Heart Disease Prediction

Vikas Ranjan

DSC680, Summer 2021

Bellevue University, NE

Introduction

For the final project of this term, I wanted to do work on healthcare domain. Hence, ended up choosing a dataset with attributes relating to heart disease. According to WHO, an estimated 17.9 million people died from heart diseases in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. On average, someone dies of CVD every 36 seconds in the US. There are 2,380 deaths from CVD each day, based on 2018 data. Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death. The early prognosis of heart diseases can help in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, fatalities, and overall healthcare cost on economy.

The purpose of this project is to identify the factors that would allow us to predict if a person might get heart disease. Using the classification methods, the intent is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). EDA and statistical analysis will also be performed on the dataset.

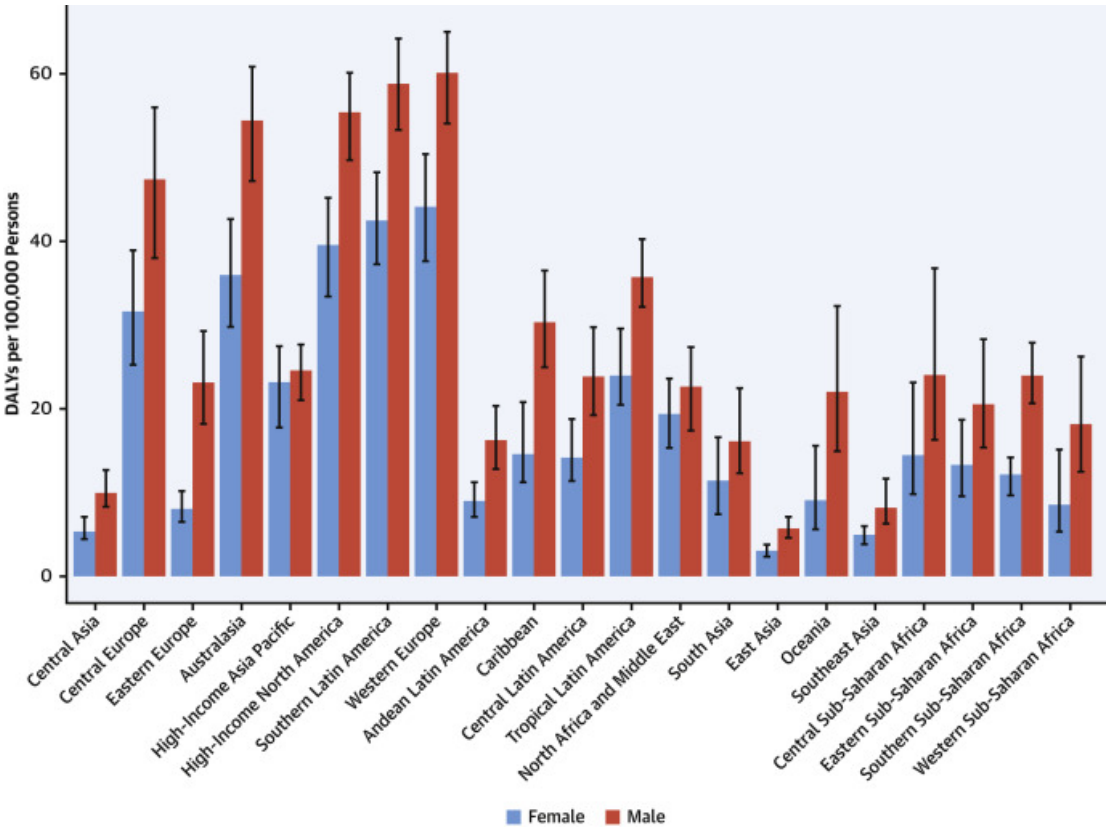
Methods

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains following steps which will be followed throughout the project.

- **Business Understanding** – Heart disease is one of the most significant health crisis that the world is facing today. It is not only the leading cause of the death in the United States but also a major cause of disability. There are many things that can raise your risk

for heart disease, and they are referred as risk factors. A large percentage of heart diseases can be prevented. Cardiovascular disease prediction is a critical challenge in the clinical data analysis. Machine learning (ML) has been showing an effective assistance in making decisions and predictions from the large quantity of data produced by the healthcare industries and hospitals.

Below are some of the charts showing the impact and severity of heart disease.



As part of this project, I'll be looking at a few patterns and insights in the heart disease dataset. I'll also be developing a few models for predicting the 10 year coronary heart disease based on the associated risk factors in the dataset. The dataset I've found and used is one of the clean & seemingly reliable dataset.

