Capstone Project -3

Cardiovascular Risk Prediction

Individual Project

Kalyani Nikam

Problem Statement

- **The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10year risk of future coronary heart disease (CHD).**

Data Description

The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variables Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Demographic:

- • **Sex**: male or female("M" or "F")
- •** Age**: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) Behavioral
- • **is smoking**: whether or not the patient is a current smoker ("YES" or "NO")
- • **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Data Description (Continued)

Medical(history):

- **BP Meds**: whether or not the patient was on blood pressure medication (Nominal)
- **Prevalent Stroke**: whether or not the patient had previously had a stroke (Nominal)
- **Prevalent Hyp**: whether or not the patient was hypertensive (Nominal)
- **Diabetes**: whether or not the patient had diabetes (Nominal)

Medical(current):

- **Tot Chol**: total cholesterol level (Continuous)
- **Sys BP**: systolic blood pressure (Continuous)
- **Dia BP**: diastolic blood pressure (Continuous)
- **BMI**: Body Mass Index (Continuous)
- **Heart Rate**: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
- **Glucose**: glucose level (Continuous)

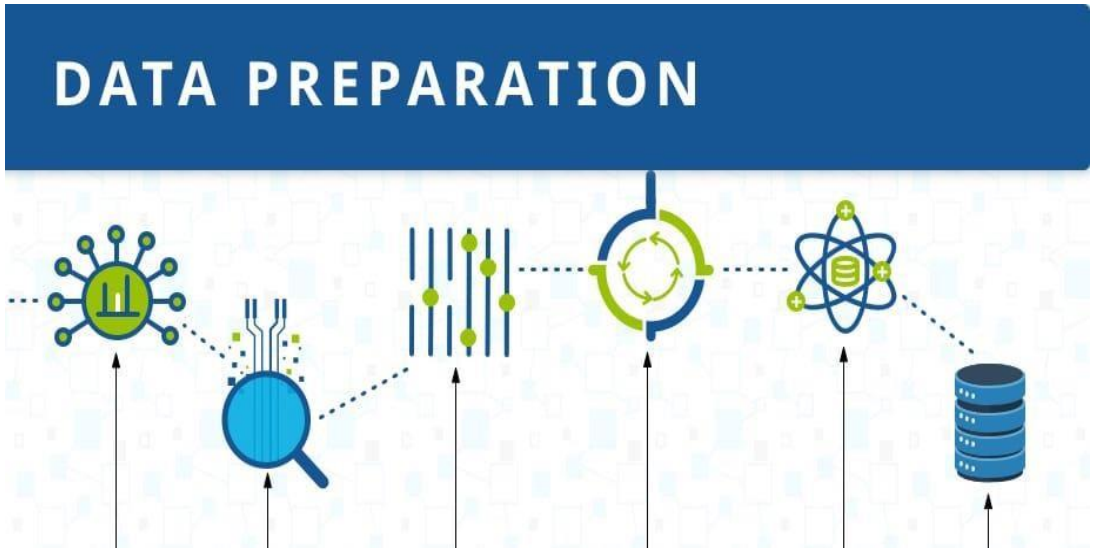
Predict variable (desired target)

- 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

Data Preparation

- **Handling missing values**
- Looking for percentage of null values of each column
- For columns containing large null values, replacing null with proper values.
- For columns containing small null values, dropping those nulls.

- **Handling duplicate values**
- No Duplicate Values In Our Dataset



- **Making of proper features**
- Removed column of very less correlated feature with target variable
- Kept only one column from strongly correlated independent variables
- Removed unnecessary features
- Standardization of independent features

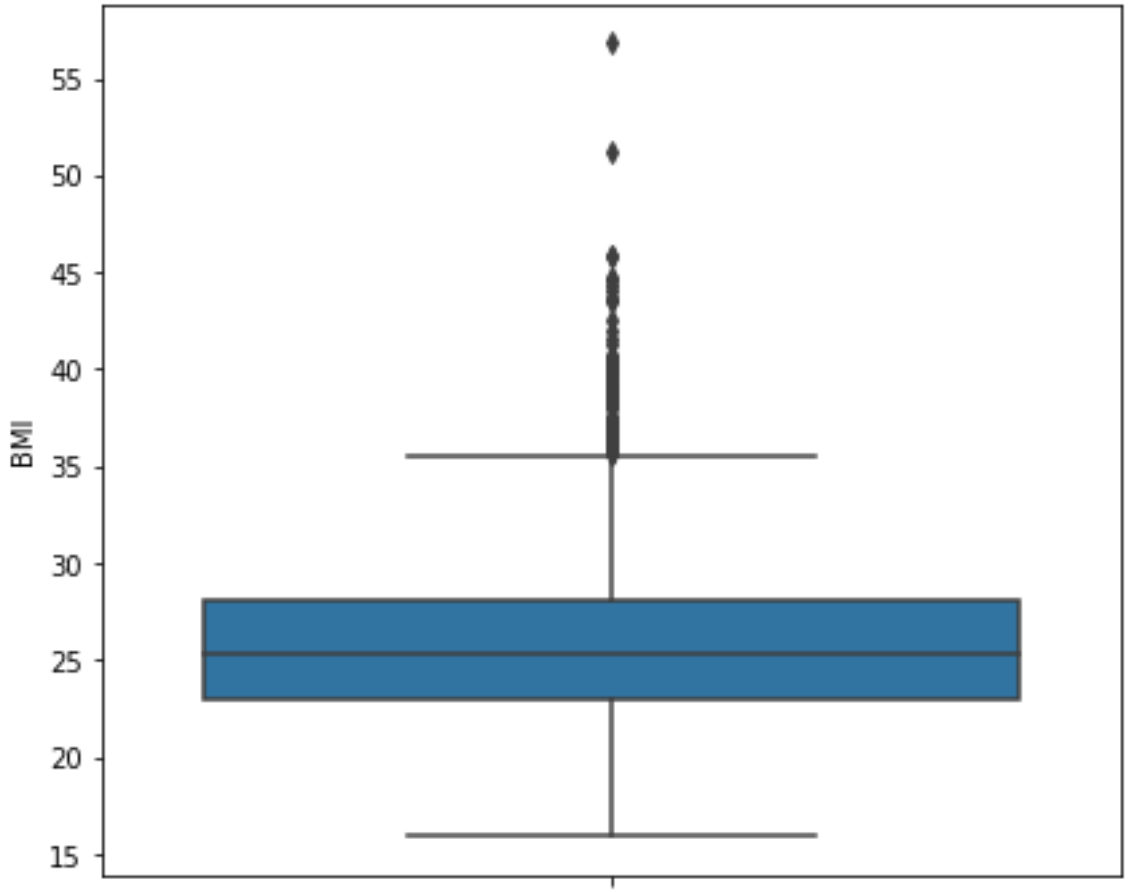
- **Handling Data Imbalance**
- Used Synthetic Minority Oversampling Technique, or SMOTE for short to handle data imbalance.

