Capstone Project - 3

CARDIOVASCULAR RISK PREDICTION

BY

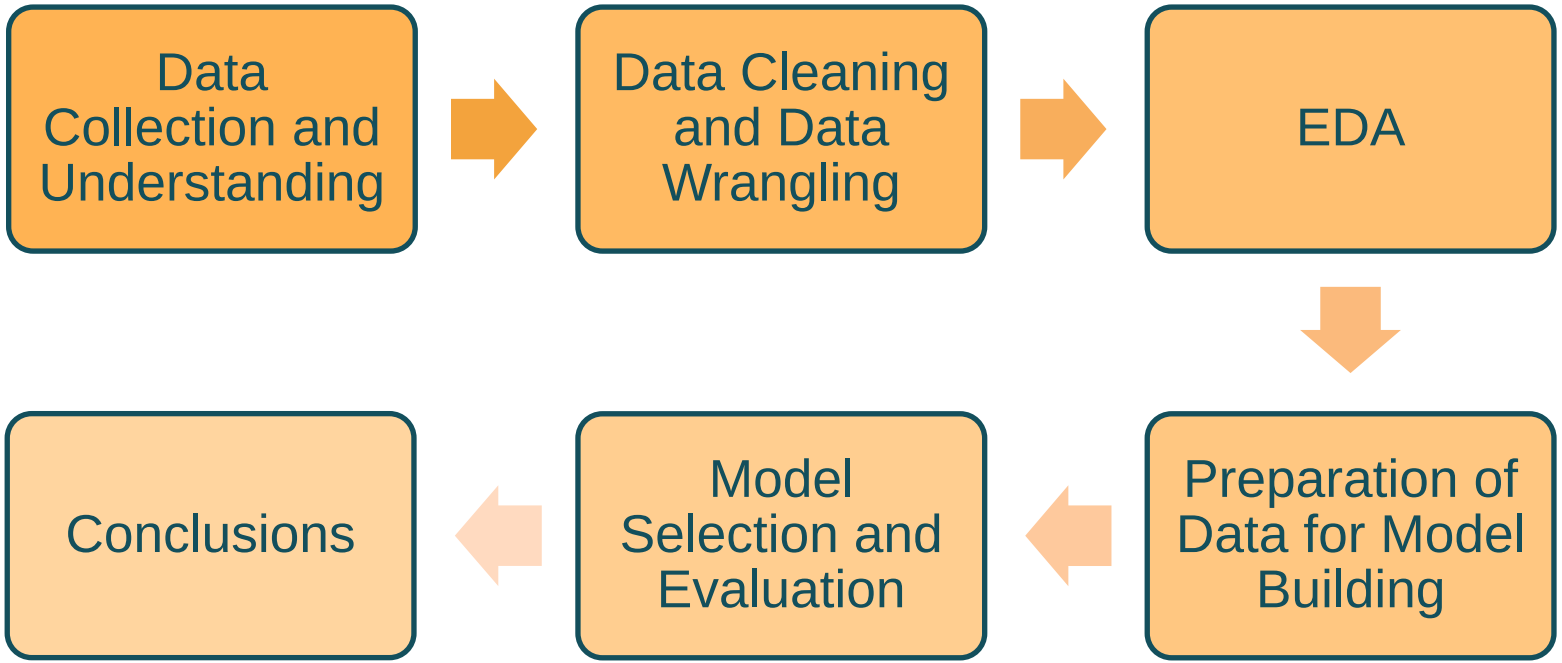
- Sumit Berde
- Omkar Desai

Problem Statement

- Cardiovascular disease(CVD's) are a group of life threatening diseases which takes millions of lives world wide.
- Early diagnosis can help in saving lives as it is said that prevention is better can cure. If our model is able to identify persons, who are at most risk of getting of CVD's then it would greatly help doctors to save lives by starting early diagnosis and reduce the chances of death.
- Our project aims at building ML models to detect high risk persons who are most vulnerable towards suffering from CVD's

Workflow

The steps involved are as follows



Data Collection And Understanding

- This project is based on medical domain dataset and it has been collected from various patients. It includes entries of over 3000 patients and has 17 different attributes for the prediction of Coronary Heart Disease in patients for the next ten years
- It includes 3390 records and 17 attributes. Variables each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors

Data Description

- Sex : male or female(‘M’ or ‘F’)
- Age : age of patient
- Is_smoking : whether or not the patient is a smoker(‘Yes’ or ‘No’)
- Cigs per day : the number of cigarettes that the person smoked on average in one day

- Bp meds : whether or not the patient was on blood pressure medication(Yes / No)
- Prevalent Stroke : whether or not the patient previously had a stroke(Yes / No)
- Prevalent Hyp : whether or not the patient was hypertensive(Yes / No)
- Diabetes : whether or not the patient had diabetes (Yes / No)
- Tot Chol : total cholesterol level
- Sys BP : systolic blood pressure
- Día BP : diastolic blood pressure
- BMI : Body Mass Index
- Heart rate : heart rate
- Glucose : glucose level
- Ten year chd : 10 year risk of coronary heart disease(Target variable)

Data Cleaning and Data Wrangling

Categorical features : is_smoking, sex, bp_meds, prevalent_hyp, prevalent_stroke. diabetes
Numerical features : age, cigs_per_day, total_cholesterol, systolic_bp, diastolic_bp, bmi, heart rate, glucose

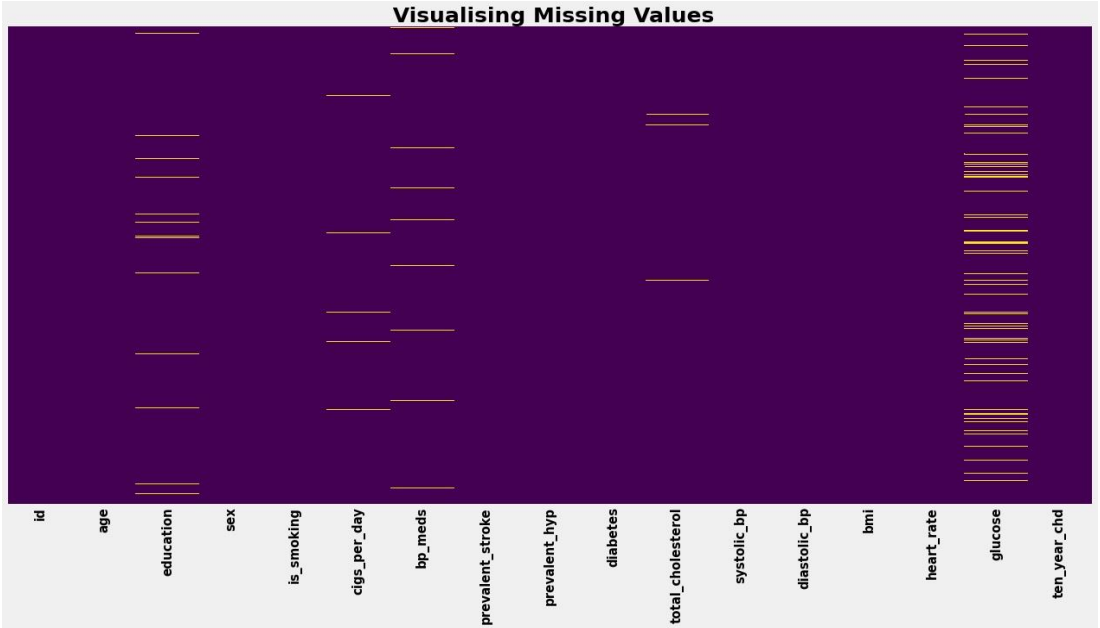
1. Rename columns: We have renamed columns to give appropriate name to features and removed spaces in-between

```
Index(['id', 'age', 'education', 'sex', 'is_smoking', 'cigs_per_day',
       'bp_meds', 'prevalent_stroke', 'prevalent_hyp', 'diabetes',
       'total_cholesterol', 'systolic_bp', 'diastolic_bp', 'bmi', 'heart_rate',
       'glucose', 'ten_year_chd'],
      dtype='object')
```