- **Data Understanding** - I extracted the dataset from Kaggle from the following url: https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression. This dataset includes 4238 rows and 15 feature variables. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors in the dataset. Following are the attributes:

  1. Sex: Male or female
  2. Age: Age of the patient
  3. Current Smoker: If the patient is a current smoker

4. Cigs Per Day: The number of cigarettes that the person smoked on average in one day
5. BP Meds: Whether or not the patient was on blood pressure medication
6. Prevalent Stroke: If the patient had previously had a stroke
7. Prevalent Hyp: Whether or not the patient was hypertensive
8. Diabetes: Whether or not the patient had diabetes
9. Tot Chol: Total cholesterol levels
10. Sys BP: Systolic blood pressure
11. Dia BP: Diastolic blood pressure
12. BMI: Body Mass Index
13. Heart Rate: Heart rate
14. Glucose: Glucose level
15. 10 year risk of coronary heart disease CHD -1 means “Yes” & 0 means “No”

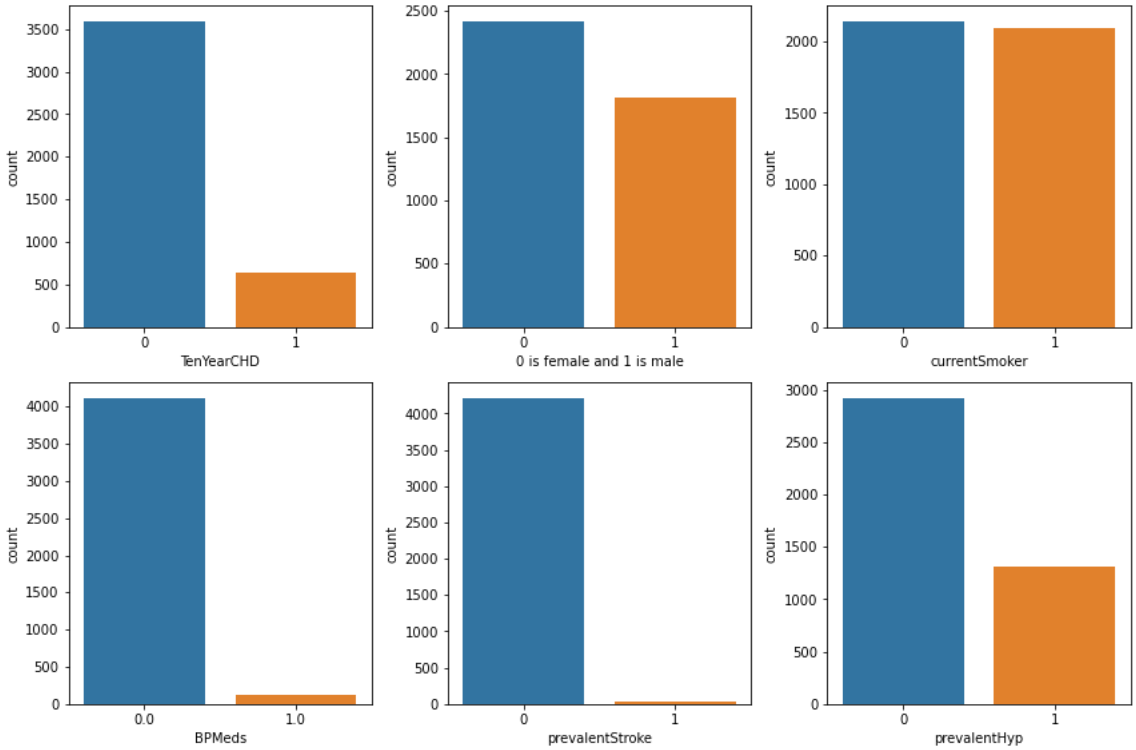
All personally identifying information has been removed from the data.

- **Data Preparation** – The source dataset was relatively clean. First step was to perform data cleanup.
    - Education attribute is not an associated risk factor, therefore removed it.
    - There were a few duplicates which were removed as well.
    - For the null values in cigsPerDay, BPMeds, totChol, BMI, heartRate, & glucose, updated them to the median values of their respective attributes.

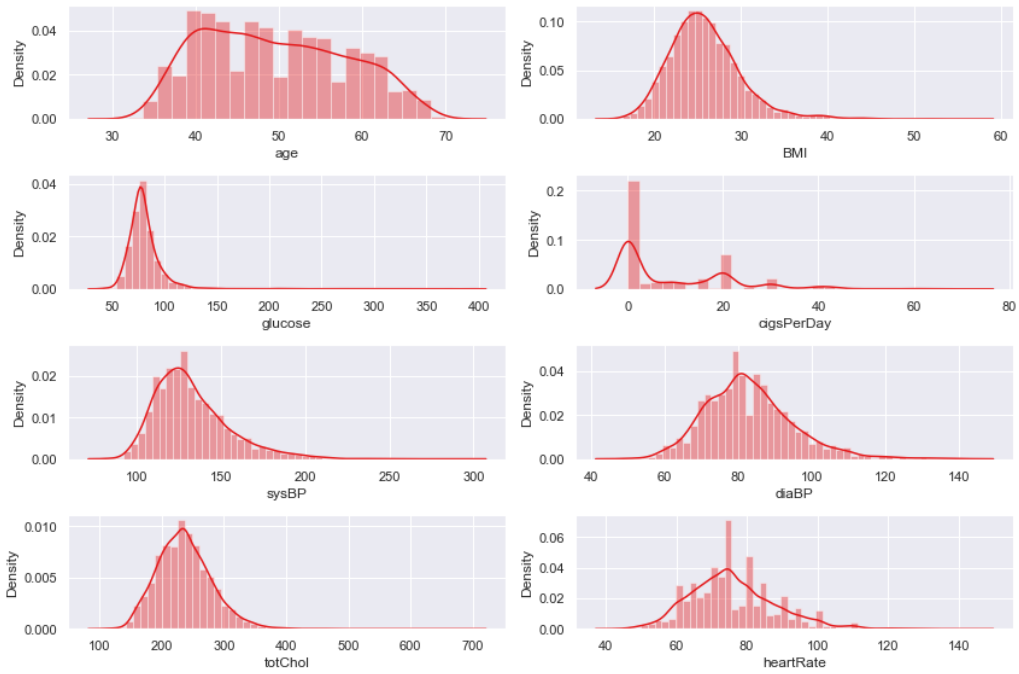
    As part of data preprocessing, I performed the following actions
    - Identified and removed the outliers from the dataset.
    - Performed scaling of data using StandardScaler which means that our values are centered around mean with a unit standard deviation.