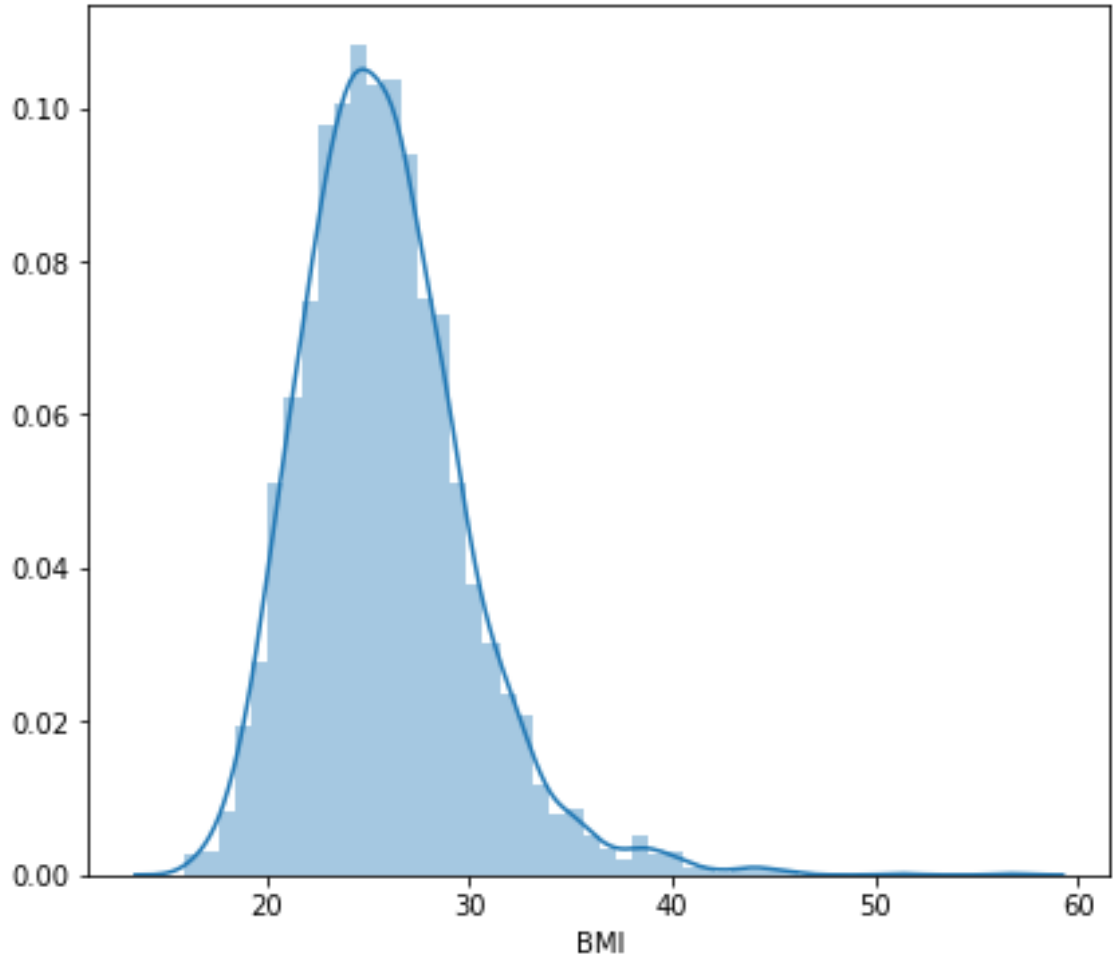
Insights Into Data

Range Index:3390 entries, 0 to 3389. Data columns (total 17 columns):

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	1	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	2	46	1.0	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	3	50	1.0	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	4	64	1.0	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0

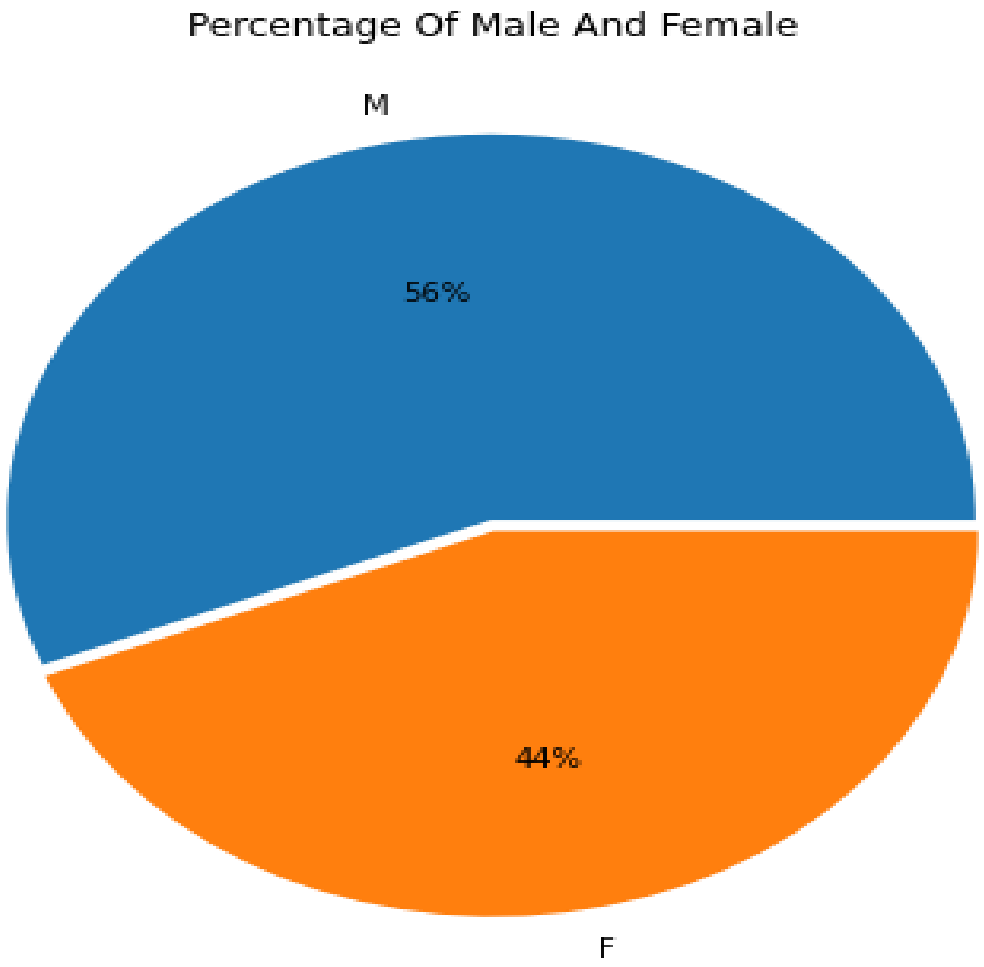
EDA- Boxplot And Distribution Plot Of BMI



