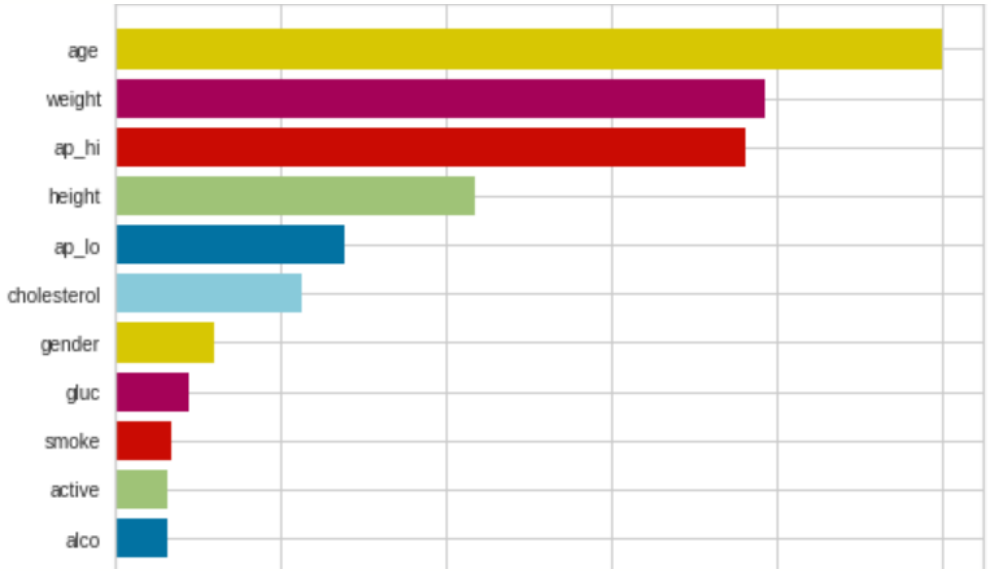
Data Visualization

Some inferences that could be made out of the dataset are as follows:

1) Correlation Heatmap



2) Feature Importance



ML Algorithms Used

1) RandomForestClassifier and Decision Tree in pipeline

Pipeline feature has been tested alongside Random Forest and Decision Tree, in which Random Forest is the best fit with a mean squared error of 32% and accuracy of 0.68

2) Support Vector Machine - SVC

SVC in SVM gave the mean squared error of 43% and accuracy score of 0.57

3) Gradient Boosting

GradientBoosting gave the mean squared error of 27% and accuracy score of 0.73

4) XGBClassifier

XGBClassifier gave the mean squared error of 28.2% and accuracy score of 0.718

5) RidgeClassifier

RidgeClassifier gave the mean squared error of 28.6% and accuracy score of 0.714

6) Xgbregressor(Model Submitted in Kaggle)

Since the metric used in Kaggle competition is RMSE, Regressor provided better results than other classifier techniques.

ML algorithm using numpy

1) Implemented Logistic regression using numpy in python

Comparing results of Task 1, 2 and 3

1) Task 1 has the accuracy of 0.73 with xgbRegressor
2) Task 2 has the accuracy of 0.74 with RidgeClassifier
3) In Task 3, I could infer that there is a slight overfitting in the model though we use ridgeclassifier .