- The dataset was moderately imbalanced, therefore applied Random oversampling to make the training dataset balanced.

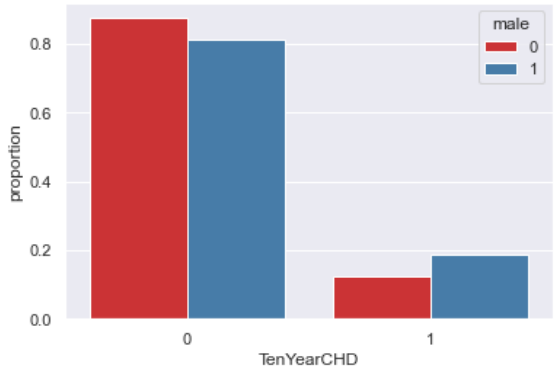
- **Exploratory Data Analysis** - Next step of understanding data is to perform graph analysis. Graph analysis helps significantly with that as it not only shows the patterns and distributions but also would also tell which attributes are corelated. They would also help to guide future business decisions.
    - Below plots shows the distribution of all the Boolean attributes of the dataset and here are some of the observations:
        - Men stand at a higher risk of getting a coronary heart disease compared to women.
        - Dataset has more number of females than males.
        - Dataset has very few number of people who are on BPMeds and had a prevalent stoke.



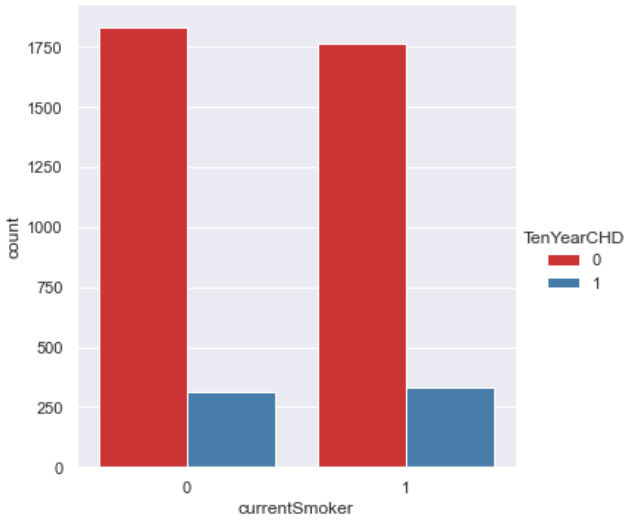
- Below set of plots show the distribution of all the continuous variables in the dataset. Below are some of the key observations:
    - Dataset contains details of people between age 35 and 70.



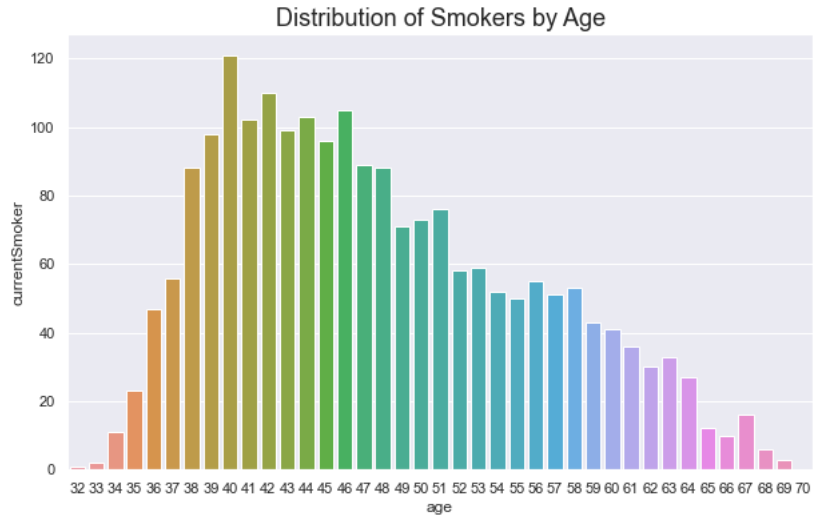
- As we can see in below plot, most of the people who don't have a 10-year risk of coronary heart disease are female while the ones who have the risk are generally male.