2. Checking for duplicate rows : We had zero duplicate rows in our dataset

3. Checking for missing values : Missing values were present in education, cigs_per_day, bp_meds, total_cholesterol, bmi, heart_rate, glucose

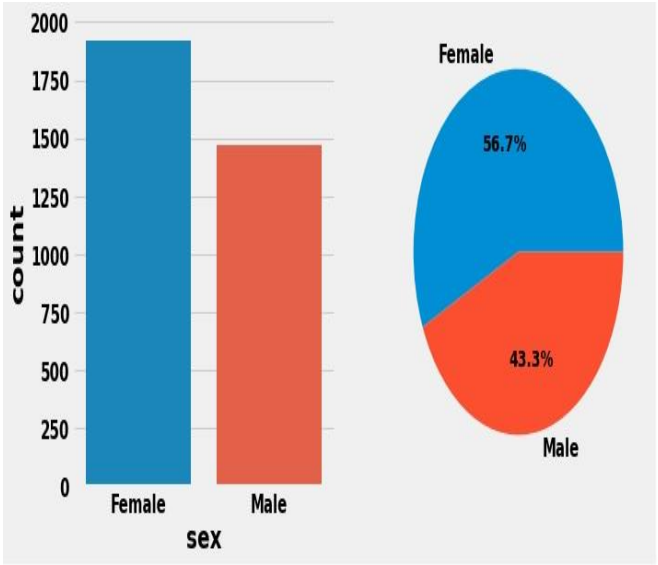
Missing value count	
id	0
age	0
education	87
sex	0
is_smoking	0
cigs_per_day	22
bp_meds	44
prevalent_stroke	0
prevalent_hyp	0
diabetes	0
total_cholesterol	38
systolic_bp	0
diastolic_bp	0
bmi	14
heart_rate	1
glucose	304
ten_year_chd	0

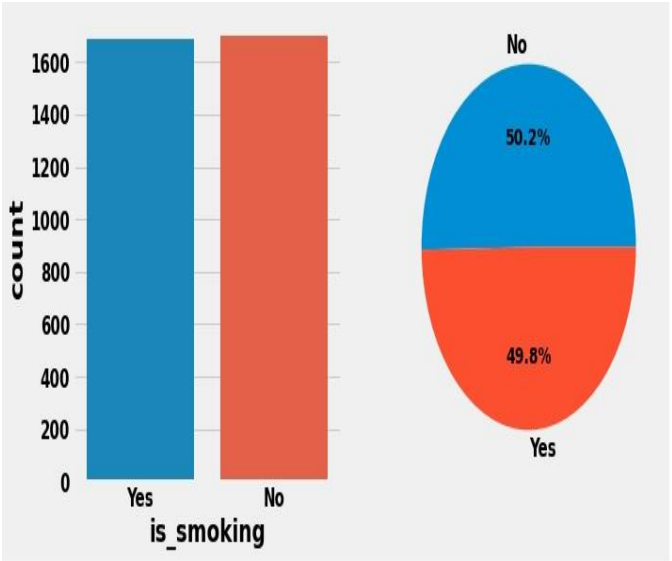


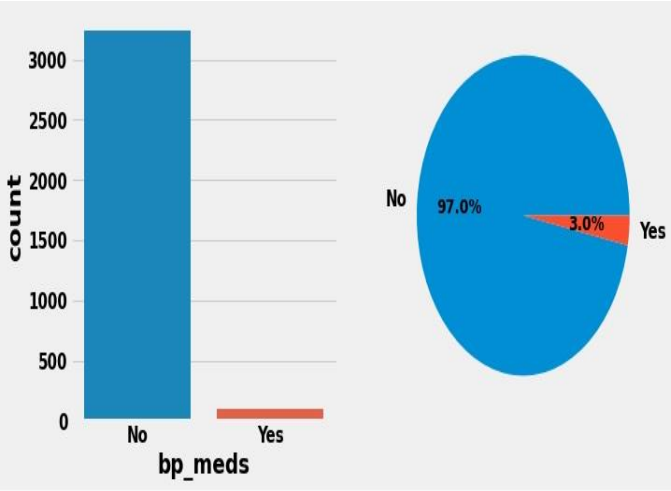
EDA(Exploratory Data Analysis)

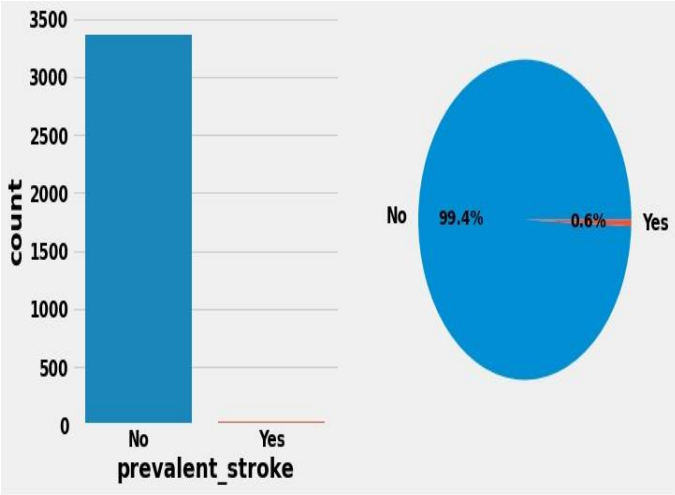
Univariate Analysis

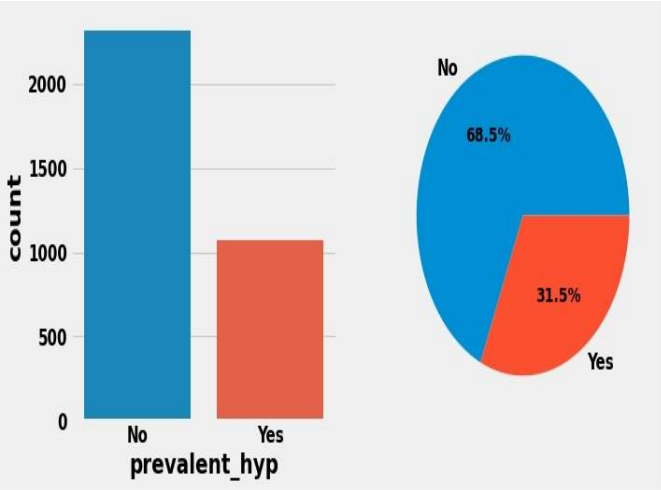
Frequency Distribution in Categorical Features