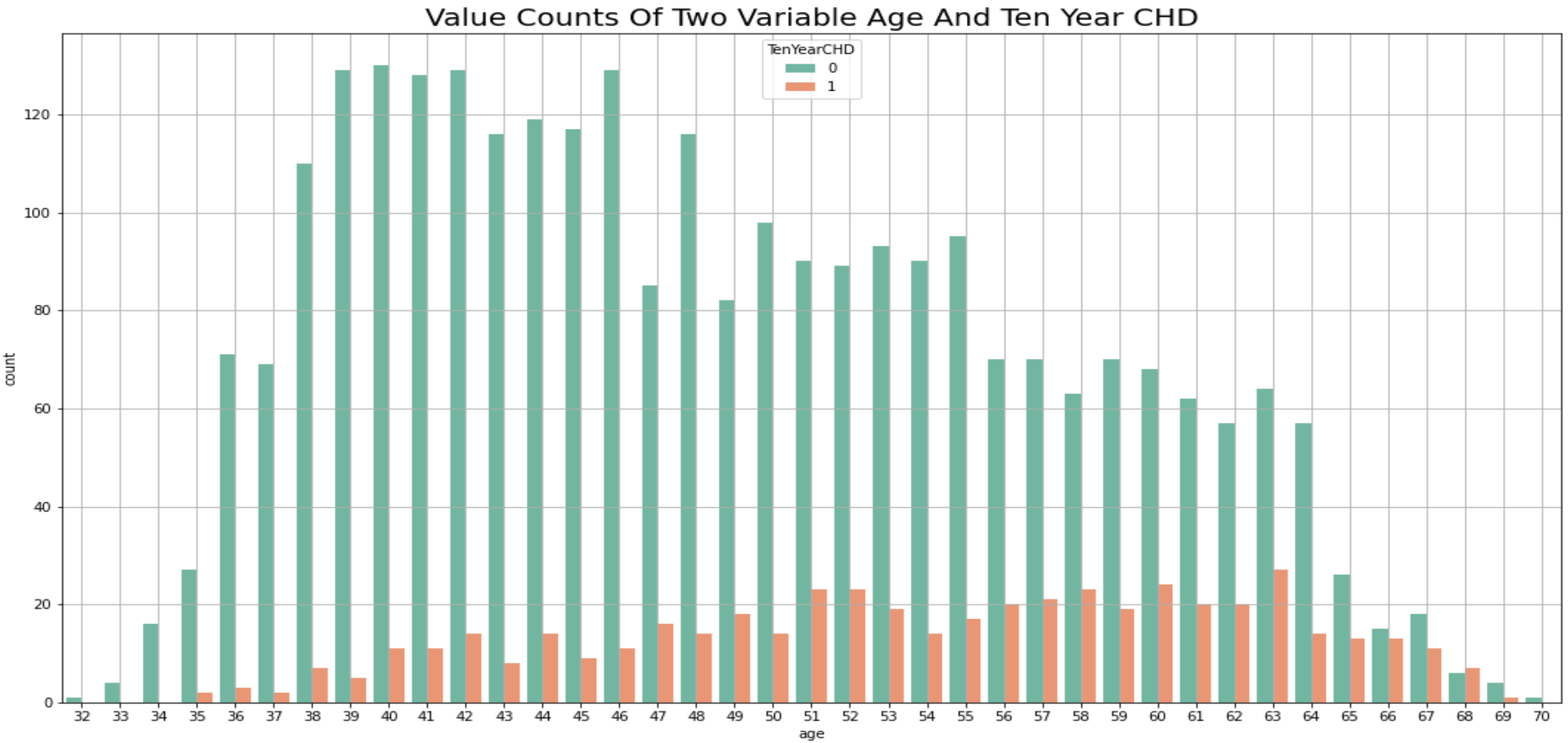


From Above Plot's We Can See That BMI Are Positively Skewed With Median BMI Of 25.38

Pie Plot of Male And Female %

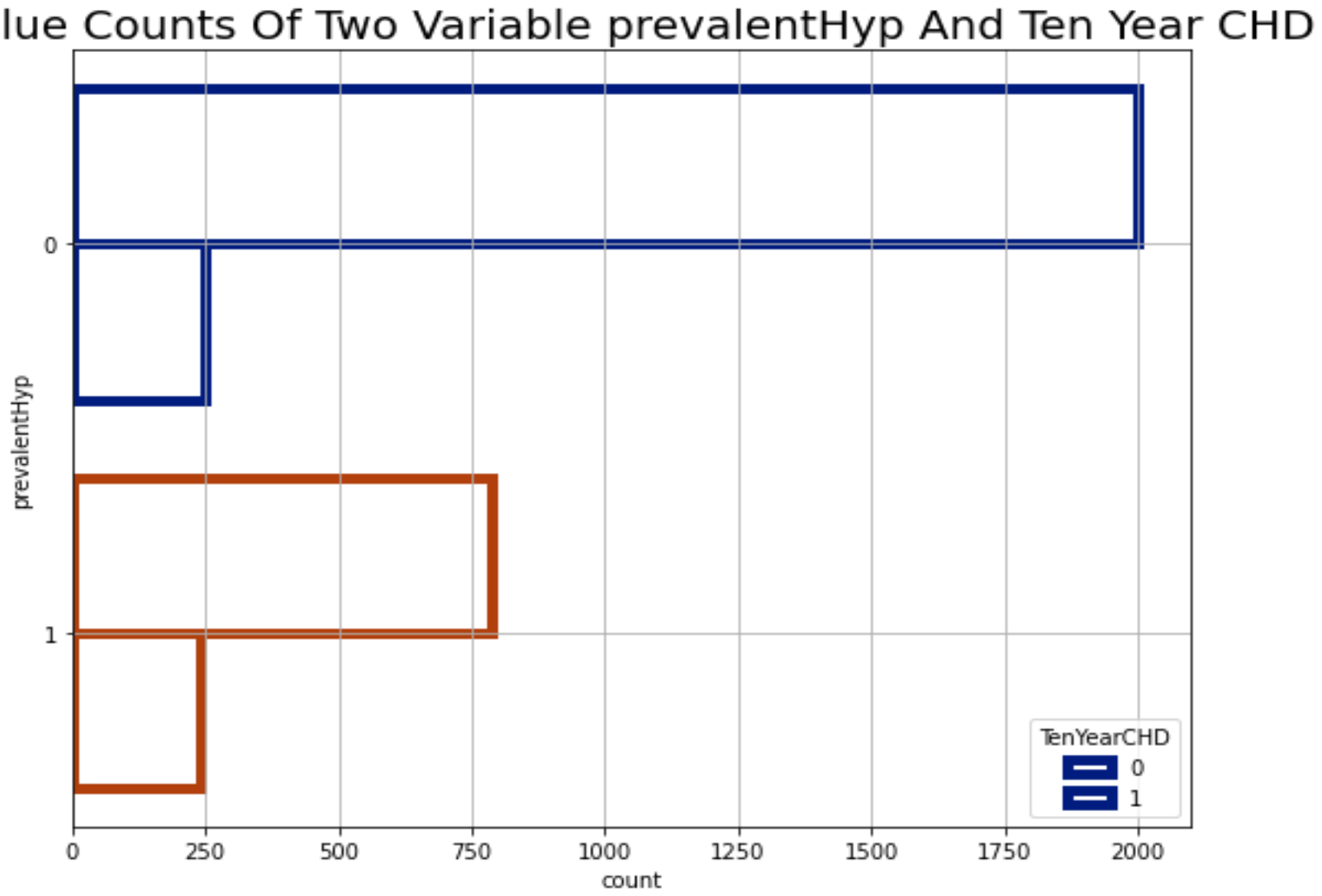


We ***We can see that Male Percentage Is Slightly Higher Than Female Percentage***



