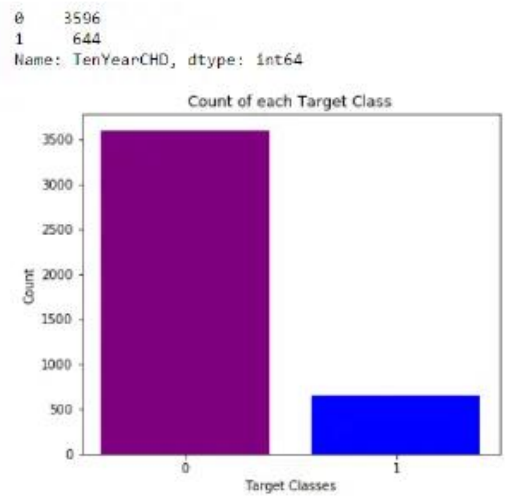
Visualizing and Predicting Heart Diseases with an Interactive Dash Board

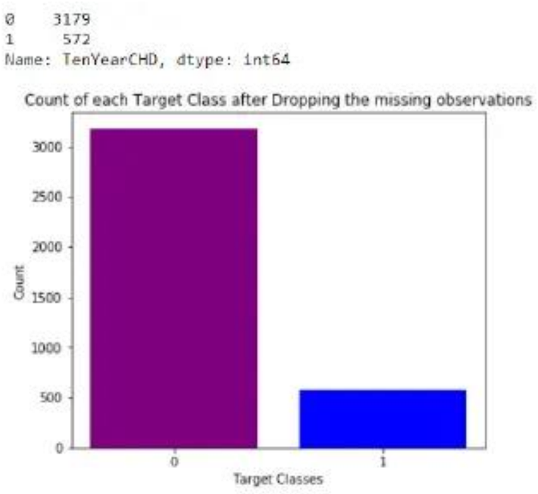
Problem Solution Fit

Data Preparation:

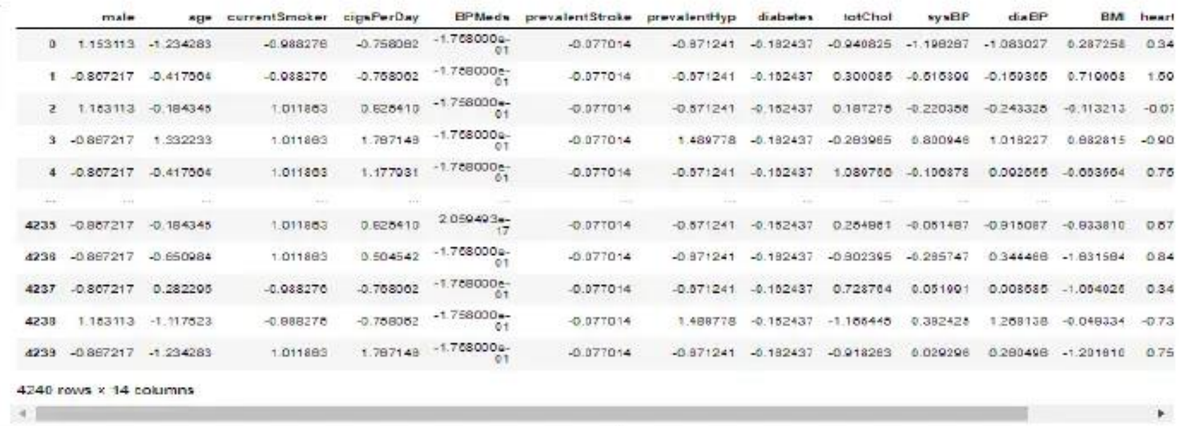
Since the dataset consists of 4240 observations with 388 missing data and 644 observations to berisked for heart disease, two different experiments were performed for data preparation. First, wechecked by dropping the missing data, leaving with only 3751 data and only 572 observationsrisked for heart disease.



```
0    3596
1     644
Name: TenYearCHD, dtype: int64
```



```
0    3179
1     572
Name: TenYearCHD, dtype: int64
```