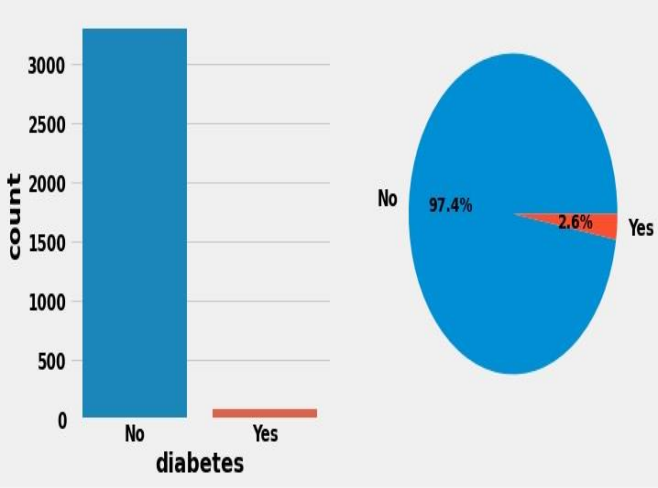




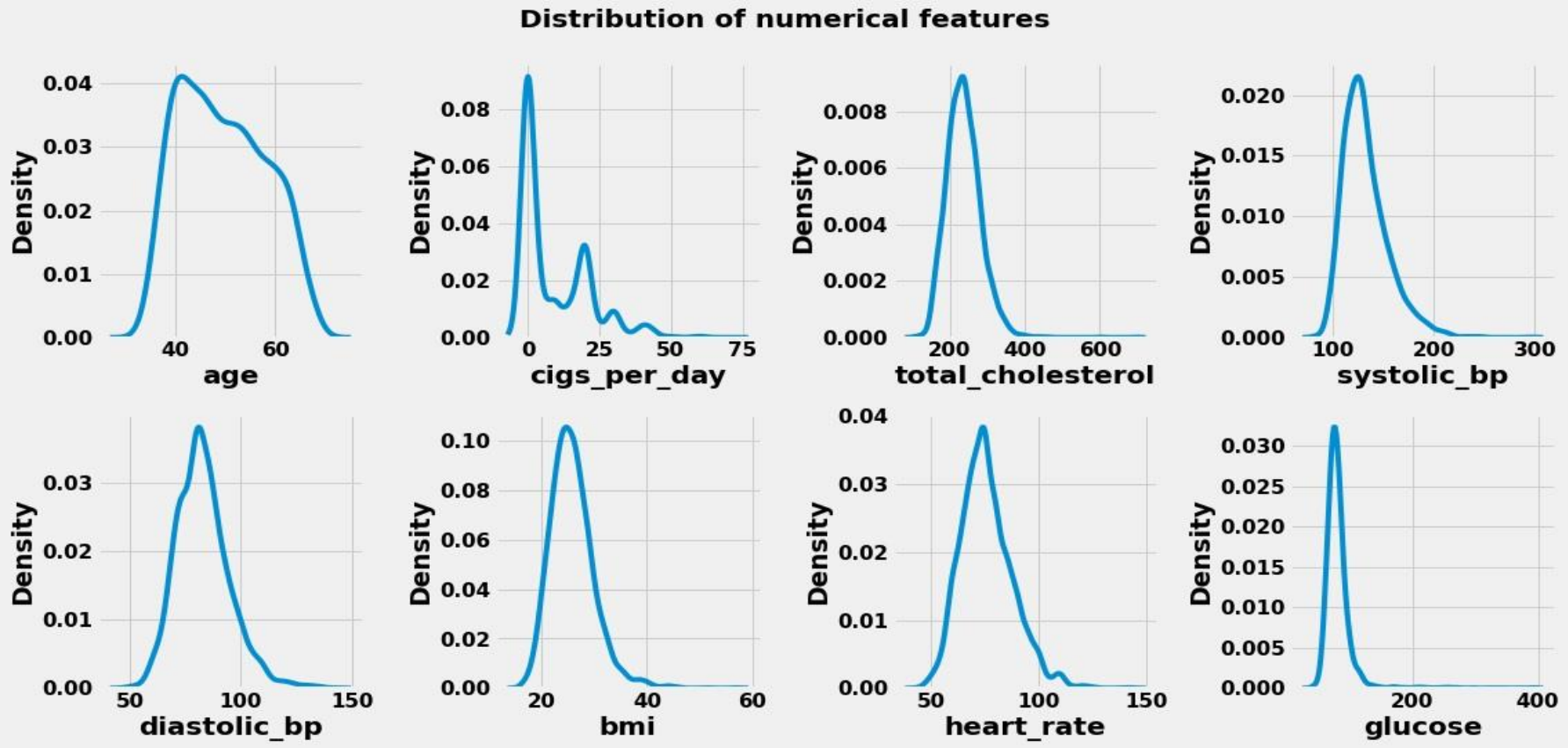

3000
2500
2000
1500
1000
500
0
count
No
Yes
bp_meds
No
97.0%
3.0%
Yes


3500
3000
2500
2000
1500
1000
500
0
count
No
Yes
prevalent_stroke
No
99.4%
0.6%
Yes

