Capstone Project 3

Cardiovascular Risk Prediction

By
Harshavardhan M. Shete

Cardiovascular Risk Prediction

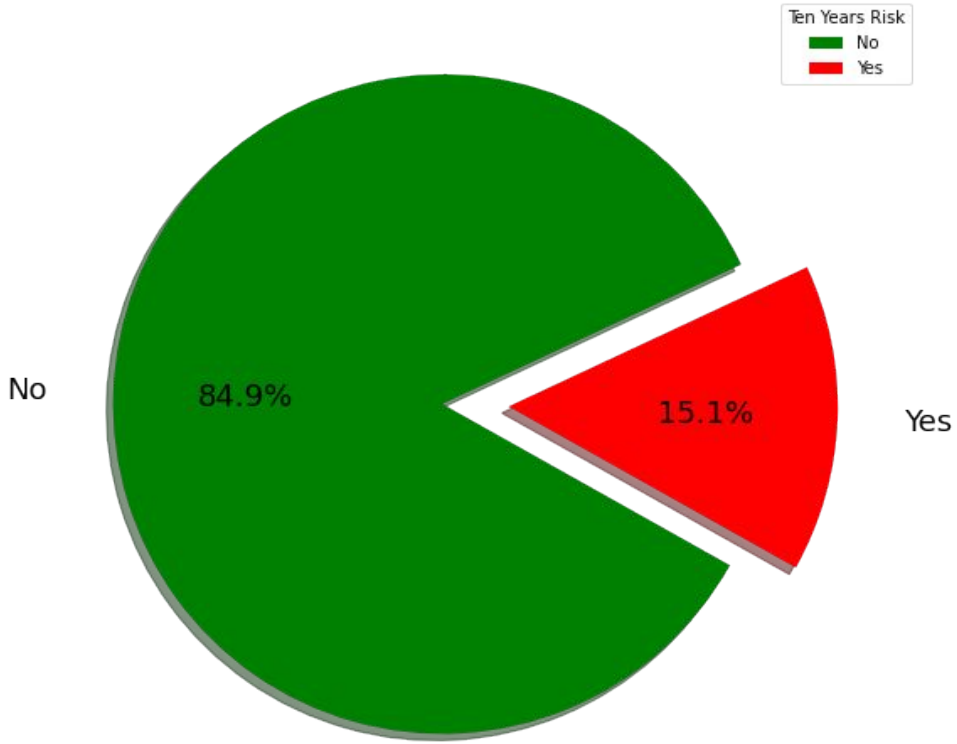
- Framingham Heart Study
- Dataset: 3390 Rows, 17 Columns
  - Demographic: Sex, Age
  - Behavioral: Smoking habits, Cigarettes per day
  - Medical (History): BP Medicines, Prevalent Stroke, Prevalent Hypertension, Diabetes
  - Medical (Current): Cholesterol Level, Systolic and diastolic blood pressure, BMI, Heart rate, Glucose Level
- Target Variable: 10-year risk of Coronary Heart Disease
- Binary Classification