- Below plot demonstrates that smoking status has insignificant effect on the risk in the data.



- Below chart shows the distribution of smokers by age.



- Below is the box and whisker plot of distribution of Glucose, Total Cholesterol, Systolic blood pressure, Siastolic blood pressure, Heart rate against TenYearCHD. And here are some observations:

Glucose vs TenYearCHD - The distributions for both the risk group and the group of people who don't have the risk are almost the same except the third quartile which

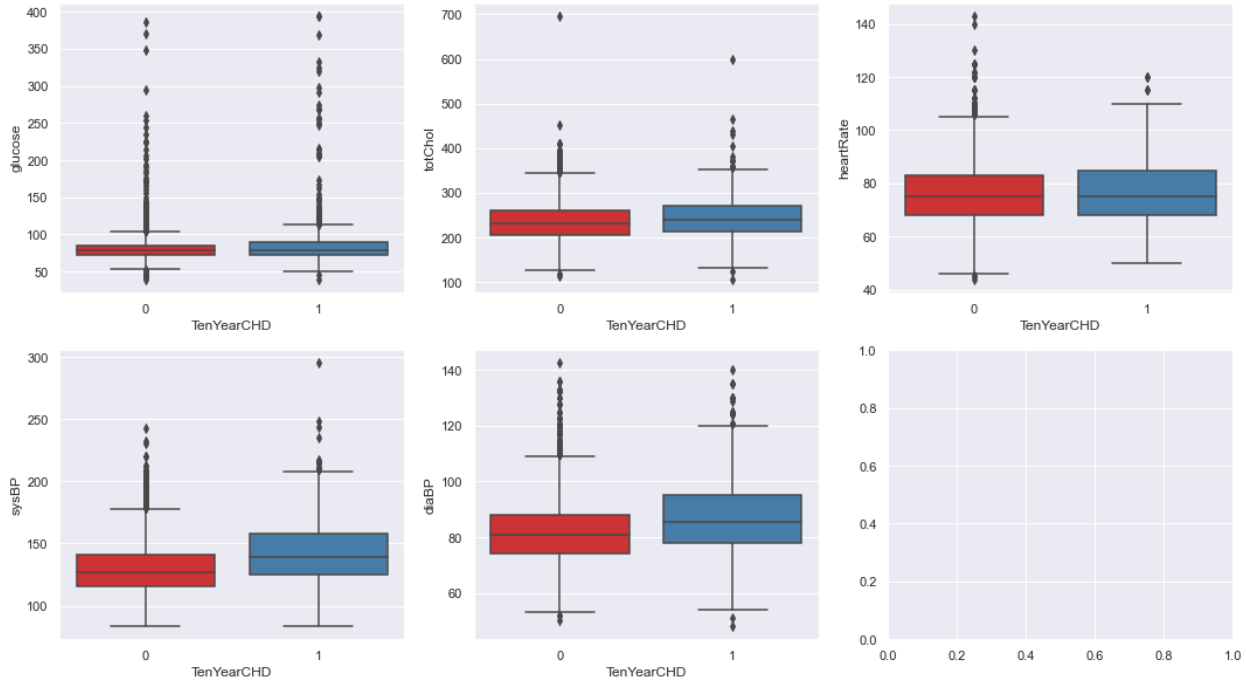
is greater and the maximum value which is slightly greater for the risk group. Both groups have so many outliers.

Total Cholesterol vs TenYearCHD - The distributions for both the risk group and the group of people who don't have the risk are almost the same. Both groups have so many outliers.

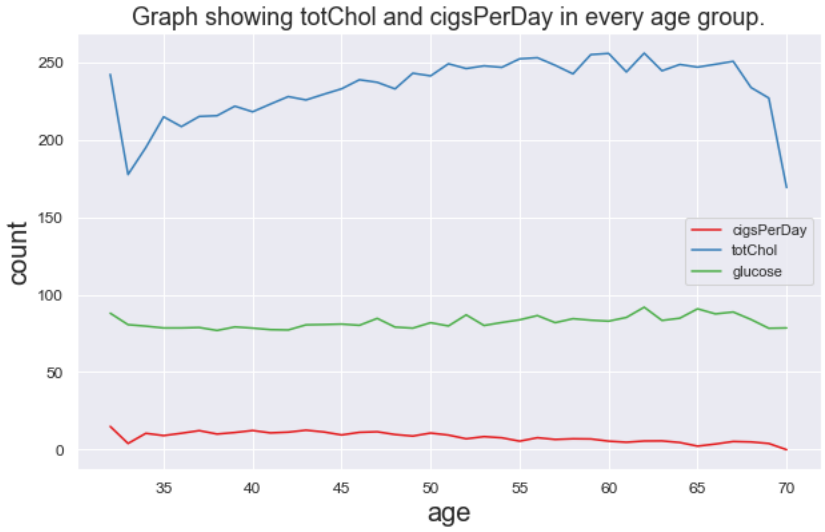
Heart Rate vs TenYearCHD - The distributions for both the risk group and the group of - people who don't have the risk are almost the same except for the third quartile, minimum and maximum values which are slightly greater for the risk group. Both groups have so many outliers especially the group of people who don't have the risk.