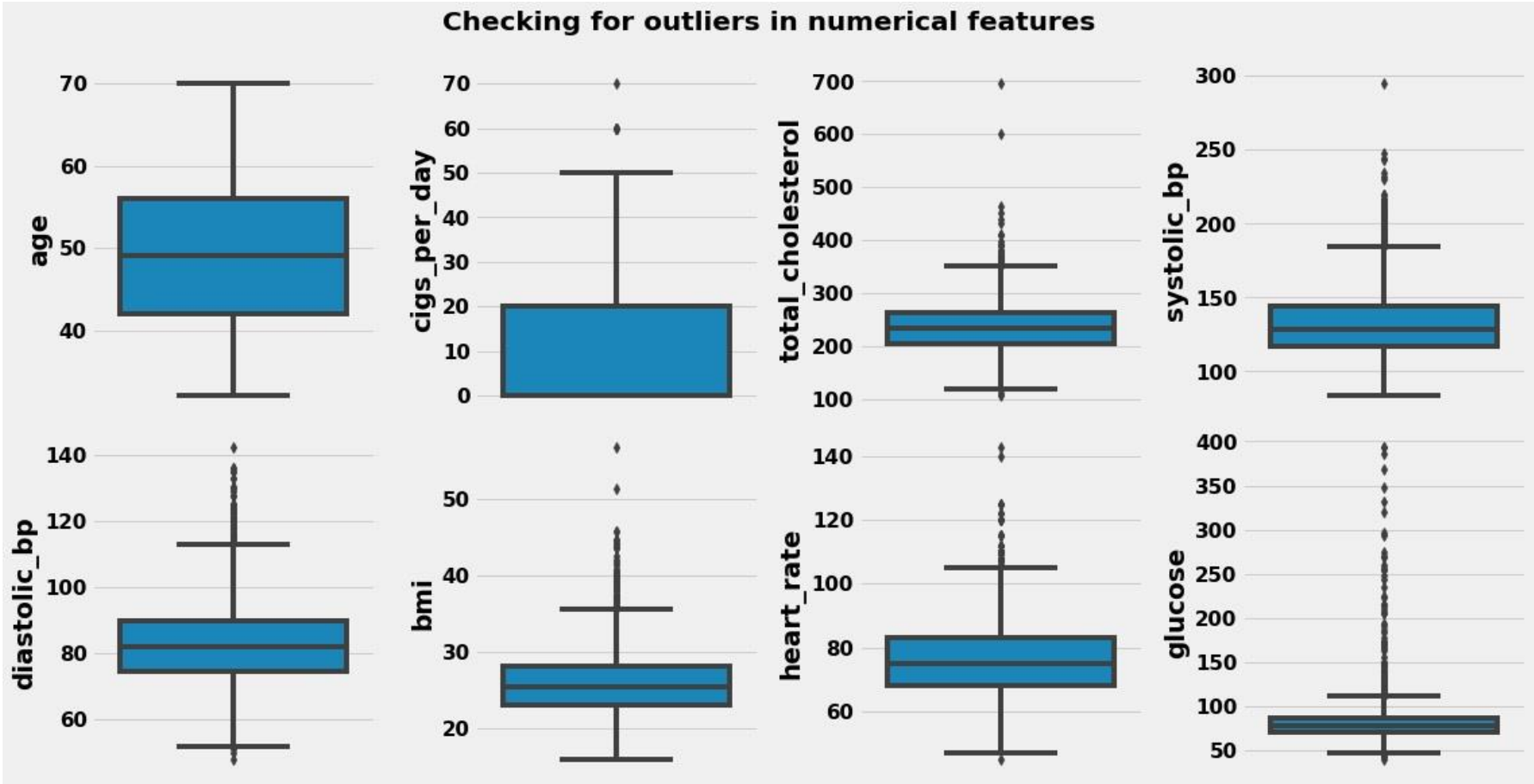

2000
1500
1000
500
0
count
No
Yes
prevalent_hyp
No
68.5%
31.5%
Yes


3000
2500
2000
1500
1000
500
0
count
No
Yes
diabetes
No
97.4%
2.6%
Yes

Checking distribution of data in numerical features

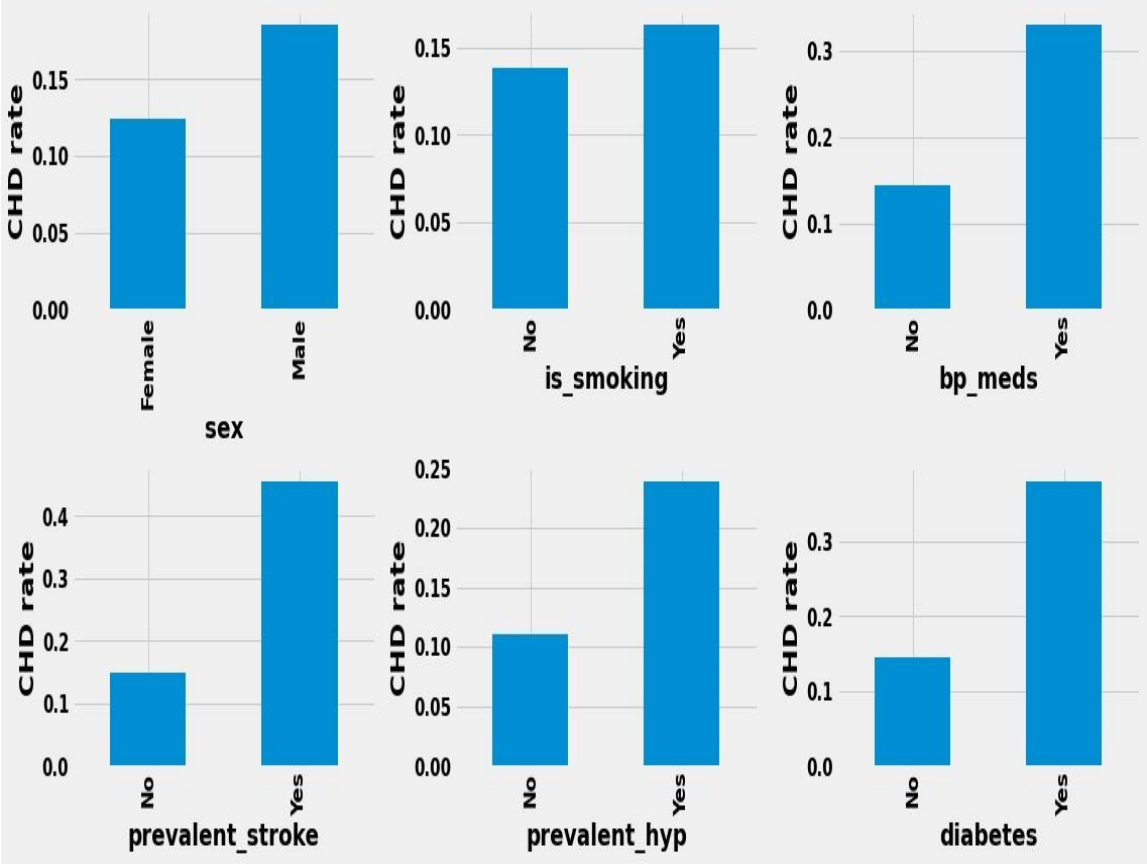


Checking for outliers in numeric features



Bivariate and Multivariate Analysis

Rate of people having cardiovascular disease within categories

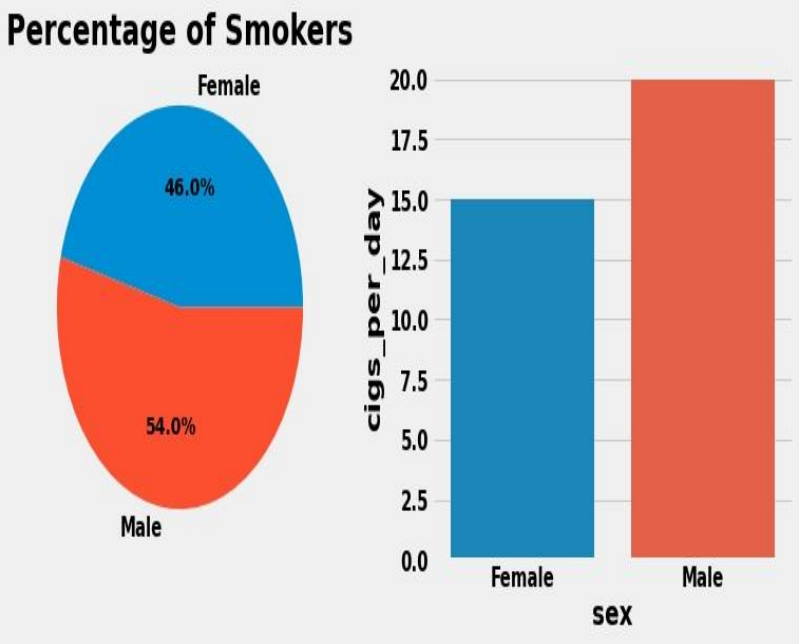


Observations:

- Males seems to have higher rate of contracting heart disease than females
- Smoking increases the chances of getting heart disease as more individuals who smoke seems to suffer from heart disease
- It looks like if someone has premedical conditions such as blood pressure medication, prevalent stroke, prevalent hypertension and diabetes his chances of getting heart disease increases

Now lets do a detailed study about how various factors like **smoking** and pre-medical conditions like **blood pressure medication**, **prevalent stroke**, **hypertension**, **diabetes** effects your chances of contracting heart disease

- Percentage of male and female smokers and then finding the average cigarettes smoked by males and females



Observations:

- Out of all the smokers, 54% are males and 46% are females
- Males on an average smoke more than females

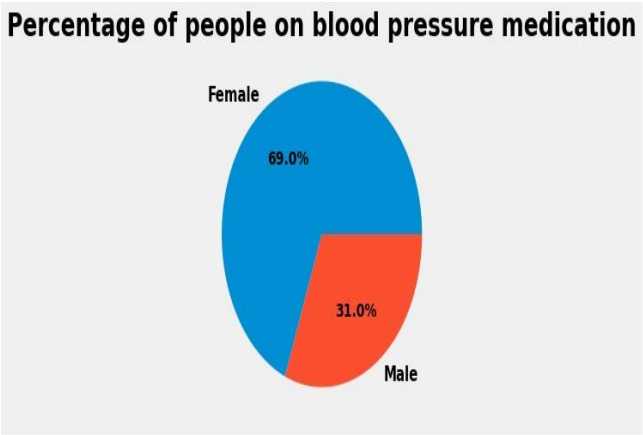
Percentage of smokers and non-smokers in males/females, who developed heart disease

sex	is_smoking	0	1
Female	No	87.18	12.82
	Yes	88.14	11.86
Male	No	83.99	16.01
	Yes	79.91	20.09

Observations:

- In males, smoking seems to have increased the chances of getting heart disease.
  Out of all male smokers, 20.09% of them eventually suffered from heart disease.
  As compared to male nonsmokers only 16.01% of them suffered from heart disease.
  There is 4% increase in chances in males of getting heart disease due to smoking
- In females there seems to be no effect of smoking. But we think it may be due to insufficient data

❖ Percentage of males and females that were on blood pressure medication. Then finding how blood pressure medication effects your chances of having heart disease