Solutioning

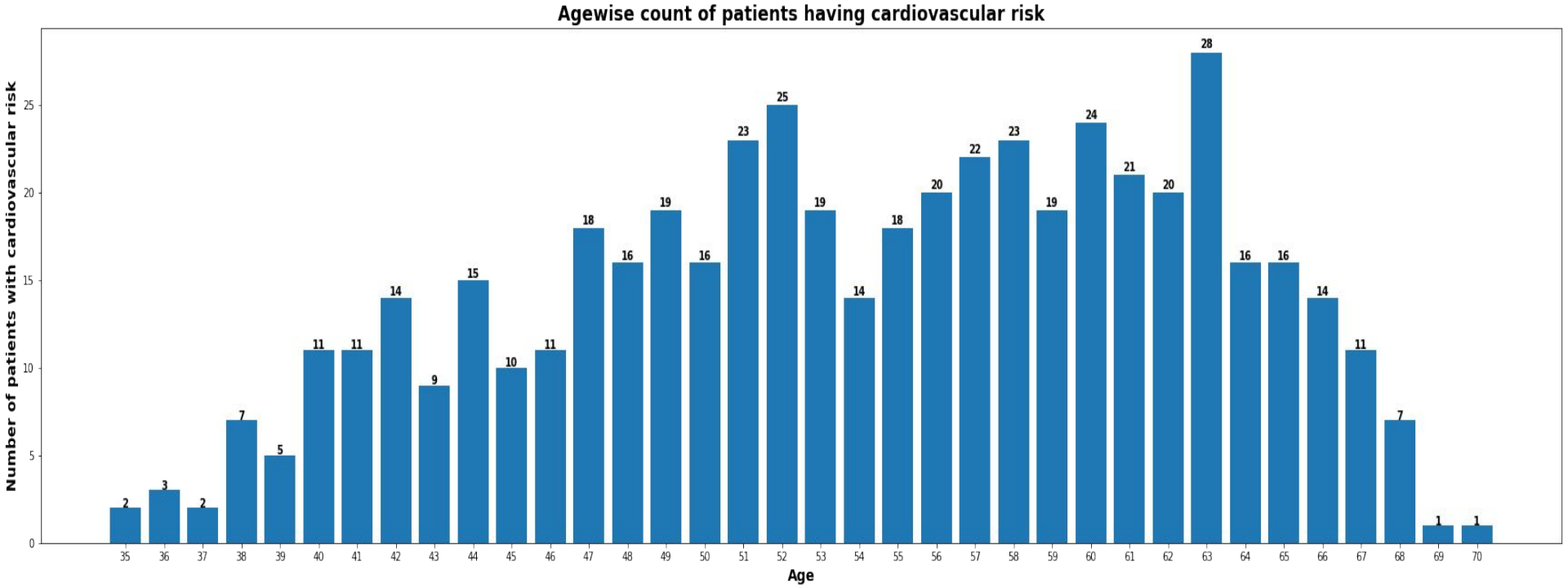
- Exploratory Data Analysis
- Handling Imbalanced Dataset
- Model Training and Performance Comparison

