Saurabh Rajeshkumar Panchal

Student ID: 015219971

**CMPE 257 Lab 1**

**Task 1**

* Task 1 comprised of performing some EDA (Exploratory Data Analysis) on the given cardiovascular data.
* There are 13 columns in our dataset. The Kaggle page describes the meaning of each column like this:

id -----> id

age -----> Age(days)

gender -----> Gender

height -----> Height(cm)

weight -----> Weight(kg)

ap\_hi -----> Systolic blood pressure

ap\_lo -----> Diastolic blood pressure

cholesterol -----> Cholesterol

gluc -----> Glucose

smoke -----> Smoking

alco -----> Alcohol intake

active -----> Physical activity

cardio -----> Presence(1) or absence(0) of cardiovascular disease (Target Variable)

* The given features can be divided into 3 categories:
  1. Nominal (Categorical) features : gender, cholesterol, gluc, smoke, alco, active
  2. Numeric (Continuous) features : age, height, weight, ap\_hi, ap\_lo
  3. Target variable : cardio
* Null values: We can see presence of NaN(missing) values in all columns except 'id' and 'cardio'. Most NaN values are in 'height' column as non-null count is very low.
* Data types: Float values are present in 'age', 'height', 'weight', 'ap\_hi', 'ap\_lo', 'smoke', 'alco', and 'active'. These columns are numeric and continuous. Int values are present in 'id' and 'cardio' columns. For columns 'gender', 'cholesterol' and 'gluc' we see data type as object because these are categorical features with string values.
* Age analysis:

age is given in days, we can convert it to years by age = age/365

As per results of Q4:

a. Oldest person in the data is : 64.0 years

b. Youngest person in the data is : 39.0 years

c. Average age of a person in the data is : 53.4 years

d. Median age of a person in the data is : 54.0 years

Relation of age vs cardio:

Chart, histogram

Description automatically generated

As we can see from the density plot, age density distribution of people with cardio disease is on the right side of density distribution of people with no cardio disease. This suggests that people with higher age(>55) have more risk for cardio disease than people with lesser age(<45).

Dividing records into age groups of "30-35", "35-40", "40-45", "45-50", "50-55", "55-60", "60-65", "65-70", "70-75":

Q4f gives an idea of how many people survived in each age group.

creating bins of 335 people ( with non-null age)

In age group 30-35 , total people: 0 , people with no cardio: 0 , survival rate: 0

In age group 35-40 , total people: 1 , people with no cardio: 1 , survival rate: 1.0

In age group 40-45 , total people: 48 , people with no cardio: 36 , survival rate: 0.75

In age group 45-50 , total people: 37 , people with no cardio: 18 , survival rate: 0.49

In age group 50-55 , total people: 96 , people with no cardio: 51 , survival rate: 0.53

In age group 55-60 , total people: 73 , people with no cardio: 34 , survival rate: 0.47

In age group 60-65 , total people: 80 , people with no cardio: 22 , survival rate: 0.28

In age group 65-70 , total people: 0 , people with no cardio: 0 , survival rate: 0

In age group 70-75 , total people: 0 , people with no cardio: 0 , survival rate: 0

Distribution of people in each age group:

People can be divided into bins from age 35-40 to 60-65 as we have people from age 39 to 64 years in our data. We have highest count of people from age group 50-55. 75% of people from age group 40-45 (36 out of 38) have cardio=0, which makes it age group with highest survival rate.

Chart, bar chart

Description automatically generated

* Processing categorical columns:

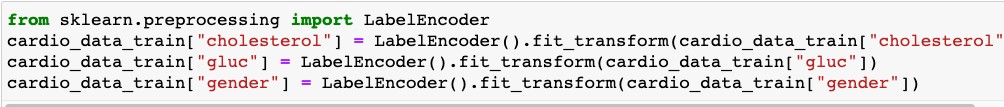
Cholesterol, gluc and gender are categorical columns in our data. We can user LabelEncoder to replace their string values by integer.

cholesterol unique values: [nan 'Normal' 'High' 'Above Normal']

glucose unique values: ['Normal' nan 'High' 'Above Normal']

gender unique values: [nan 'Men' 'Women']

Example of LabelEncoder:



Alternatively, we can use pandas map lambda function to replace each “Men” with 0, “Women” with 1 and in similar manner replace “Normal” with 0, “Above Normal” with 1, and “High” with 2.

* Graph of distribution of records with cardio = 0 and cardio = 1:

Chart, bar chart

Description automatically generated

* “id” column is not needed for prediction task since its just identifier for each person and is unique for every record. We can drop this column before feeding training data to classifier.
* For further questions (density plots), we replace numeric columns with mean of that column and replace categorical columns with mode (most occurring value) of the column. This way we can visualize distribution on entire training data.
* In question 4g, we plot relationships between a few features against cardio. The graphs and observations are as mentioned below:

Age vs Cardio:

Chart, histogram

Description automatically generated

We can see in the kde plot that people arround the age 35-45 have higher chance of cardio=0. People around the age 50-55 have high survival rate as well as high fatality rate. People with age 57-70 have high chance of having cardio=1.