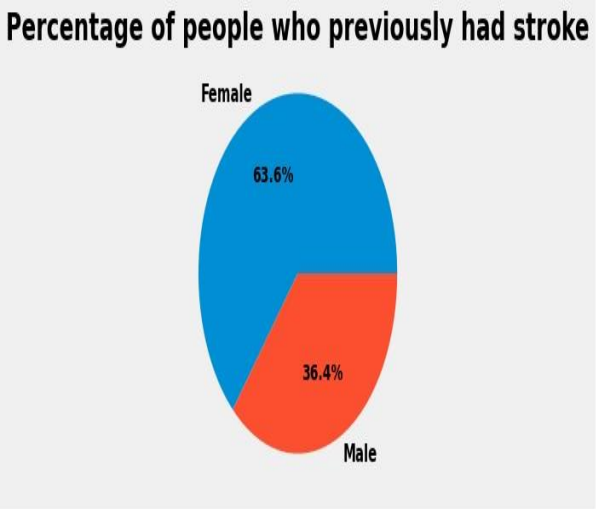


Percentage of males and females who were on blood pressure medication who eventually developed heart disease

	ten_year_chd	0	1
sex	bp_meds		
Female	No	88.36	11.64
	Yes	68.12	31.88
Male	No	81.82	18.18
	Yes	64.52	35.48

- Out of all those people those who were on blood pressure medication 69% were females and 31% were males
- In both males and females, the effect of blood pressure medication can be seen. In males out of all people who were on blood pressure medication 35.48% eventually contracted heart disease, whereas in females 31.88% contracted heart disease
- In males those who were not on blood pressure medication only 18.18% people contracted heart disease and in females 11.64% contracted heart disease. Being on blood pressure medication increases your chances of getting heart disease

❖ Percentage of males and females who had prevalent stroke. Then finding out how having prevalent stroke effects your chances of having heart disease

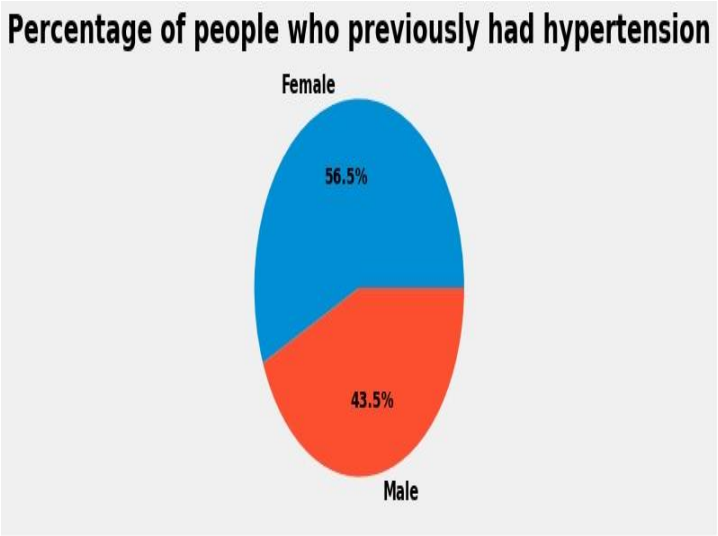


Percentage of males and females who previously had stroke, who eventually developed heart disease

	ten_year_chd	0	1
sex	prevalent_stroke		
Female	No	87.79	12.21
	Yes	57.14	42.86
Male	No	81.63	18.37
	Yes	50.00	50.00

- Out of those people those who had prevalent stroke 63.6% were females and 36.4% were males
- 50% of males who had stroke eventually developed heart disease and 42.86% of females who had stroke developed heart disease
- Having prevalent stroke, significantly increases your chances of having heart disease

❖ Plotting to find percentage of males and females who had prevalent hypertension. Then finding out how having prevalent hypertension effects your chances of having heart disease

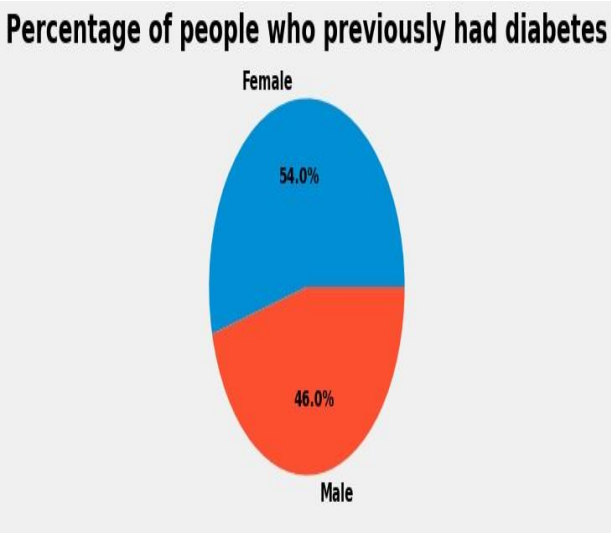


Percentage of males and females who previously had hypertension, who eventually developed heart disease

	ten_year_chd	0	1
sex	prevalent_hyp		
Female	No	91.74	8.26
	Yes	78.48	21.52
Male	No	85.33	14.67
	Yes	73.12	26.88

- Out of all those people who had hypertension, 56.5% were females and 43.5% were males
- 26.88% of males who previously had hypertension developed heart disease and 21.52% of females who had hypertension developed heart disease
- So having hypertension increases your chances of having heart disease

❖ Percentage of males and females who had diabetes. Then finding out how having diabetes effects your chances of having heart disease

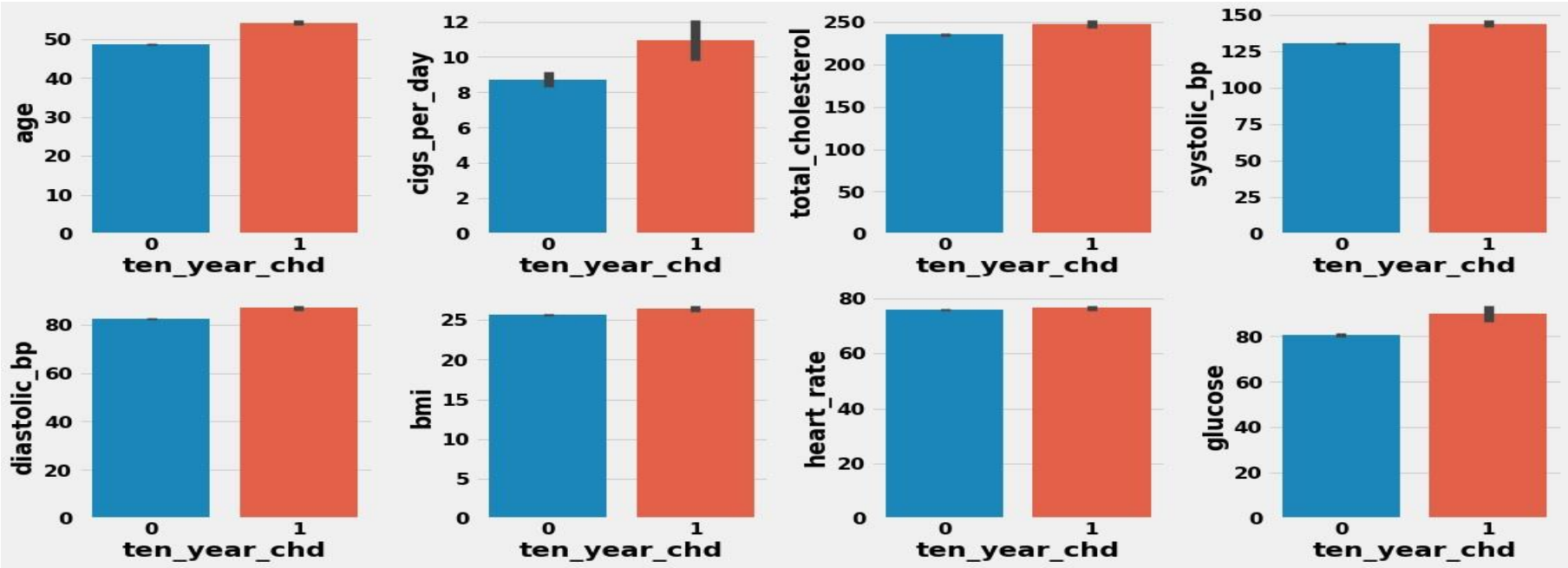


Percentage of males and females who previously had diabetes, who eventually developed heart disease

sex	diabetes	ten_year_chd 0	1
Female	No	88.06	11.94
	Yes	68.09	31.91
Male	No	82.20	17.80
	Yes	55.00	45.00