Exploratory Data Analysis

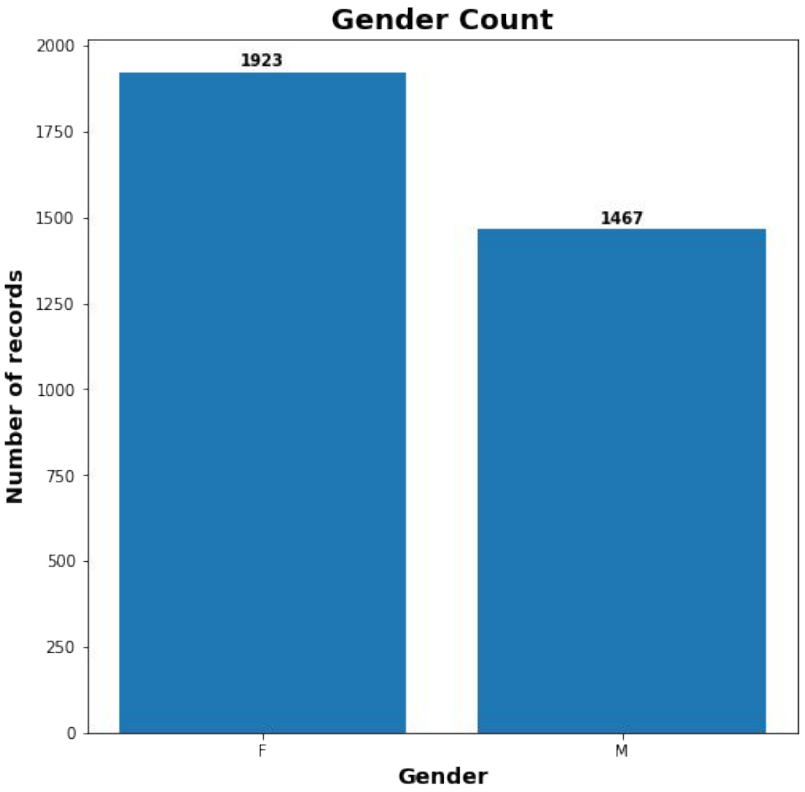
Target Variable Class Proportion

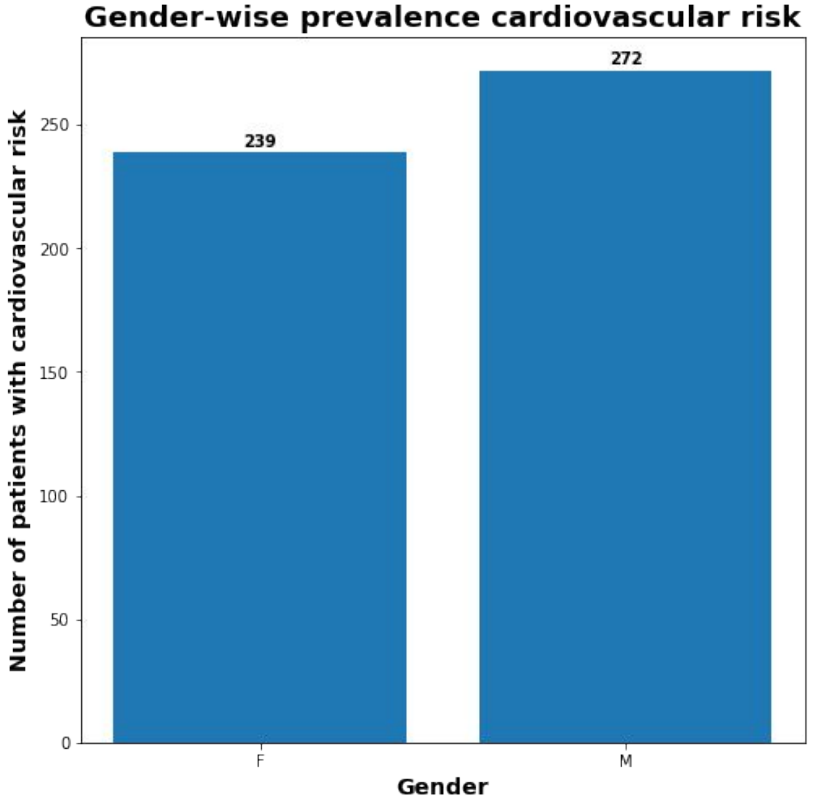


Age-wise count of people with CHD risk



Gender count and gender-wise risk prevalence

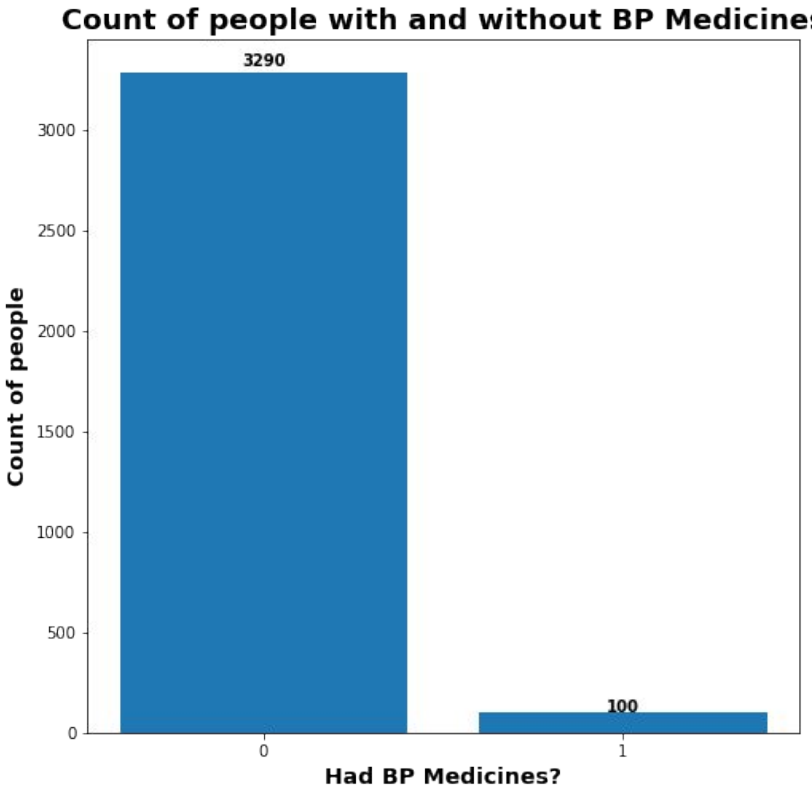


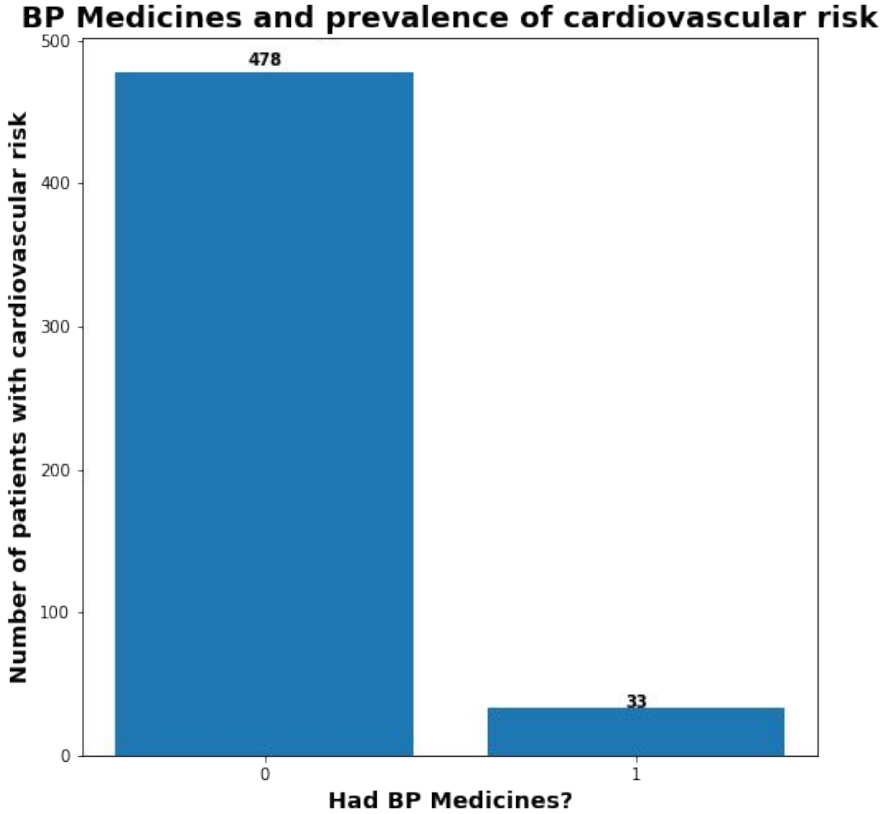


Similar analysis for Education, Smoking Habits and Cigarettes per day

- The education level and risk prevalence shows correlation, as education level increases the number of people with risk have decreased
- The number of people with and without smoking habits are almost equal in the sample, but there are higher number of people with risk who have smoking habits
- Cigarettes per day and risk prevalence does not show any correlation