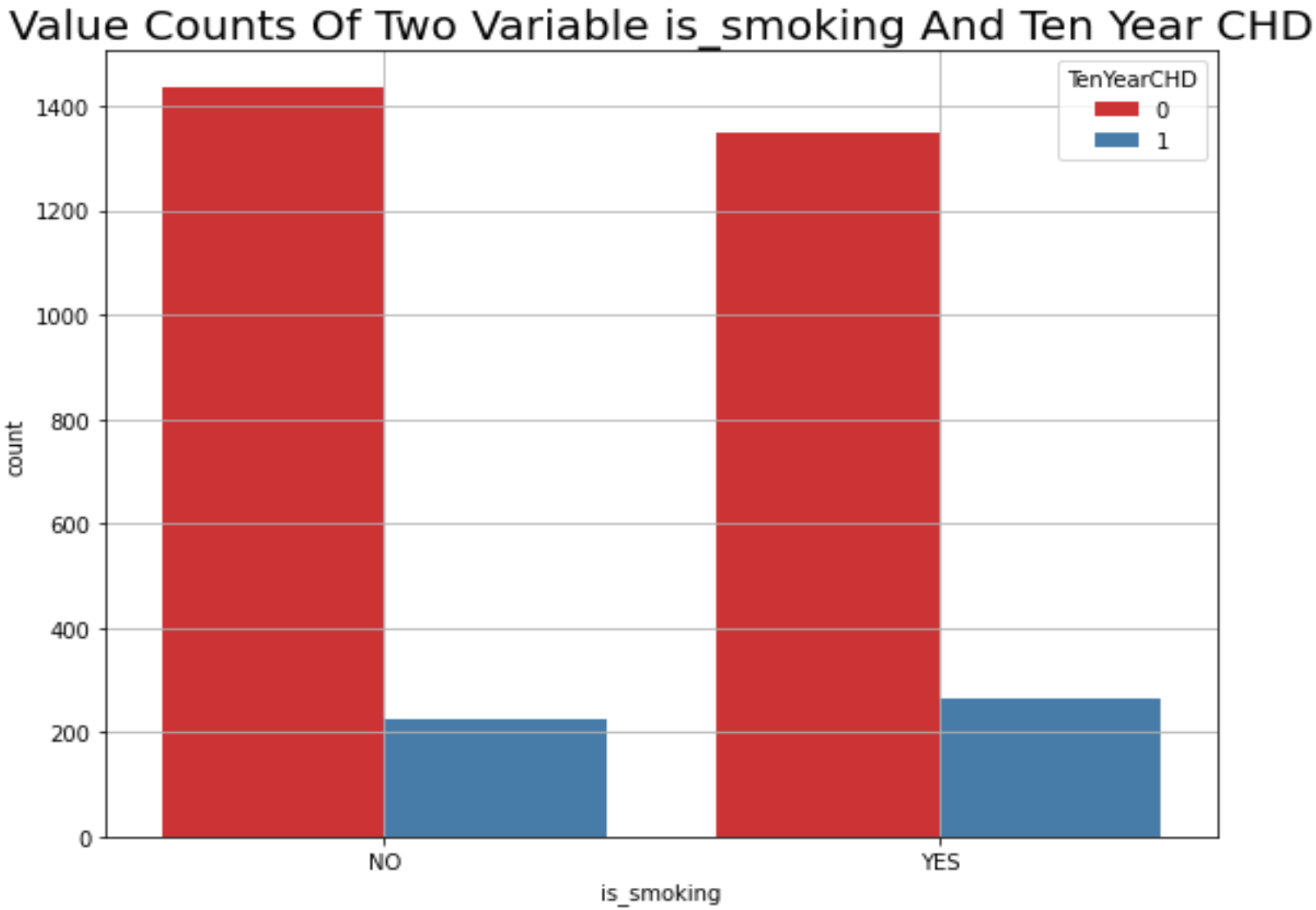
From above plot we can see that as age increases proportion of people having CHDalso increases

Count Plot Of prevalentHyp And Ten Year CHD



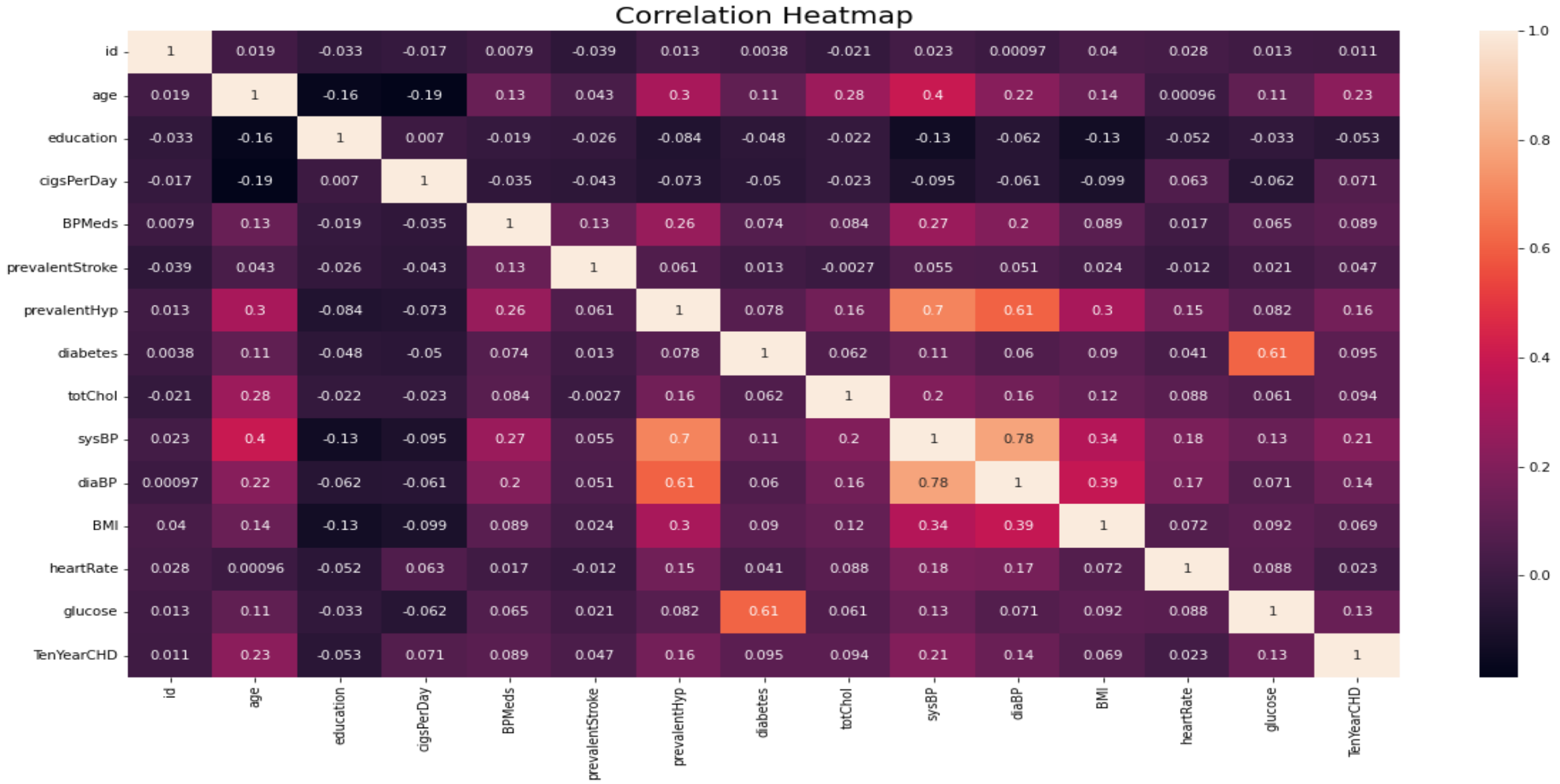
proportion of prevalentHyp people having CHD is larger than non prevalentHyp people