- Out of all those people who had diabetes, 54% were females and 46% were males
- 45% of males who had diabetes developed heart disease and 31.91% of females who had diabetes developed heart disease
- So having diabetes plays a role in person developing heart disease

❖ Plotting barplot to find the average value of numerical features when people contract heart disease and when people do not contract heart disease



- Age , cigs_per_day , total_cholesterol , systolic_bp , diastolic_bp , glucose generally have high values for people developing heart disease as compared to people who not develop heart disease
- Heart rate and bmi seems to be same for people developing heart disease and for people not developing heart disease

Preparation Of Data For Model Building

- **1) Data Splitting:** Before performing any feature engineering step, we need to split the data into a training and testing dataset to prevent data leakage.

- Since the data is imbalanced, a stratified split was employed to get an almost equal proportion of dependent variables in the training and test sets.

```
[52] # Creating a copy of dataset
     df3 = df.copy()

     # Creating df of only independent features
     X = df3.drop(columns='ten_year_chd',axis=1)

     # Creating a df of only dependent feature
     y = df3['ten_year_chd']

     # Train test split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y, random_state=42)
```

```
[53] # Printing shape of X_train and X_test
     print(f'The shape of train dataset is {X_train.shape}')
     print(f'The shape of test dataset is {X_test.shape}')

     The shape of train dataset is (2373, 14)
     The shape of test dataset is (1017, 14)
```

2) Handling Missing values : As in our dataset, some features contain missing values, so before model building we need to take care of them.

Missing Values/Null Values

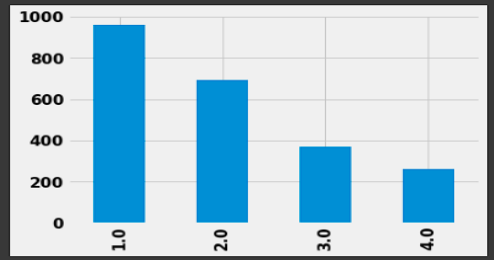
```
# Missing Values/Null Values Count
df.isnull().sum()
```

```
id                   0
age                  0
education           87
sex                  0
is_smoking           0
cigs_per_day        22
bp_meds             44
prevalent_stroke     0
prevalent_hyp        0
diabetes             0
total_cholesterol    38
systolic_bp          0
diastolic_bp         0
bmi                  14
heart_rate           1
glucose            304
ten_year_chd         0
dtype: int64
```

Each feature will have a different impact on our model, so we have used different imputation technique to fill in missing values in each feature.

1) **In the education** column, we have filled missing values with the most frequent category.
2) We have filled missing values in **cigs_per_day** by taking the median of the cigs per day column who actually smoke.

```
# Count of frequencies in education
X_train['education'].value_counts().plot(kind='bar');
```




```
[57] # Filling missing values in education with most frequent values
     X_train['education'] = X_train['education'].fillna(1)
     X_test['education'] = X_test['education'].fillna(1)
```

```
# Finding median of X_train data in cigs_per_day
median_cigs = X_train[X_train['is_smoking']=='YES']['cigs_per_day'].median()

# Filling missing values in X_train
X_train['cigs_per_day'] = X_train['cigs_per_day'].fillna(median_cigs)

# Filling missing values in X_test
X_test['cigs_per_day'] = X_test['cigs_per_day'].fillna(median_cigs)
```

Continued

- 3) For the **bp_meds** column, if we impute values with the most frequent values, i.e., "No," we are making a huge assumption that the person was not on blood pressure medication. Hence, it's better to drop missing rows.
- 4) For remaining columns such as **total_cholesterol**, **bmi**, and **heart_rate** we have imputed missing values with the respective column median value.
- 5) As the glucose column has 8% missing values. If we use the median to impute, there is a great chance of disturbing the distribution of glucose. Also, glucose is an important feature, so we use the KNN imputer for accurate results.

```
def filling_missing_values(dataset1,dataset2,feature_list):
    '''This function fills missing values in 'total_cholesterol','bmi','heart_rate' with median'''

    for feature in feature_list:
      # Finding median of X_train data of provided feature
      median_value_of_feature = dataset1[feature].median()

      # Filling missing values in X_train
      dataset1[feature] = dataset1[feature].fillna(median_value_of_feature)

      # Filling missing values in X_test
      dataset2[feature] = dataset2[feature].fillna(median_value_of_feature)
```

```
[62] # Filling missing values in 'total_cholesterol','bmi','heart_rate'
     filling_missing_values(X_train,X_test,['total_cholesterol','bmi','heart_rate','glucose'])
```

```
[65] # Creating an instance of KNNImputer class
     knn_impute = KNNImputer(n_neighbors=10)

     # Imputing values using KNN imputer
     X_train = pd.DataFrame(knn_impute.fit_transform(X_train),columns=X_train.columns)
     X_test = pd.DataFrame(knn_impute.transform(X_test),columns=X_test.columns)
```

3) Categorical Encoding: As all our categorical features have binary labels so we have use binary label encoding technique.

- For '**sex**' column,(male=1,Female = 0)
- For '**is_smoking**' column (yes =1 ,No=0)
- For '**bp_meds**' column (yes =1 ,No=0)
- For '**prevalent_stroke**' column (yes =1 ,No=0)
- For '**prevalent_hyp**' column (yes =1 ,No=0)
- For '**diabetes**' column (yes =1 ,No=0)

```
def encoding(dataset):
  ''' This function binary encodes 'sex','is_smoking','bp_meds','prevalent_stroke','prevalent_hyp','diabetes'
      columns in dataset '''

  # Encoding 'sex' feature
  dataset['sex'] = dataset['sex'].map({'Male':1,'Female':0})

  # Encoding required list of columns
  for col in ['is_smoking','bp_meds','prevalent_stroke','prevalent_hyp','diabetes']:
    dataset[col] = dataset[col].map({'Yes':1,'No':0})
```

4) Handling Skew and Outliers:

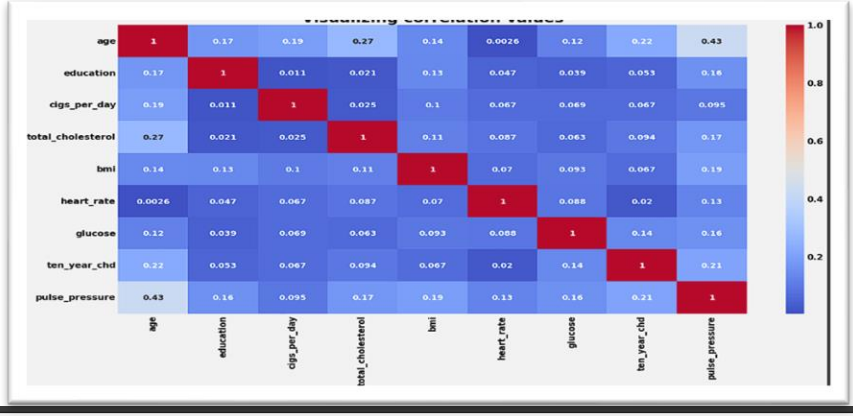
- The skew in numeric variables is reduced by performing **log transformation.**