Systolic Blood Pressure vs TenYearCHD - In the group that doesn't have a 10-year risk of coronary heart disease, the median is about 130 while in the other group that has a 10-year risk of coronary heart disease it is almost 150. The minimum systolic blood pressure value for both two groups are the same while the maximum value is much higher in the risk group. Also, the first and third quartiles are so much higher in the risk group.

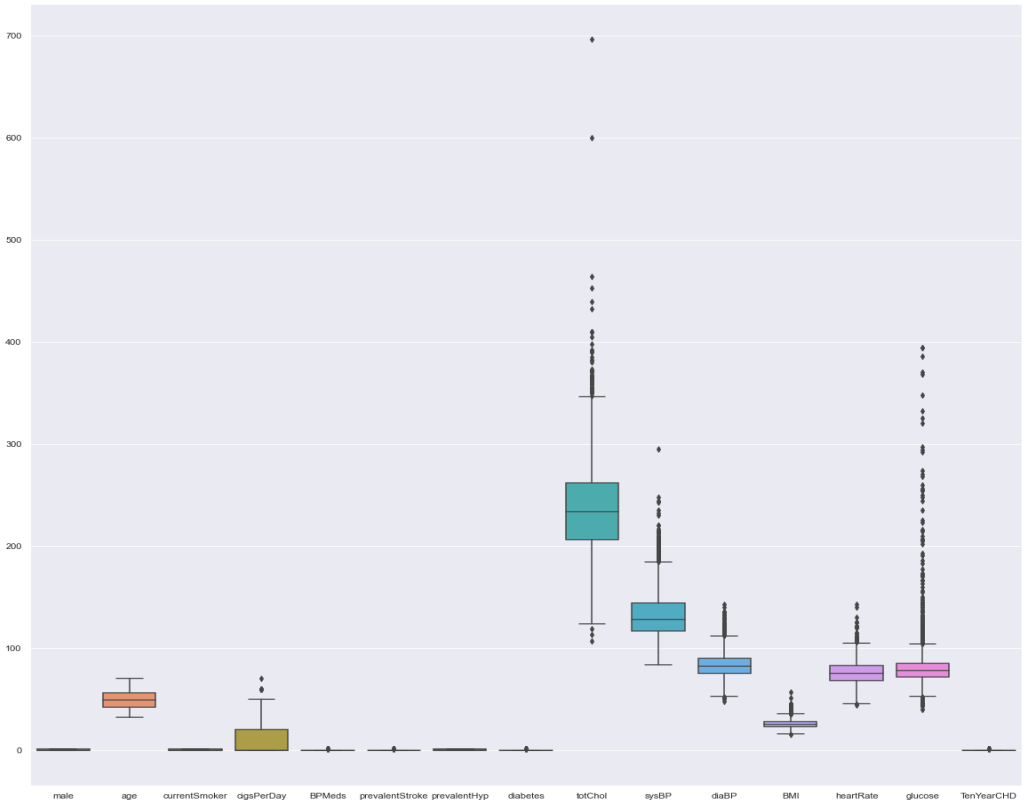
Diastolic Blood Pressure vs TenYearCHD - Like systolic blood pressure, diastolic blood pressure's median, max, first quartile, and third quartile values are higher for the risk group.



- Below plot shows the line graph comparison of totChol, cigPerDay & glucose.