Count Plot Of Is Smoking And Ten Year CHD



Proportion of smokers having CHD is slightly more than non smokers

Correlation Heatmap



diaBP is positively correlated with SysBP with correlation value of 0.78

Model's Used

- Logistic Model
- Decision Tree
- Decision Tree With Hyperparameter Tuning
- Random Forest
- Random Forest With Hyperparameter Tuning
- XG Boost
- XG Boost With Hyperparameter Tuning
- KNN
- KNN With Hyperparameter Tuning
- Naive Bayes Classifier
- Naive Bayes Classifier With Hyperparameter Tuning
- SVM
- SVM With Hyperparameter Tuning

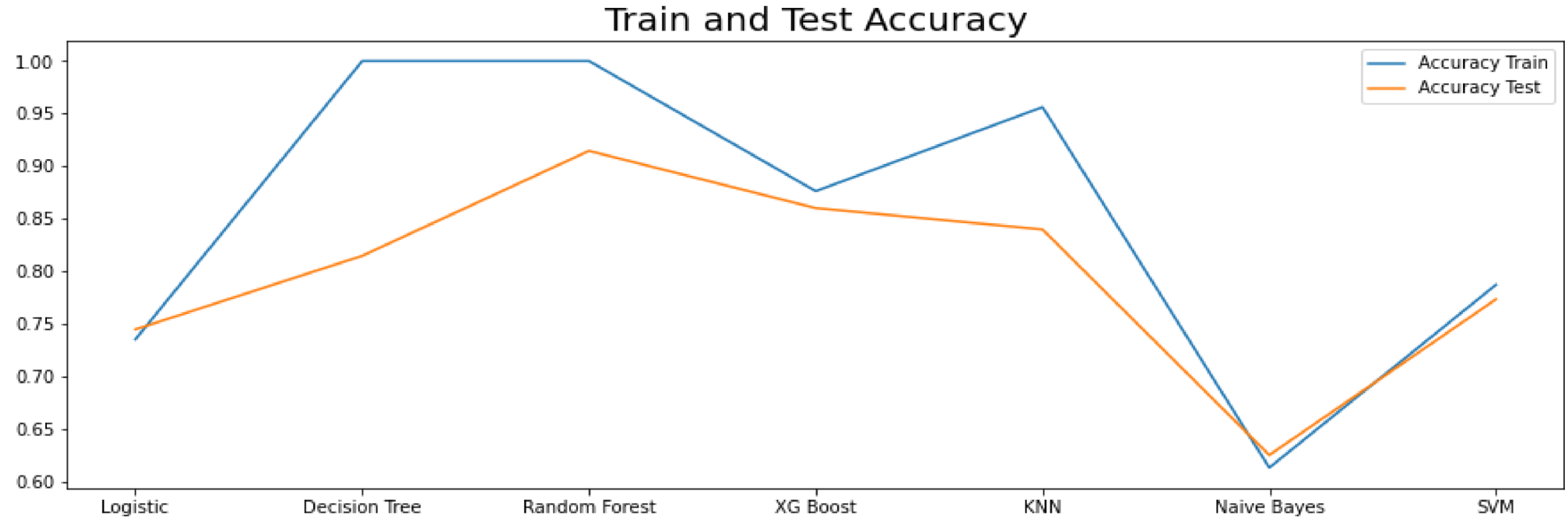
Evaluation of all models without Hyperparameter Tuning

	Logistic	Decision Tree	Random Forest	XG Boost	KNN	Naive Bayes	SVM
Accuracy Train	0.734924	1.000000	1.000000	0.876059	0.956120	0.612779	0.787016
Accuracy Test	0.744464	0.814482	0.914423	0.859964	0.839617	0.624776	0.773190
ROC AUC Train	0.734358	1.000000	1.000000	0.875420	0.955791	0.610827	0.786469
ROC AUC Test	0.745728	0.814489	0.914864	0.861418	0.840067	0.629452	0.774441
Precision	0.751102	0.814522	0.915706	0.871349	0.840900	0.682978	0.780228
Recall	0.744464	0.814482	0.914423	0.859964	0.839617	0.624776	0.773190
F1 Score	0.743280	0.814492	0.914400	0.859130	0.839565	0.596097	0.772175

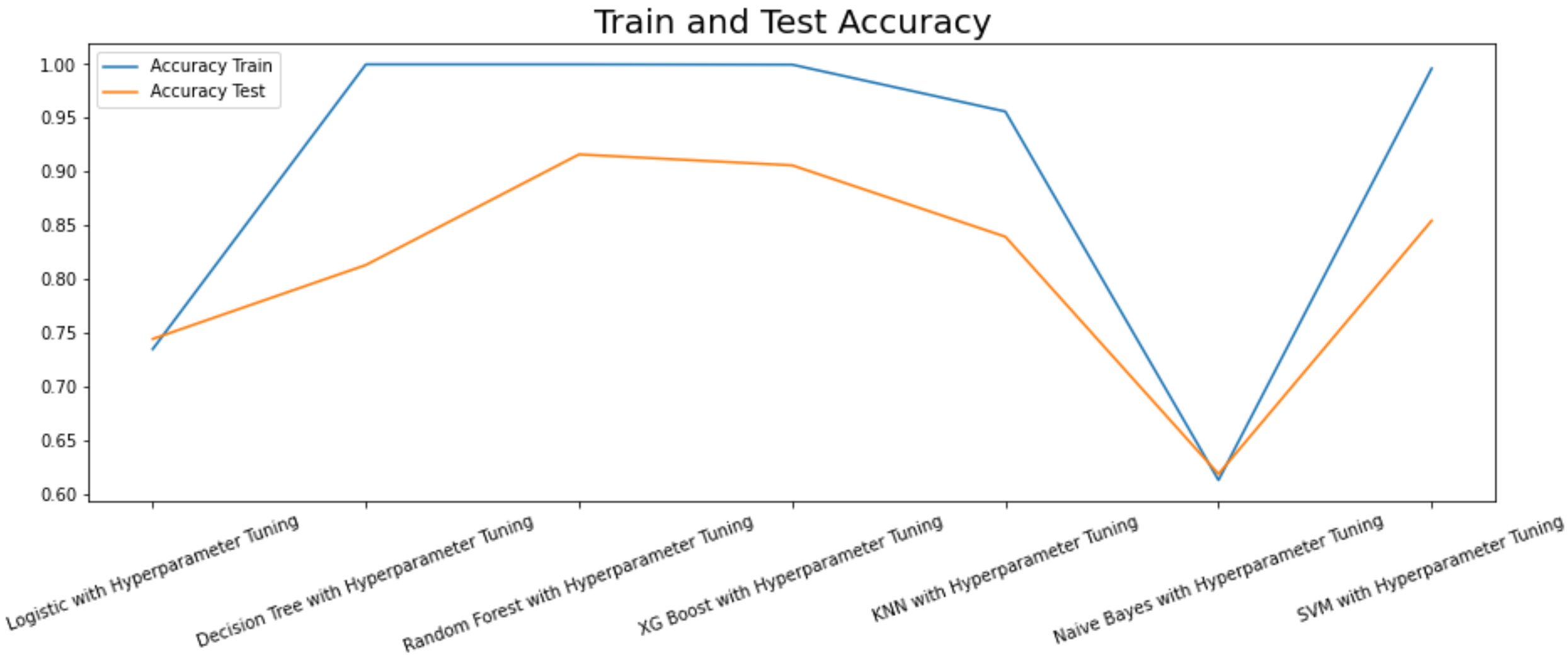
Evaluation of all models with Hyperparameter Tuning