- The outliers beyond 3 **standard deviations** from the mean were **imputed** with the **median** value

Attribute	Original skew	Skew After Transformation
age	0.208197	-0.036952
cigs_per_day	1.249114	0.274297
total_cholesterol	1.122804	-0.041904
pulse_pressure	1.407770	0.295810
bmi	0.982133	0.178421
heart_rate	0.675533	0.086530
glucose	6.529257	0.431254

5) Feature Manipulation and Selection:

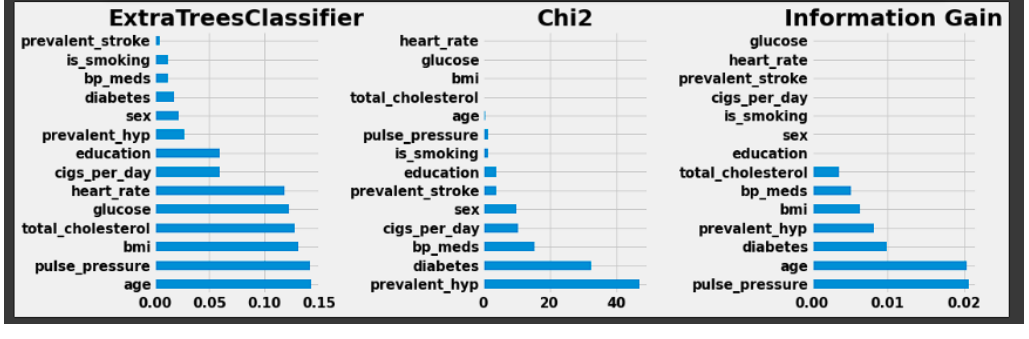
- ***removing multicollinearity*:** As in our dataset Systolic blood pressure and diastolic blood pressure are highly correlated. Hence to solve multicollinearity we have created new feature.
- **Pulse pressure = Systolic BP – Diastolic BP**



Different feature selection methods we used such as :

- Gini Impurity
- Chi2 Test
- Information Gain

From **chi2** test we can clearly see that '**is_smoking**' column has the highest p-value so it is the least relevant feature. As a result, we have **dropped** it.

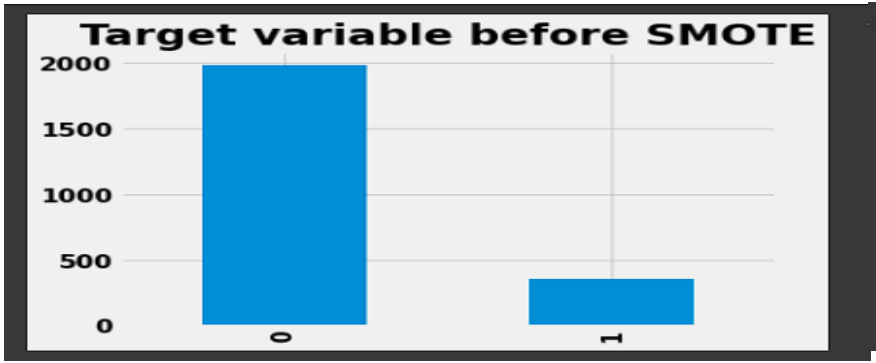


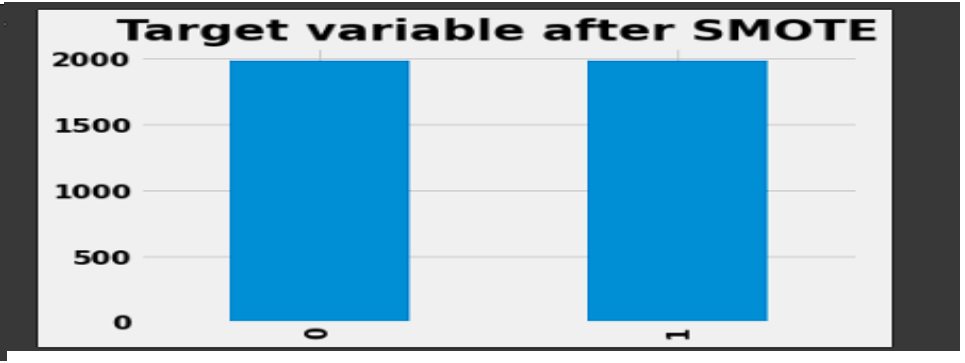
6) Data Scaling:

- Since predictions from the distance-based models will be affected if the attributes are in different ranges, we use StandardScaler to scale down the variables.

7) Handling Imbalanced Dataset:

Since we are dealing with unbalanced data, i.e., only 15% of the patients were diagnosed with coronary heart disease, we have oversampled the training dataset using SMOTE (Synthetic Minority Oversampling Technique).
This will ensure that our model has been trained equally on all kinds of results and is not biased towards one particular result.





Model Selection And Evaluation

Before starting model building, it's important to choose the evaluation metric. We have chosen **recall** as the **evaluation metric**. Because this is a health-care dataset, it is critical to predict who is most likely to develop heart disease. We are fine with a false positive (the model predicted heart disease and the person did not get heart disease), but a false negative (the model did not predict heart disease and the person got heart disease) is very dangerous as the person may lose his life.

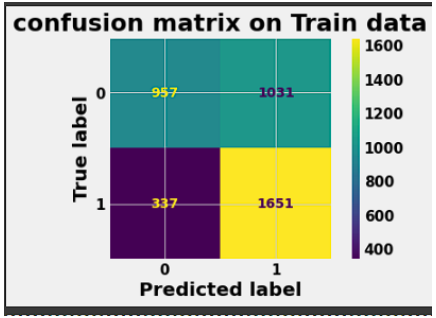
$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

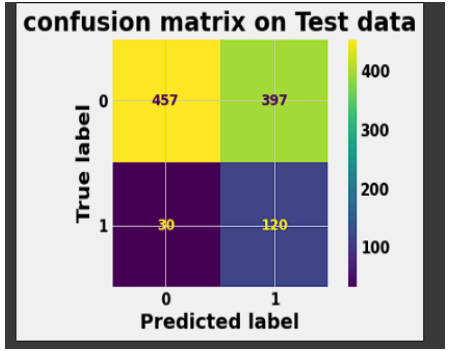
Continued

By fitting the data into various classification models and evaluating them with test and training data, we got the following results for the recall of different models on the training and test data, from which we can conclude that **"logistic regression"** and **“Decision tree"** were the two best performing models. We have plotted a confusion matrix for logistic regression and decision trees for training and testing data.

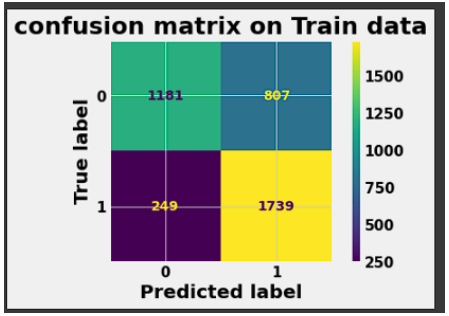
	Train_recall	Test_recall
Support Vector Machines	1.00	0.23
Naive Bayes	0.38	0.34
XG Boosting	0.78	0.46
K Nearest Neighbors	1.00	0.65
Random Forest	0.94	0.66
Decision Tree	0.89	0.73
Logistic Regression	0.84	0.81