BP Medicines vs. Risk Prevalence

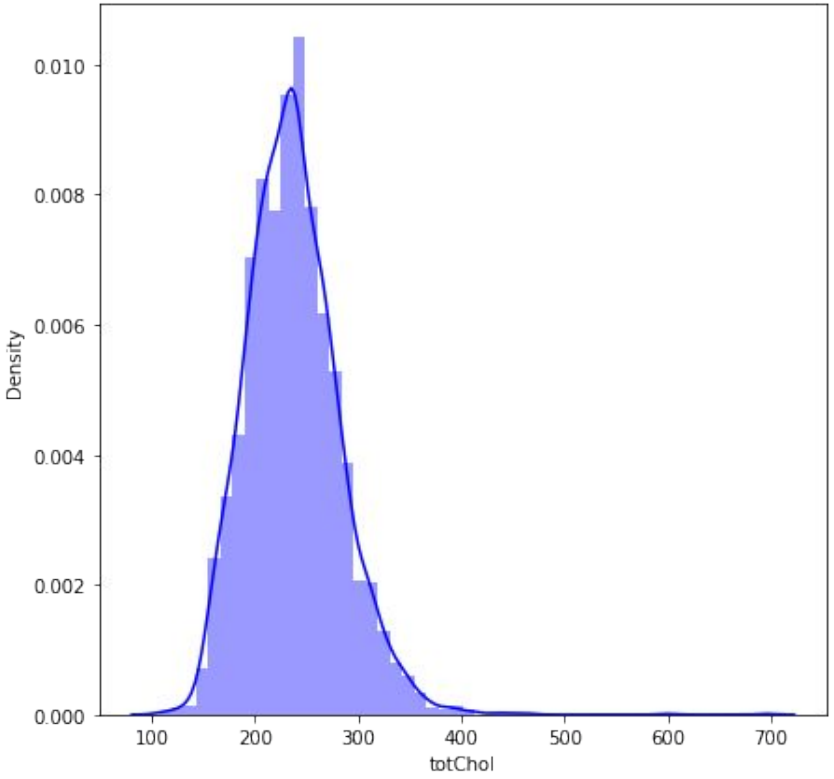


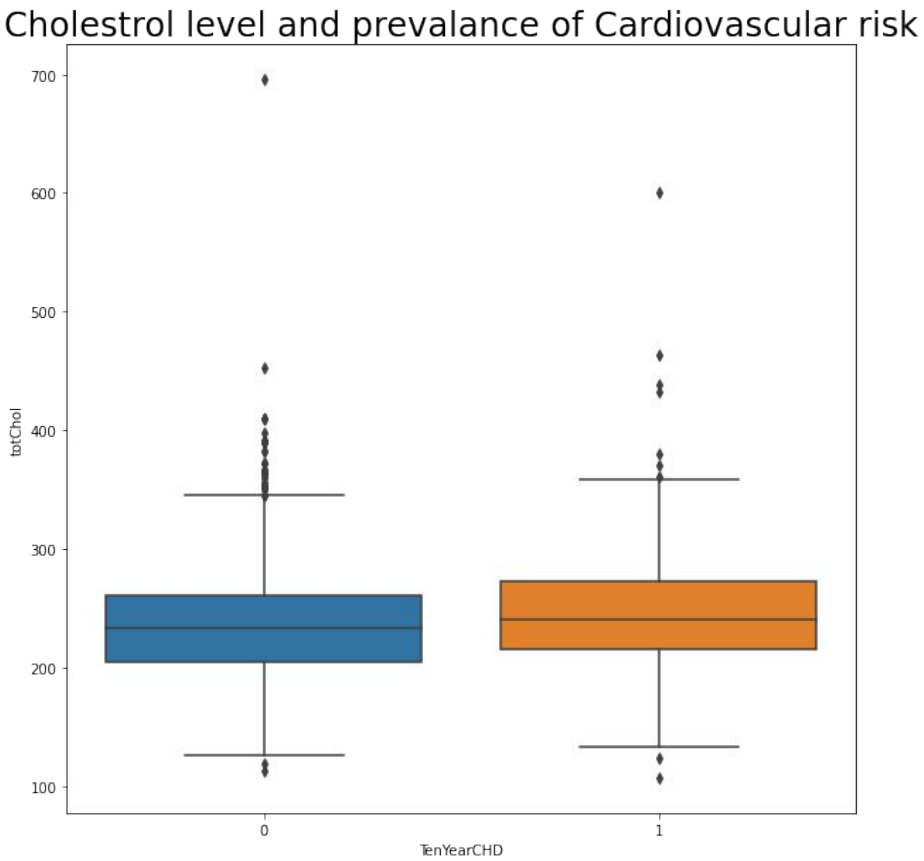


Similar analysis for Stroke History, Hypertension History, Diabetes History

- The sample size of people with stroke history is very small as compared to people without stroke history, thus unable to come up with any generalized conclusion
- Cardiovascular risk for people with hypertension is on higher side as compared to people with no history of hypertension
- Disproportionate sample size of people with and without diabetes history, cannot draw any generalized conclusion