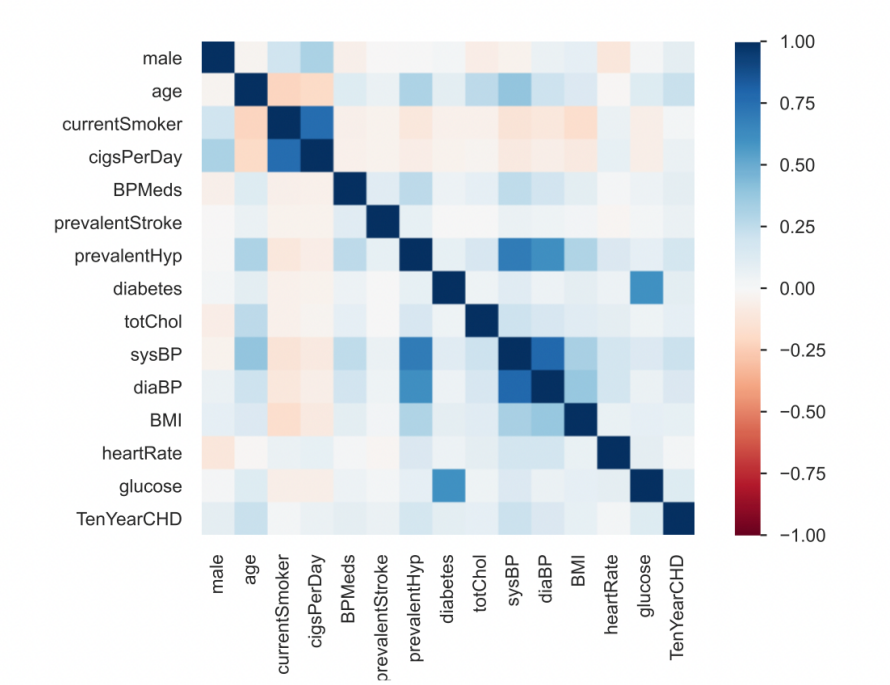


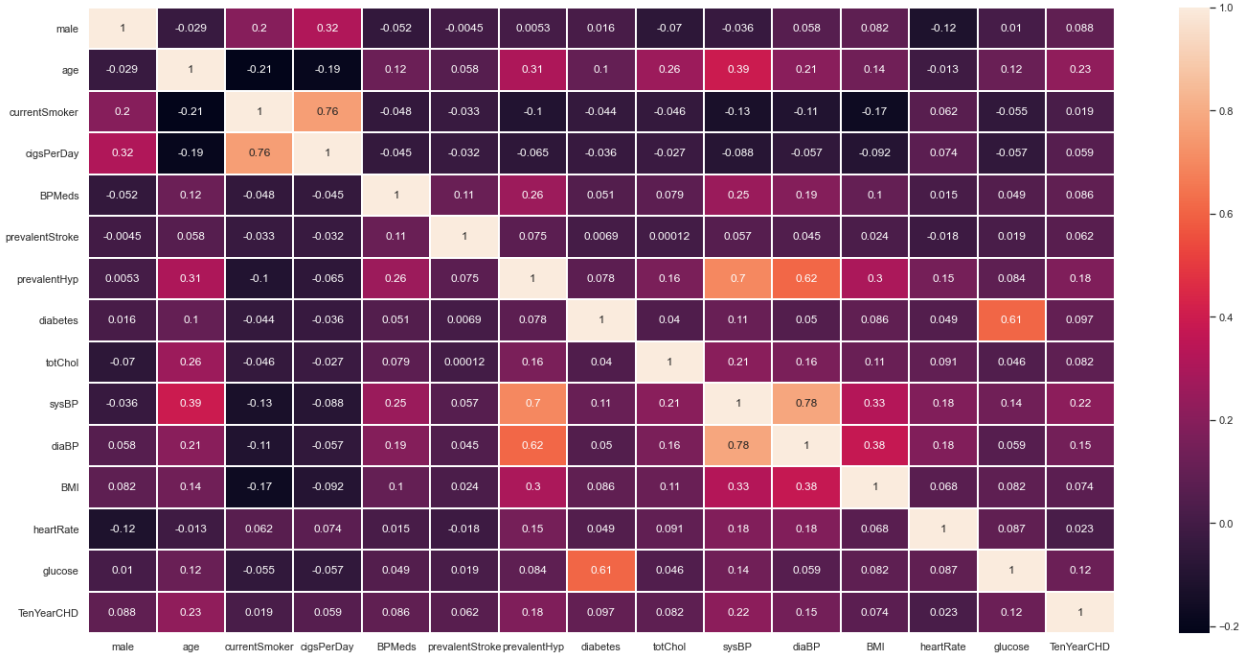
- Below box-whisker plot helps to identify the outliers in the dataset.



- Heatmap of the correlation

  Pearson Correlation:






```
TenYearCHD          1.000000
age                 0.225256
sysBP               0.216429
prevalentHyp        0.177603
diaBP               0.145299
glucose             0.121277
diabetes            0.097317
male                0.088428
BPMeds              0.086417
totChol             0.081566
BMI                 0.074217
prevalentStroke     0.061810
cigsPerDay          0.058859
heartRate           0.022857
currentSmoker       0.019456
Name: TenYearCHD, dtype: float64
```

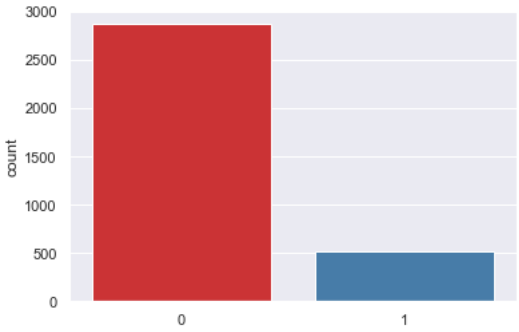
Correlations Analyis - TenYearCHD:

- The strongest positive correlations of TenYearCHD are with age and sysBP.
- CurrentSmoker has the least correlation with TenYearCHD.