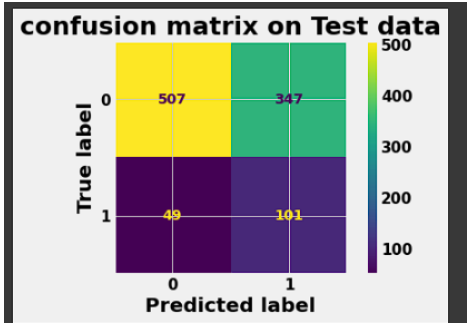
Logistic Regression





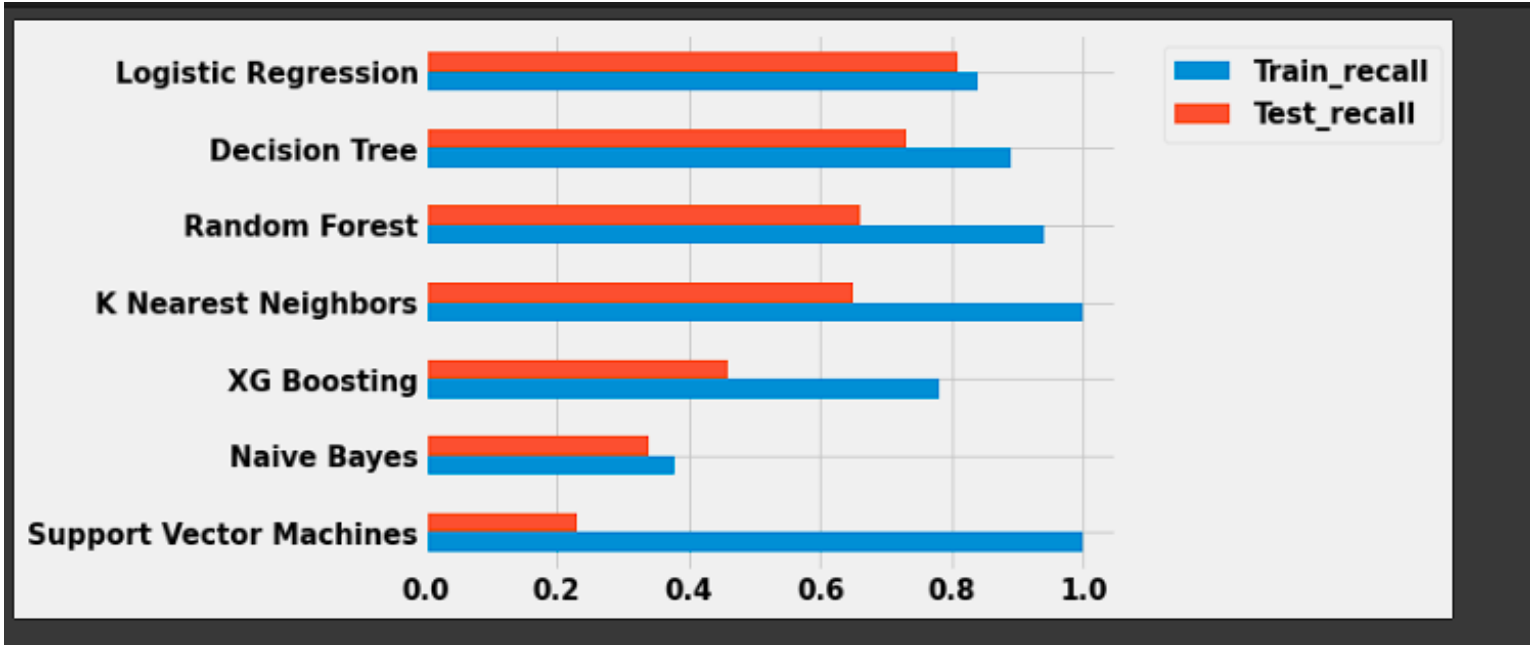
Decision Tree



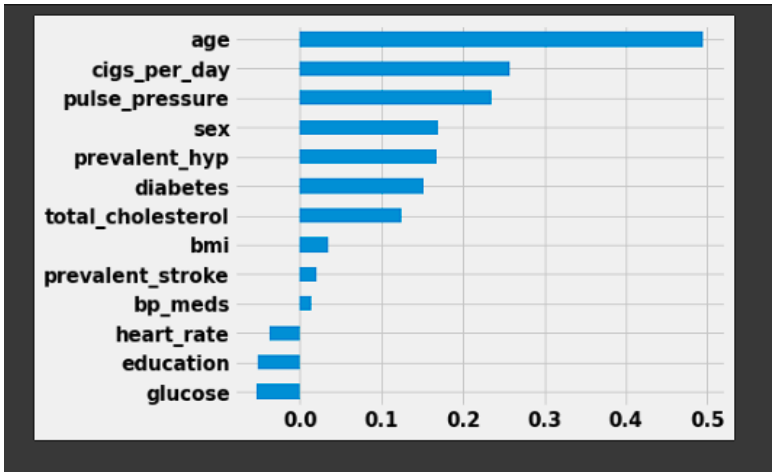


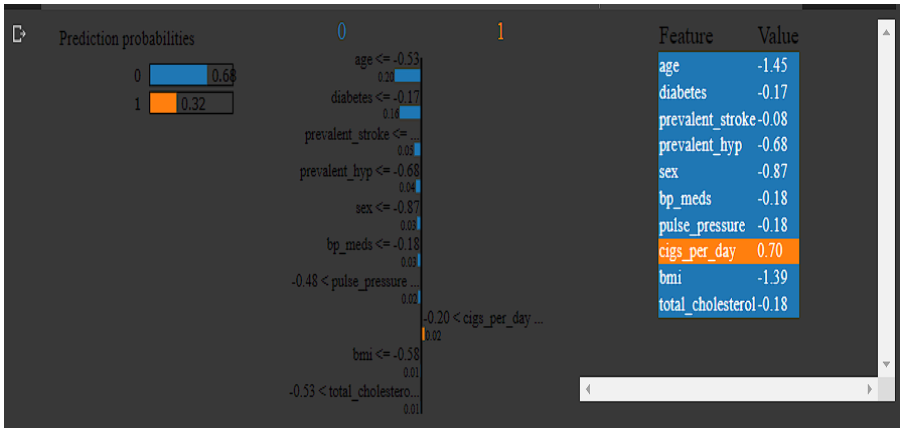
Continued

- But here we can observe that Decision Tree has overfitted as it has relatively large difference between train recall and test recall
- Logistic Regression seems to have generalized well on the given data.
- Hence we have selected **'Logistic Regression**‘ as model for deployment



Feature Importance and Model Explainability





- **'Age'** is the most important feature which influences the probability of getting a he art disease.Olderly people are most at risk of getting a heart attack
- **'Cigs per day'**, **'Prevalent hypertension'** and **'Pulse Pressure'** are next important features.

Smoking and having high pulse pressure increases the chances of getting a heart attack

Conclusion

1) We trained 7 machine learning models using the training dataset, and hyperparameter tuning was used in some models to improve the model's performance.
2) Missing values were handled, feature engineering and feature selection were carried out, and the training dataset was oversampled using SMOTE to reduce bias on one outcome.
3) We chose recall as the model evaluation metric because it was critical that we reduce false negatives.
4) Predicting the risk of coronary heart disease is critical for reducing fatalities caused by this illness. We can avert deaths by taking the required medications and precautions if we can foresee the danger of this sickness ahead of time.
5) It is critical that the model we develop has a high recall score. It is OK if the model incorrectly identifies a healthy patient as a high risk patient because it will not result in death, but if a high risk patient is incorrectly labelled as healthy, it may result in fatality. We were able to create a model with a recall of just 0.81 because of limitated data available and limited computational power available.
6) A recall score of 0.81 indicates that out of 100 individuals with the illness, our model will be able to classify only 81 as high risk patients, while the remaining 19 will be misclassified.
7) Future developments must include a strategy to improve the model recall score, enabling us to save even more lives from this disease. This includes involving more people in the study, and include people with different medical history, etc. build an application with better recall score.
8) From our analysis, it is also found that the age of a person was the most important feature in determining the risk of a patient getting infected with CHD, followed by pulse pressure, prevalent hypertension and total cholesterol. Diabetes, prevalent stroke and BP medication were the least important features in determining the risk of CHD.

Thank You