Cholesterol level vs. Cardiovascular risk prevalence

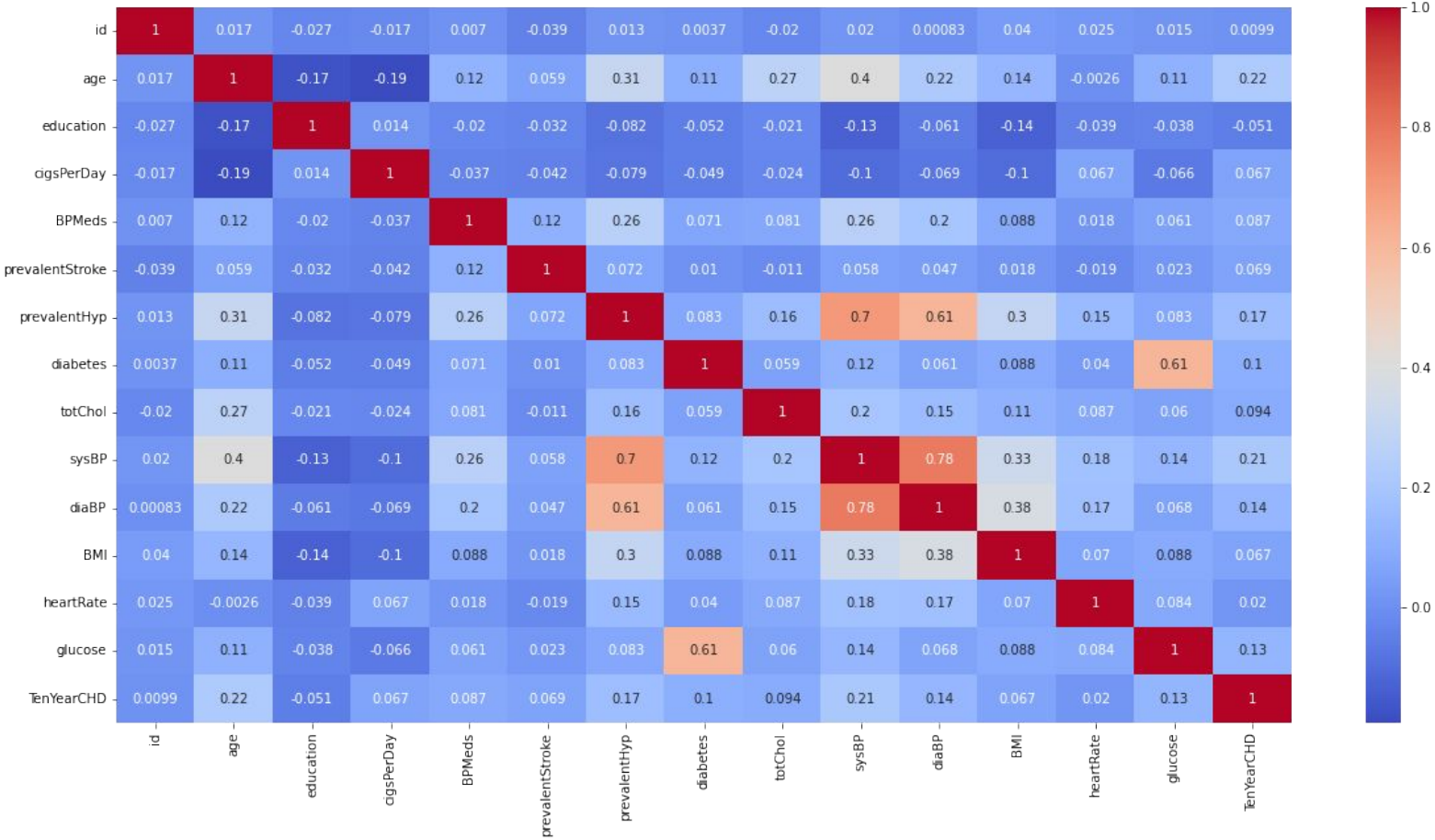




Similar analysis for Systolic and Diastolic BP, BMI, Heart Rate and Glucose level