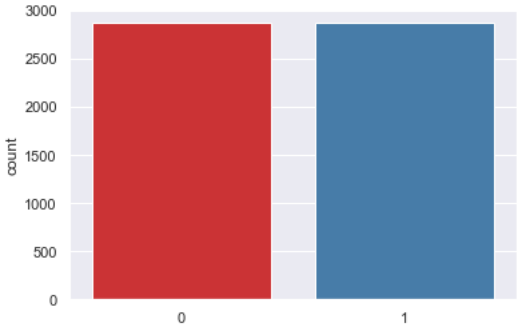
Modeling – As part of modeling, I did notice that this dataset is moderately imbalanced. If not taken care of, results may be biased as the algorithms are much likely to classify new observations to the majority class and high accuracy won't tell us anything. To

address the problem of imbalanced dataset, I choose to use oversampling data approach technique. Oversampling increases the number of minority class members in the training set. In order to make our data set balanced, I'm using a type of oversampling called Random Oversampling to overcome using resample. RandomOverSample class is part of imblearn which allows us to over-sample the minority class by picking samples at random with replacement. One of the other popular class would be SMOTE.

Training dataset before oversampling:



Training dataset after oversampling:



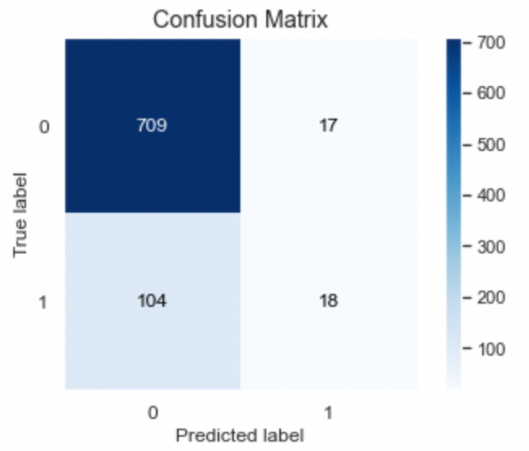
As we can see above, after oversampling training dataset is now balanced and ready for the classification models to be applied on it.

Classification models:

1. **Random forest model** creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. It provides us a high true prediction values for our dataset. This model when applied on the dataset has an AUROC of 0.8573 which means that the model has very good discriminatory ability.

```
Confusion Matrix:
[[709  17]
 [104  18]]

Classification report:
              precision    recall  f1-score   support

           0       0.87      0.98      0.92       726
           1       0.51      0.15      0.23       122

    accuracy                           0.86       848
   macro avg       0.69      0.56      0.58       848
weighted avg       0.82      0.86      0.82       848

Accuracy of the model:  0.8573113207547169
```

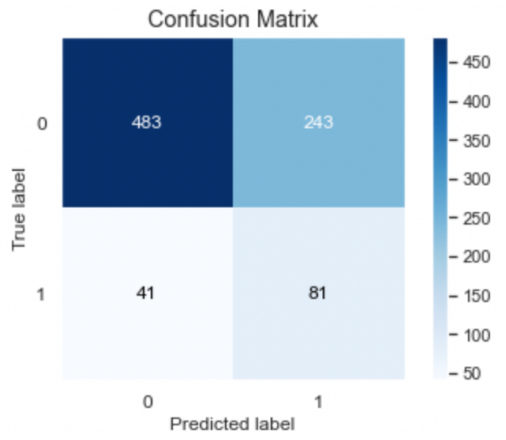
Confusion Matrix:



2. **Logistic regression model** takes a linear equation as input and use logistic function and log odds to perform a binary classification task. I tested this model with the test dataset which resulted in an AUROC of 0.6650. Below is the snapshot of the results.

```
Confusion Matrix:
[[483 243]
 [ 41  81]]

 Classification report:
              precision    recall  f1-score   support

           0       0.92      0.67      0.77       726
           1       0.25      0.66      0.36       122

    accuracy                           0.67       848
   macro avg       0.59      0.66      0.57       848
weighted avg       0.83      0.67      0.71       848

Accuracy of the model:  0.6650943396226415
```

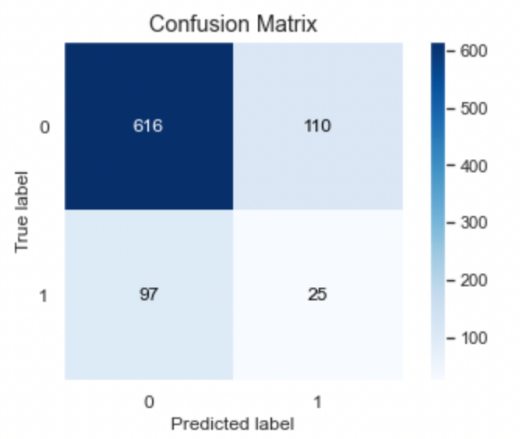
Confusion Matrix:



3. **Decision Tree model** - I tested this model with the test dataset which resulted in an AUROC of 0.7559. Below is the snapshot of results.

```
Confusion Matrix:
[[616 110]
 [ 97  25]]

 Classification report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.86       726
           1       0.19      0.20      0.19       122

    accuracy                           0.76       848
   macro avg       0.52      0.53      0.53       848
weighted avg       0.77      0.76      0.76       848

Accuracy of the model:  0.7558962264150944
```

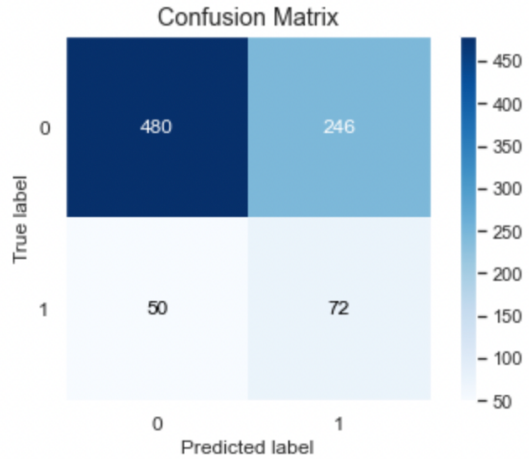
Confusion Matrix:



4. **KNN classification** - Applied KNN model on the test dataset which resulted in an AUROC of 0.6509. Below is the snapshot of results.

```
Confusion Matrix:
[[480 246]
 [ 50  72]]


 Classification report:
              precision    recall  f1-score   support

           0       0.91      0.66      0.76       726
           1       0.23      0.59      0.33       122

    accuracy                           0.65       848
   macro avg       0.57      0.63      0.55       848
weighted avg       0.81      0.65      0.70       848

Accuracy of the model:  0.6509433962264151
```

Confusion Matrix:



5. **SVM (Support Vector Machine) Classification -** Applied SVM model on the training dataset with different parameters, below are results for each of those:

   - SVM with Gamma = 0.1, and C = 1.0

	precision	recall	f1-score	support
0	0.90	0.73	0.81	726
1	0.25	0.54	0.34	122
accuracy			0.70	848
macro avg	0.58	0.64	0.58	848
weighted avg	0.81	0.70	0.74	848

Accuracy of the model: 0.7770223152022315

   - SVM with Gamma = 0.01, and C = 1.0

	precision	recall	f1-score	support
0	0.93	0.66	0.77	726
1	0.25	0.69	0.37	122
accuracy			0.66	848
macro avg	0.59	0.67	0.57	848
weighted avg	0.83	0.66	0.71	848

Accuracy of the model: 0.6795676429567643

   - SVM with Gamma = 0.1, and C = 80

	precision	recall	f1-score	support
0	0.88	0.82	0.85	726
1	0.23	0.32	0.26	122
accuracy			0.74	848
macro avg	0.55	0.57	0.55	848
weighted avg	0.78	0.74	0.76	848

Accuracy of the model: 0.9494421199442212

- **<u>Models Evaluation</u> –** Comparison of the implemented models in terms of accuracy. Here, we can see that SVM classification model has the best accuracy score (94.944), followed by Random Forest classification model (85.73) had best accuracy scores when applied on this dataset.

	Model	Accuracy
0	Logistic Regression	66.509434
1	K-Nearest Neighbour	65.094340
2	SVM	94.944212
3	Decision Tree	75.589623
4	Random Forrest	85.731132

Conclusion

- The strongest positive correlations of TenYearCHD are with age and sysBP.
- Performed Standardization technique to scale the data before running it through the ML algorithms.
- Since the training dataset was moderately imbalanced, applied Random oversampling technique to overcome the imbalanced datasets challenge.
- Based on the accuracy of the implemented classification models to predict TenYearCHD, SVM classification model shows highest accuracy followed by Random Forest classification model.

References

1. Dileep. (2019, June 7). Logistic regression to predict heart disease. Kaggle. https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression.
2. Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. (2020, September 9). Artificial intelligence, machine learning, and cardiovascular disease. Clinical Medicine Insights. Cardiology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7485162/.
3. Shah, D., Patel, S., & Bharti, S. K. (2020, October 16). Heart disease prediction using machine learning techniques. SN Computer Science. https://link.springer.com/article/10.1007/s42979-020-00365-y.

4. HealthITAnalytics. (2020, August 5). Artificial intelligence may accelerate heart failure diagnosis. HealthITAnalytics. https://healthitanalytics.com/news/artificial-intelligence-may-accelerate-heart-failure-diagnosis.

5. Centers for Disease Control and Prevention. (2020, September 8). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm.

6. World Health Organization. (n.d.). Cardiovascular diseases. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

7. Krittanawong, C., Zhang, H. J., Wang, Z., Aydar, M., & Kitai, T. (2017, May 22). Artificial intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology. https://www.sciencedirect.com/science/article/pii/S0735109717368456.

8. Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda

9. Machine Learning with Python Cookbook, Chris Albon

10. Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H. J., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. H. W. (2020, September 29). Machine learning prediction in cardiovascular diseases: A meta-analysis. Nature News. https://www.nature.com/articles/s41598-020-72685-1.