Ap\_hi vs cardio:

Chart, histogram

Description automatically generated

People with average Systolic blood pressure(ap\_hi) have higher chance of having cardio=0, i.e. there is higher chance that they don't have cardio disease.

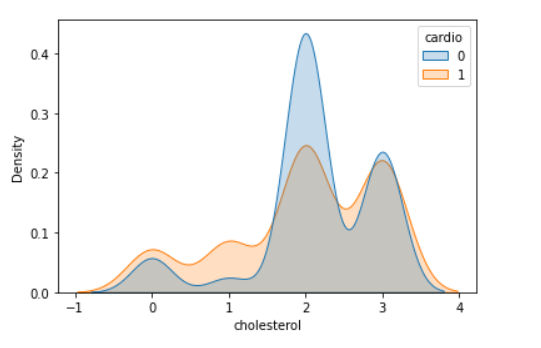
Ap\_lo vs cardio:

Chart, histogram

Description automatically generated

We can see from density plot of ap\_lo that people with average Diastolic blood pressure(ap\_lo) have very high probability of not having cardio disease.

Cholesterol vs cardio:



From density plot of cholesterol, we can see that people with Normal(0) cholesterol have higher chance of cardio=0. i.e. there is high chance that they will not have cardio disease, than people who have more cholesterol.

Gluc vs cardio:

Chart, histogram

Description automatically generated

In the "gluc" kde plot, we can see that, people who have Normal(0) glucose have higher chance of cardio=0, i.e there is a high chance that they wont have cardiovascular disease, than the people who have more glucose.

Active vs cardio:

Chart, histogram

Description automatically generated

In the "active" kde plot, we can see that, people who are more physically active have a slightly higher chance of cardio = 0, i.e there is a high chance that they won't have cardiovascular disease, than the people who have less physically active.

Height vs cardio:

Chart, histogram

Description automatically generated

In kde plot of height, the plot of cardio=0 and cardio=1 is roughly the same and from this we can also tell height column does not add much importance, so we can drop height column when we are building our model.

Weight vs cardio:

Chart

Description automatically generated

In kde plot of weight, the plot of cardio=0 and cardio=1 is roughly the same and from this we can also tell weight column does not add much importance, so we can drop weight column when we are building our model.

* PlotCorrelationMatrix – Plotting the correlation matrix, I can see correlations of each

column with itself and other columns (This is a symmetric matrix). From the correlation matrix I can see that, “ap\_hi”, “age” and “cholesterol” has the highest correlation with “cardio” when compared to other columns (shown below).

Timeline

Description automatically generated

* PlotScatterMatrix – Plotting the Scatter Matrix, we can see the scatter matrix of each column with other column and we can see the kde plot of each column too. (shown below).

A picture containing chart

Description automatically generated

* PlotPerColumnDistribution – plots the histograms of each columns

Chart, box and whisker chart

Description automatically generated

* Missing value imputation:

I am loading the data again to try out different missing data imputation techniques.

Just to recap, numbers of missing values in all columns are :

Text

Description automatically generated

Heatmap of missing values:

Plotting the heatmap for missing values, In this plot, I can see that, The more the number of missing values in the column, the lighter the column is, for example, column "id" and column "cardio" have zero missing values, so they have a full dark shade, meanwhile, height has the highest missing values, so it has a very light shade (shown below).

A picture containing chart

Description automatically generated

Different strategies of replacing NaN values:

All these strategies are evaluated by training and evaluating a LogisticRegression model.

i. Use dropna

Text

Description automatically generated with medium confidence

When I used dropna as an imputation strategy, I got Train accuracy as 1.0, Test

accuracy as 0.5 and f1 score as 0, by looking at these numbers, I can say that the model is overfitting

ii. Use replace NA with zero or max value

Text, letter

Description automatically generated with medium confidence

Replacing NaN with 0 gives test accuracy of 56%, which is improvement from dropna, but it is very bad accuracy for binary classification problem.

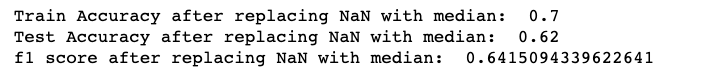
iii. Use replace NA with mean

Text

Description automatically generated with medium confidence

After replacing NaN values with mean of individual columns, we get test accuracy of 68% and f1 score of 69%. which is better result than previous 2 strategies. This method of replacement is better than dropna and replace with 0.

iv. Use replace NA with median



After replacing NaN values with median of individual columns, we get test accuracy of 62% and f1 score of 64%. Replacing NaN with mean is better strategy than replacing NaN with median.

v. Use replace NA with mode

Text

Description automatically generated with low confidence

After replacing NaN values with mode of individual columns, we get test accuracy of 71% and f1 score of 71%. Replacing NaN with mode is better strategy than replacing NaN with mean or median.

**Task 2**

**Task 3**