	Logistic with Hyperparameter Tuning	Decision Tree with Hyperparameter Tuning	Random Forest with Hyperparameter Tuning	XG Boost with Hyperparameter Tuning	KNN with Hyperparameter Tuning	Naive Bayes with Hyperparameter Tuning	SVM with Hyperparameter Tuning
Accuracy Train	0.734924	1.000000	1.000000	0.999743	0.956120	0.613036	0.996151
Accuracy Test	0.744464	0.813285	0.916218	0.906044	0.839617	0.618791	0.854578
ROC AUC Train	0.734350	1.000000	1.000000	0.999741	0.955791	0.611013	0.996153
ROC AUC Test	0.745748	0.813001	0.916670	0.906587	0.840067	0.623715	0.854212
Precision	0.751303	0.813418	0.917561	0.907916	0.840900	0.683033	0.854974
Recall	0.744464	0.813285	0.916218	0.906044	0.839617	0.618791	0.854578
F1 Score	0.743237	0.813210	0.916194	0.905996	0.839565	0.586168	0.854481

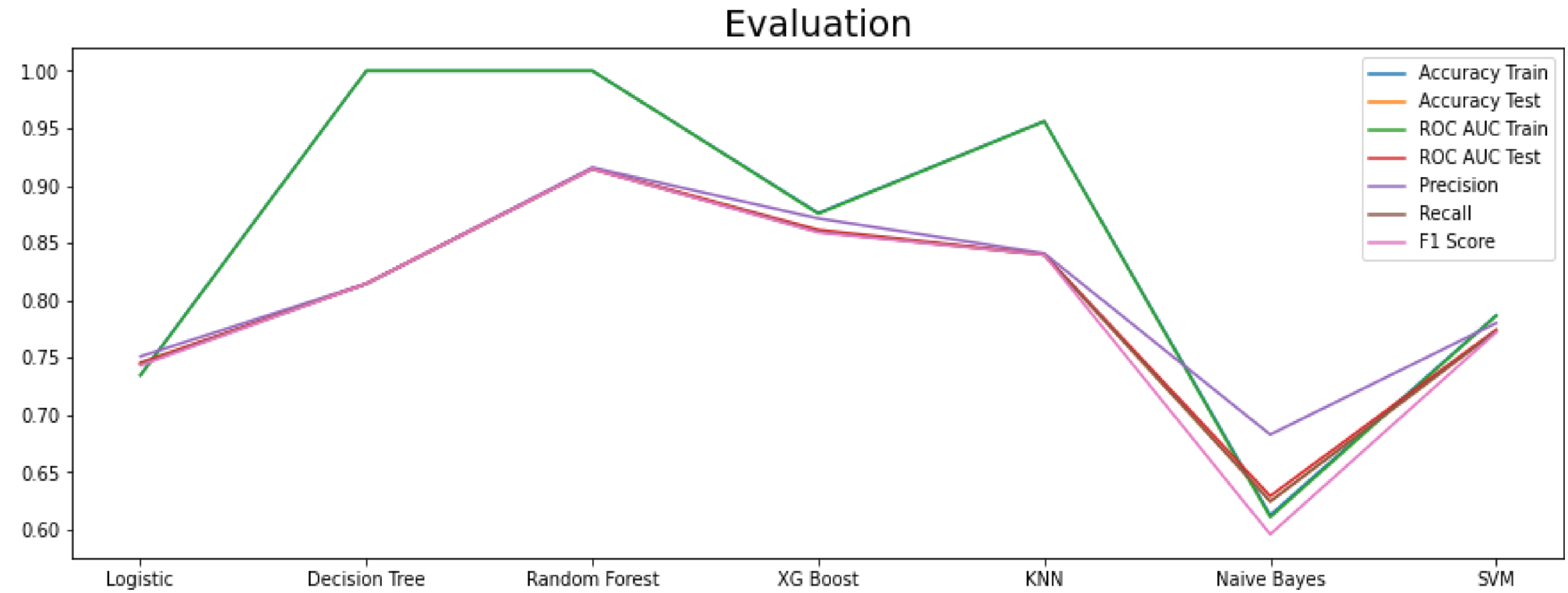
Train and Test Accuracy Of All Models Without Hyperparameter Tuning



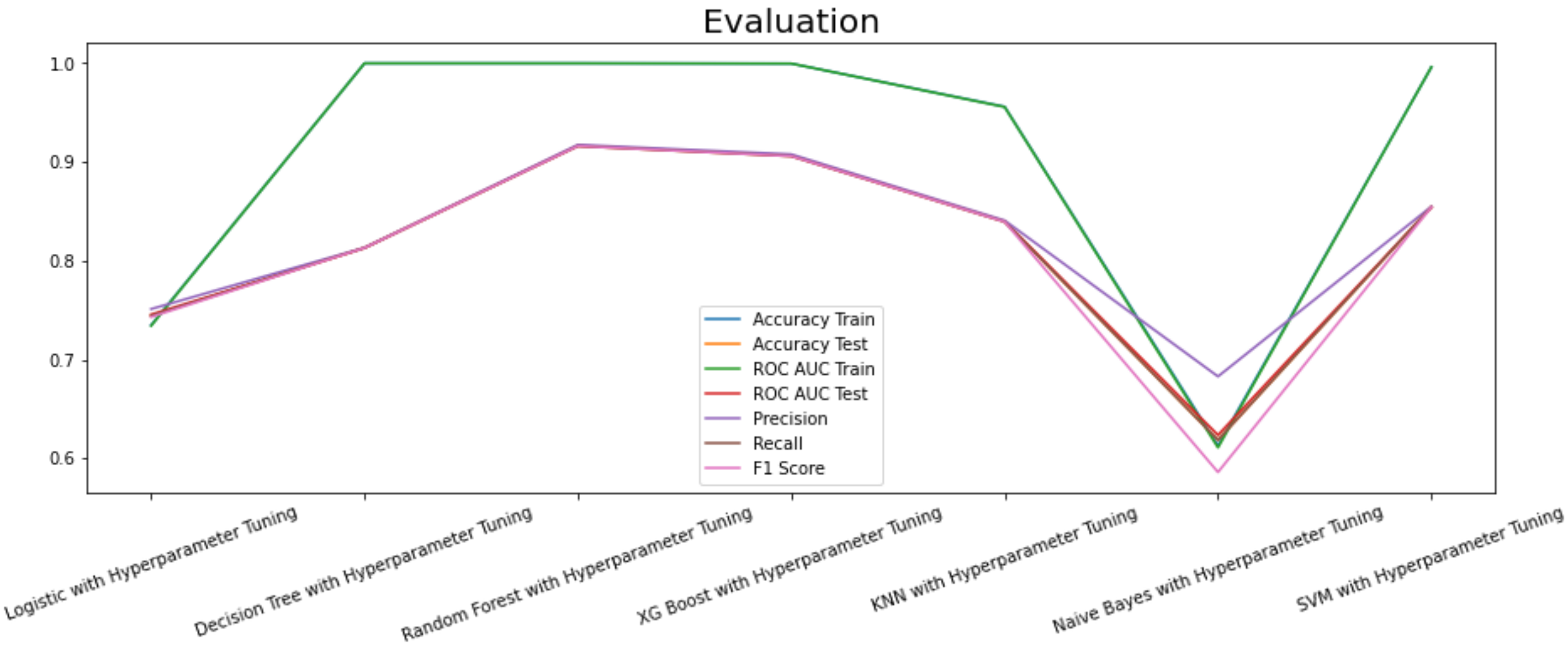
Train and Test Accuracy Of All Models With Hyperparameter Tuning