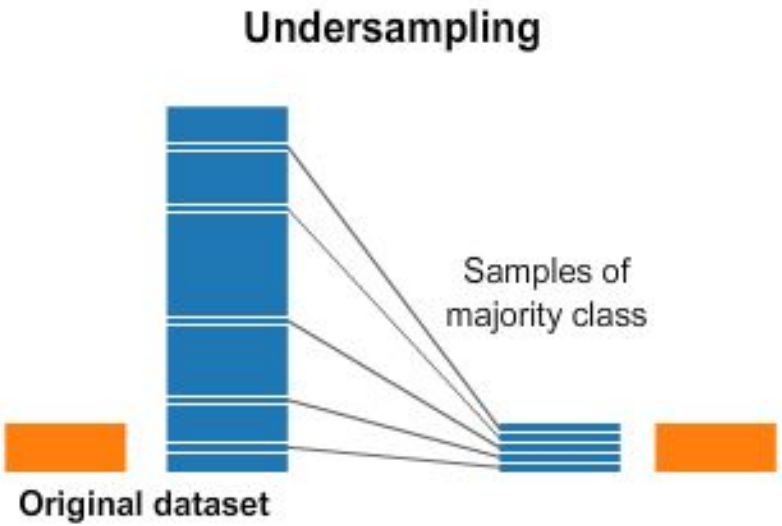
- People with higher systolic and diastolic blood pressure tend to have higher risk of cardiovascular disease, although there are exceptions
- BMI distribution of both people with and without risk prevalence is almost similar
- Mean heart rate for both the classes (having and not having cardiovascular risk) is almost same, seems like the Heart Rate has very less effect on cardiovascular risk prevalence
- Glucose level distribution is almost same for both the classes and there are too many outliers in both cases

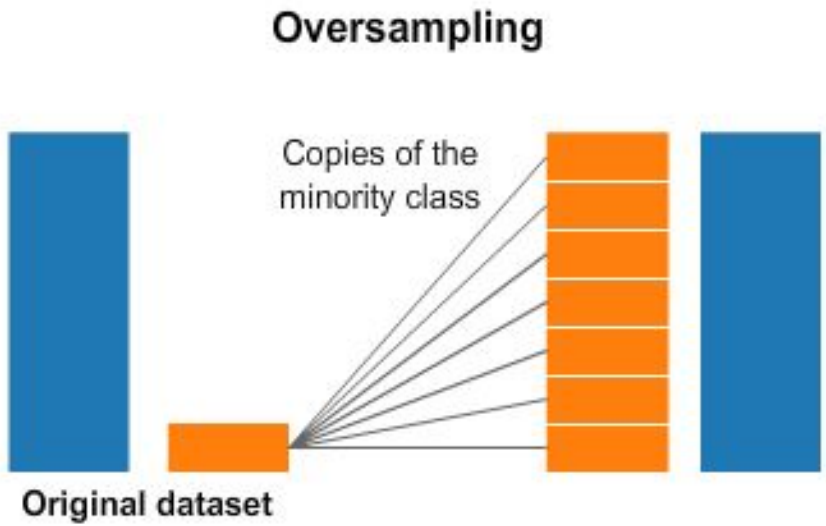
Correlation Matrix



Handling Imbalanced Dataset
And
Model Training

Resampling





SMOTE (Synthetic Minority Over Sampling Technique)

Models Trained

1. Logistic Regression Classifier
2. Support Vector Machine Classifier
3. Decision Tree Classifier
4. K Nearest Neighbor Classifier

Performance Metrics

1. Accuracy
2. Precision
3. Recall
4. F1 Score

Performance Comparison

	Model	Accuracy	Precision	Recall	F1 Score
2	K Nearest Neighbour	0.872222	0.872498	0.872222	0.872278
1	Support Vector Machines	0.858333	0.858523	0.858333	0.858294
0	Decision Tree	0.812500	0.819507	0.812500	0.812826
3	Logistic Regression	0.681944	0.682437	0.681944	0.682107

Challenges

- Imbalance Dataset
- Imbalanced variables like BP Medicines, Stroke History etc.
- Longer execution time

Conclusion

- Males have higher CHD risk
- Smoking habits may lead to higher CHD risk
- People with hypertension have higher risk of CHD
- For the given classification problem, out of the four tried algorithms **‘K Nearest Neighbour’** classifier has the best performance