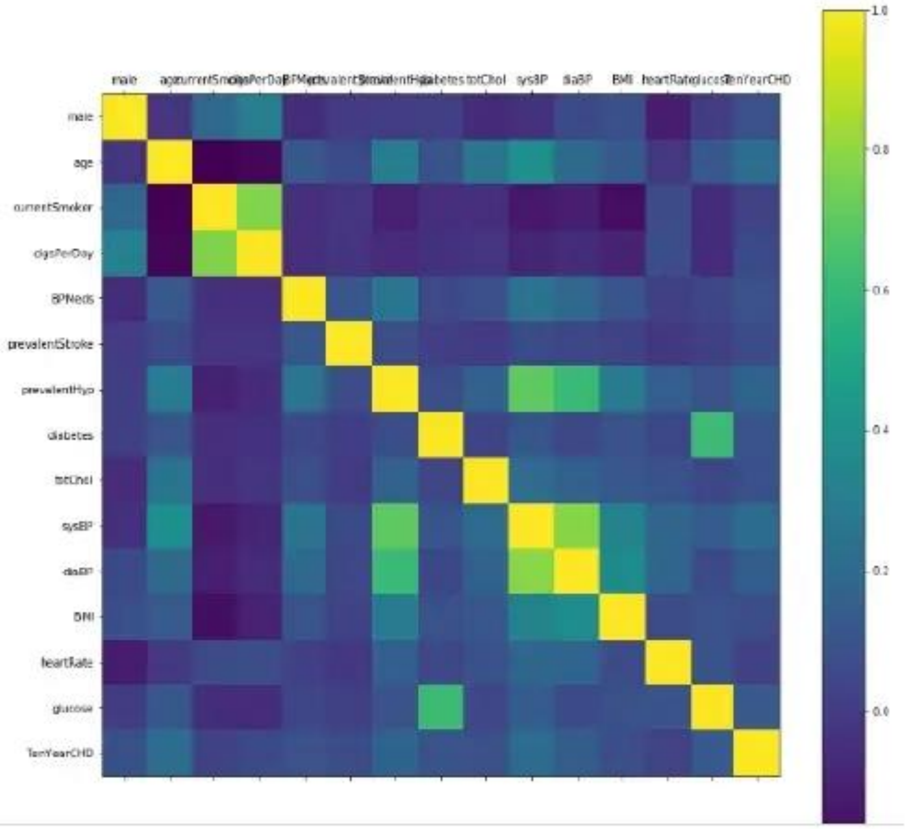
	male	age	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heart
0	1.153113	-1.234283	-0.988276	-0.758082	-1.768000e-01	-0.077014	-0.871241	-0.182437	-0.940825	-1.198287	-1.083027	0.287258	0.34
1	-0.867217	-0.417564	-0.988276	-0.758062	-1.768000e-01	-0.077014	-0.871241	-0.182437	0.300086	-0.515396	-0.159365	0.719668	1.59
2	1.153113	-0.184345	1.011863	0.926410	-1.758000e-01	-0.077014	-0.871241	-0.182437	0.187276	-0.220356	-0.243326	-0.113213	-0.07
3	-0.867217	1.332233	1.011863	1.787149	-1.768000e-01	-0.077014	1.489778	-0.182437	-0.283965	0.800946	1.018227	0.682815	-0.90
4	-0.867217	-0.417564	1.011863	1.177931	-1.768000e-01	-0.077014	-0.871241	-0.182437	1.089760	-0.106878	0.092665	-0.663664	0.76
...	...	...	...	...	...	...	...	...	...	...	...	...	
4235	-0.867217	-0.184345	1.011863	0.926410	2.059493e-17	-0.077014	-0.871241	-0.182437	0.264861	-0.081487	-0.916087	-0.833810	0.67
4236	-0.867217	-0.650984	1.011863	0.504542	-1.768000e-01	-0.077014	-0.871241	-0.182437	-0.802395	-0.285747	0.344468	-1.831584	0.84
4237	-0.867217	0.282295	-0.988276	-0.768062	-1.768000e-01	-0.077014	-0.871241	-0.182437	0.728764	0.051991	0.008585	-1.064026	0.34
4238	1.163113	-1.117523	-0.888276	-0.758062	-1.758000e-01	-0.077014	1.488778	-0.182437	-1.166446	0.382428	1.268138	-0.048334	-0.73
4239	-0.867217	-1.234283	1.011863	1.787149	-1.768000e-01	-0.077014	-0.871241	-0.182437	-0.918263	0.029296	0.280498	-1.201816	0.75

4240 rows × 14 columns

This leads to reduced number of the observations providing irrelevant training to our model. So,we progressed with imputation of data with the mean value of the observations and scaling themusing SimpleImputer and StandardScaler modules of Sklearn

Exploratory Analysis Document:

Correlation Matrix visualization Before Feature Selection shows

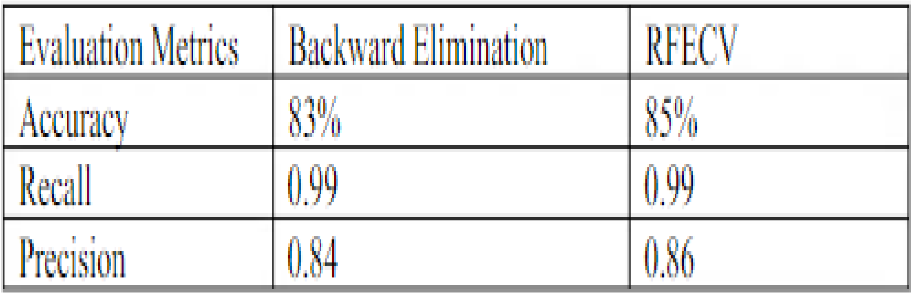


It shows that there is no single feature that has a very high correlation with our target value.Also, some of the features have a negative correlation with the target value and some have positive. The data was also visualized through plots and bar graphs.