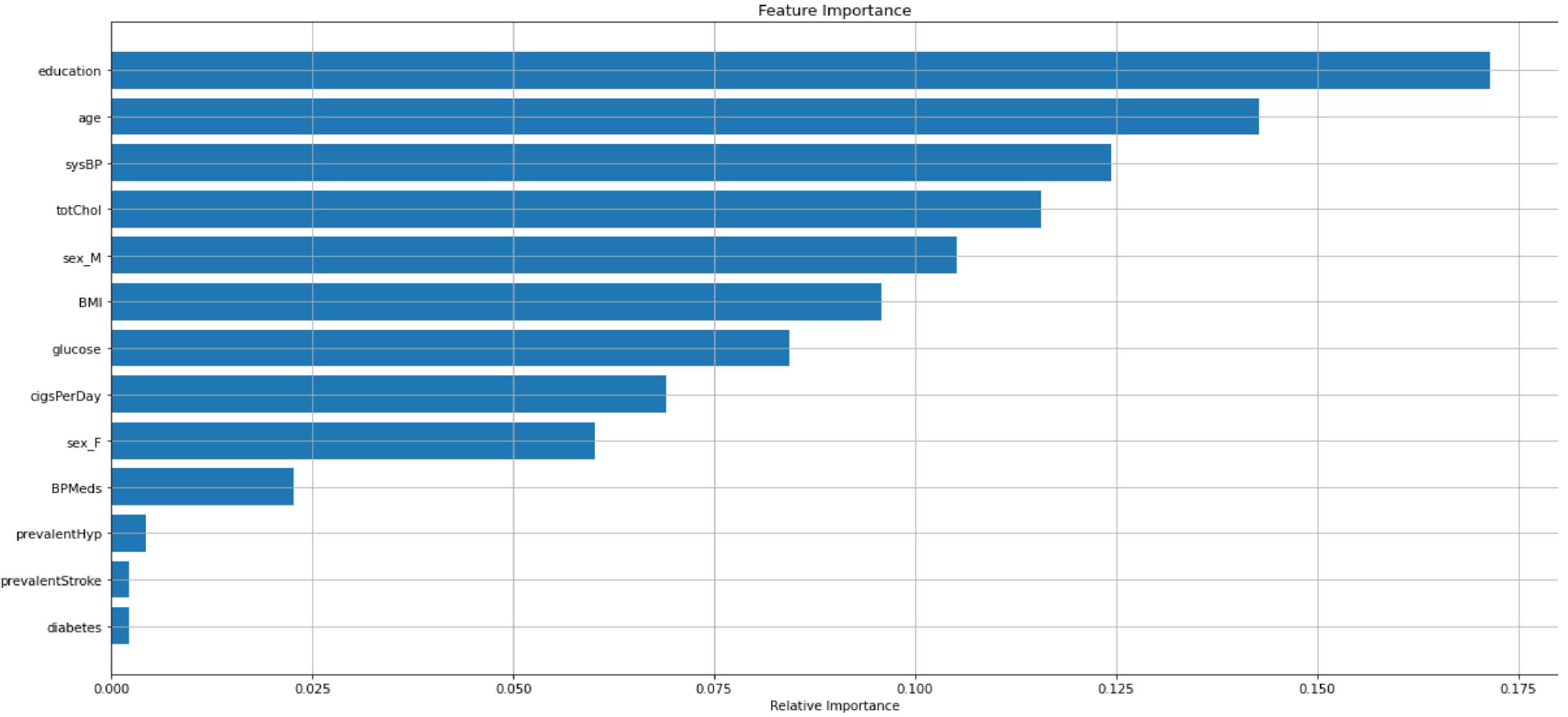
Line Plot Of Different Types of Evaluations Of All Models Without Hyperparameter Tuning



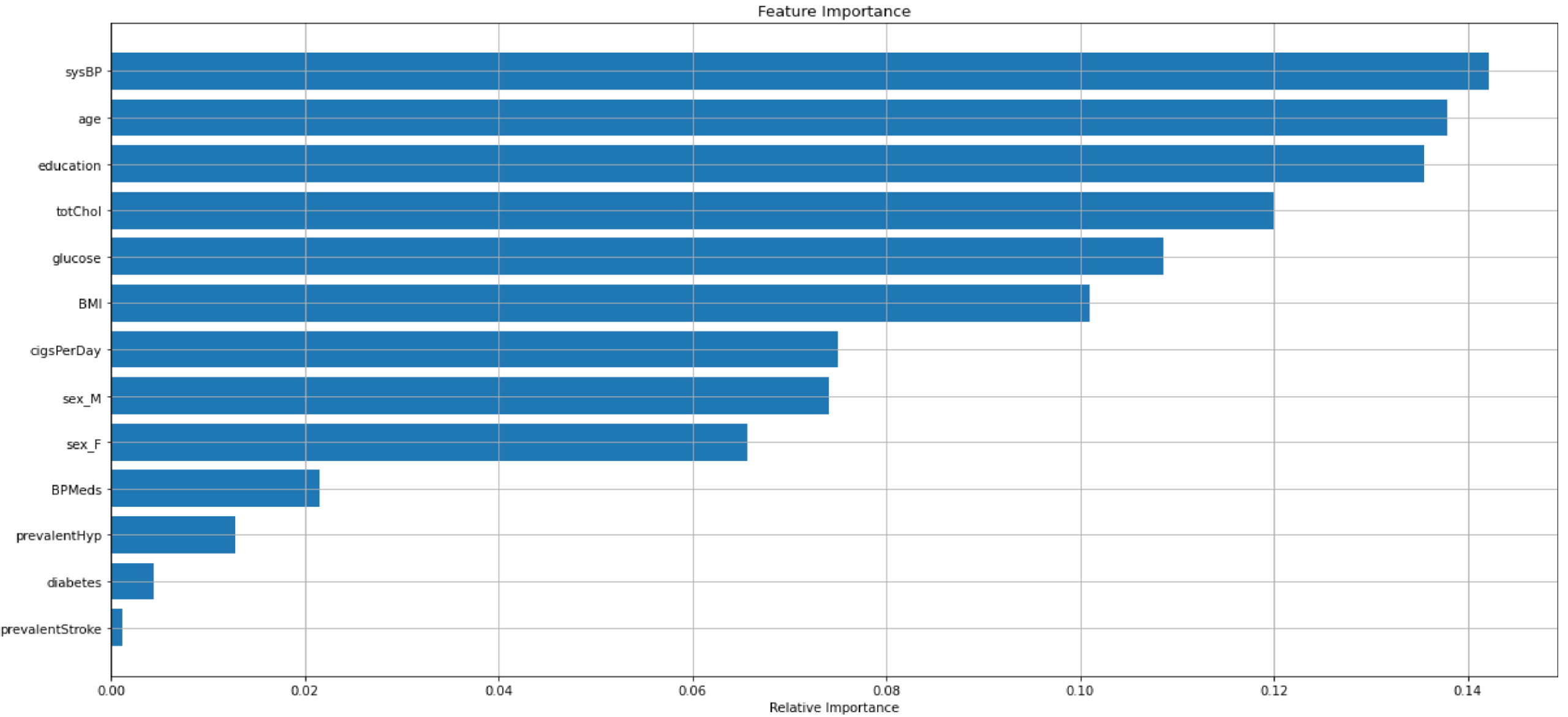
Line Plot Of Different Types of Evaluations Of All Models With Hyperparameter Tuning



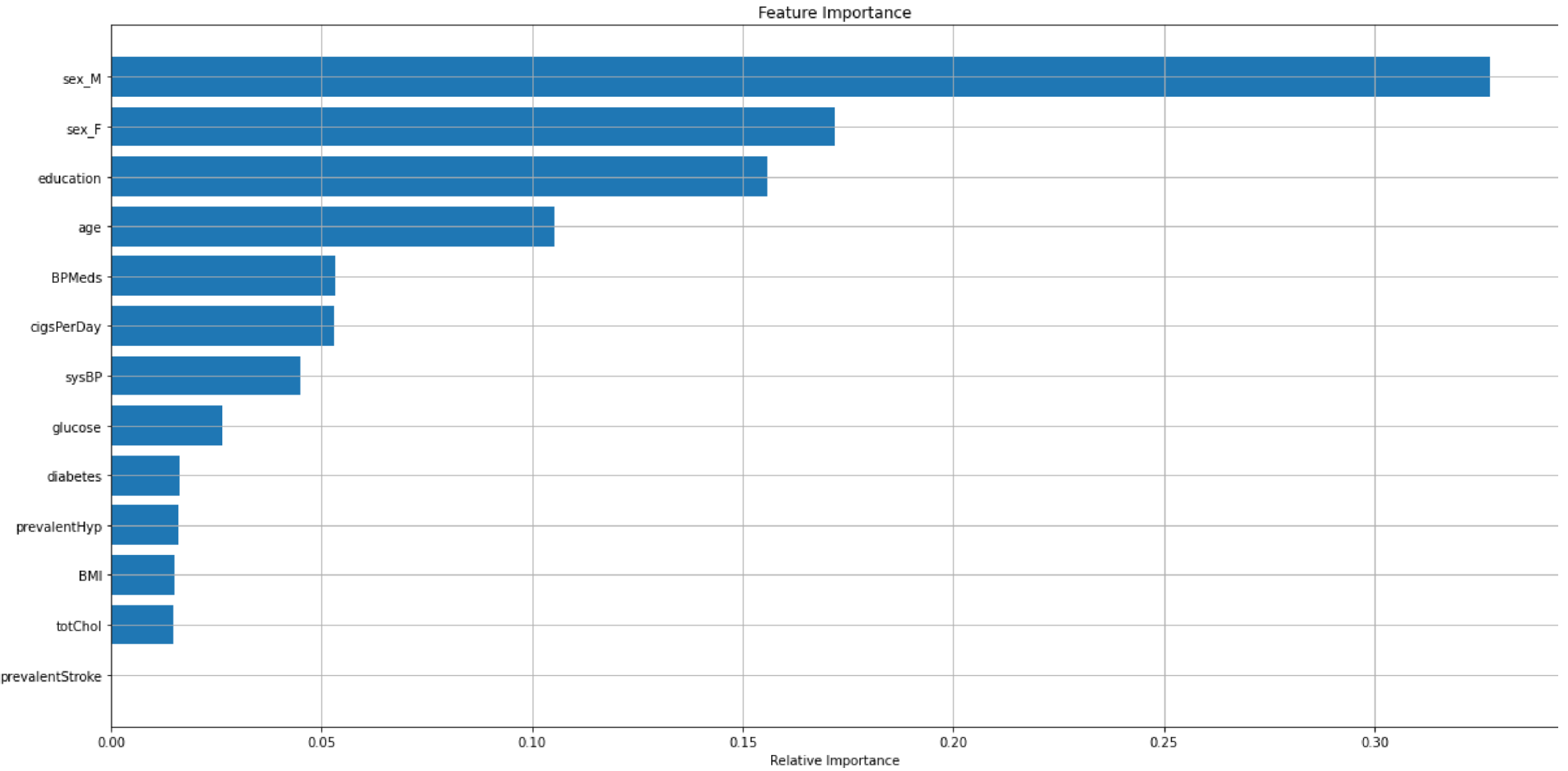
Features Importance Of Decision Tree Model



Features Importance Of Random Forest Model



Features Importance Of XG Boost Model



Conclusion's

- *Logistic Model Has Accuracy Of 74%*
- *Logistic with Hyperparameter Tuning Has Accuracy Of 74%*
- *Decision Tree Has Accuracy Of 81%*
- *Decision Tree with Hyperparameter Tuning Has Accuracy Of 81%*
- *Random Forest Has Accuracy Of 91%*
- *Random Forest with Hyperparameter Tuning Has Accuracy Of 92%*
- *XG Boost Has Accuracy Of 86%*
- *XG Boost with Hyperparameter Tuning Has Accuracy Of 91%*
- *KNN Has Accuracy Of 84%*
- *KNN with Hyperparameter Tuning Has Accuracy Of 84%*

Conclusion's (Continued)

- *Naive Bayes Has Accuracy Of 62%*
- *Naive Bayes with Hyperparameter Tuning Has Accuracy Of 62%*
- *SVM Has Accuracy Of 77%*
- *SVM with Hyperparameter Tuning Has Accuracy Of 85%*
- **From Above We Can Conclude That Random Forest With Hyperparameter Tuning And XG Boost With Hyperparameter Tuning Is The Best Fitted Model To Our Data.**
- **Random Forest With Hyperparameter Tuning has highest precision,recall and f1 score among all models.**
- *According to Random Forest Model SysBP, Age And education are the most important features which affects our Target variable.*


THANK
YOU