Discussion On Results:

When performing various methods of feature selection, testing it was found that backwardelimination gave us the best results among others. The various methods tried were BackwardElimination with and without KFold, Recursive Feature Elimination with Cross Validation. Theaccuracy that was seen in them ranged around 85% with 85.5% being maximum. Though bothmethods gave similar accuracy but it was seen that in Backward Elimination we found that thenumber of misclassifications of True Negative was more and it was observed that the accuracy hadmore variance compared to RFEV. The precision of Backward Elimination and RFEV are 84%and 86% respectively. And the recalls are 0.99

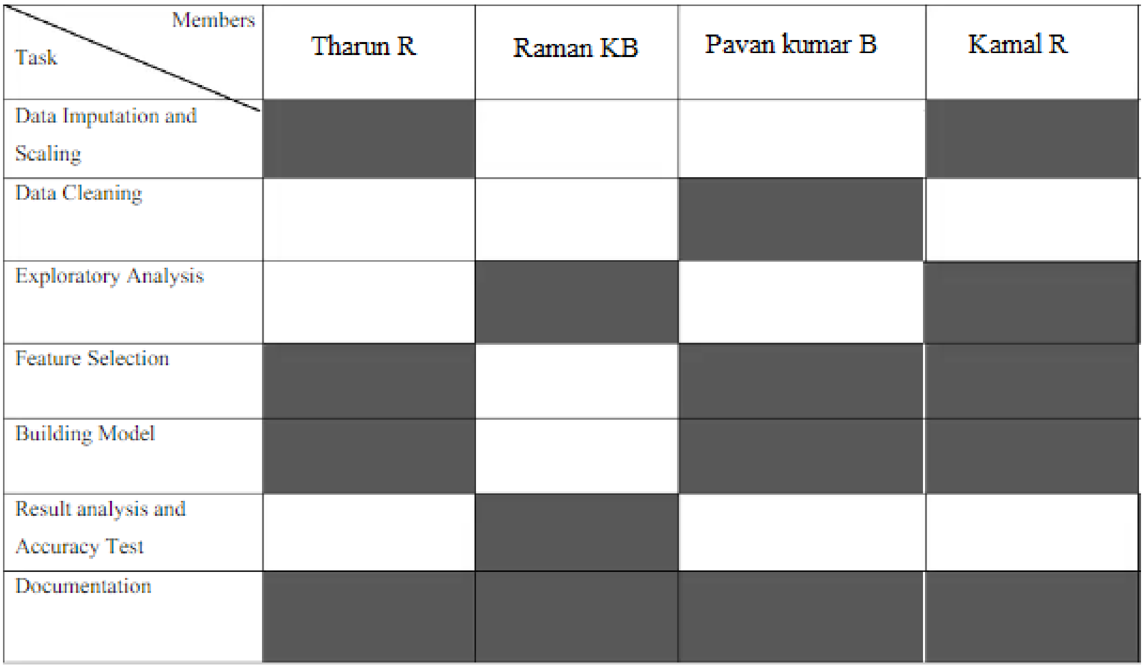
and 1 respectively. The precision and recall also shows that the number of misclassifications is less in RFECV than in Backward Elimination.



Evaluation Metrics	Backward Elimination	RFECV
Accuracy	83%	85%
Recall	0.99	0.99
Precision	0.84	0.86

Table 3: Comparison between the feature selection models after training and testing through LogisticRegression model

Contributions:



Members Task	Tharun R	Raman KB	Pavan kumar B	Kamal R
Data Imputation and Scaling	■			■
Data Cleaning			■	
Exploratory Analysis		■		■
Feature Selection	■		■	■
Building Model	■		■	■
Result analysis and Accuracy Test		■		
Documentation	■	■	■	■

Table 4: Work Division

Conclusion:

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes inhigh risk patients and in turn reduce the complications, which can be a great milestone in the fieldof medicine. This project resolved the feature selection i.e. backward elimination and RFECV behind the models and successfully predict the heart disease, with 85% accuracy. The model usedwas Logistic Regression. Further for its enhancement, we can train on models and predict thetypes of cardiovascular diseases providing recommendations to the users, and also use moreenhanced models