IST707 – Applied Machine Learning

Research Project Title: Heart Disease Prediction

Pranali Shenvi

pshenvi@syr.edu

Kirat Saran

ksaran@syr.edu

Abstract

According to the Centers for Disease Control and Prevention, heart disease is one of the major causes of death for people of all races in the United States (African Americans, American Indians and Alaska Natives, and white people). Approximately half of all Americans (47%) have at least one of three important risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other important indicators include diabetes status, obesity (high BMI), a lack of physical exercise, and excessive alcohol consumption. It is critical in healthcare to identify and avoid the variables that have the greatest influence on heart disease. Computational advances, in turn, enable the use of machine learning approaches to find patterns in data that might forecast a patient's condition. Understanding the fundamental components that are the major causes of a heart attack allows us to bring this to the people, hence lowering the risk of heart disorders.

Objective

Our objective is to present a method for determining the degree of key variables that contribute to heart disease prediction. Our study's goal is to predict heart disease based on major feature ratings. This will in turn help to create a correct prediction of heart disease which can save a person's life. In this study, we plan to use exploratory analysis to summarize the dataset and understand the different attributes in the dataset. We are planning to predict if a person will have a heart disease or not based on the key indicators. We have used several machine learning algorithms like K-Nearest Neighbors, Decision Tree, XGBoost Classifier, Random Forest Classifier and a few more for the purpose of modelling.

Data Description

The dataset for this study is taken from Kaggle. The dataset originally is from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS). This is the most recent dataset (as of February 2022) which includes data from 2020. It consists of 401,958 rows and 279 columns. The great majority of columns are questions regarding respondents' health, such as "Do you have major difficulties walking or climbing stairs?" or "Have you ever smoked at least 100 cigarettes in your life?" The original dataset of nearly 300 variables was reduced to about 18 attributes and 319,795 transactions, which is the dataset we are using for the purpose of our study.

Data Preparation

The first step for examining the dataset is to carry out exploratory data analysis. For that, we have first checked the dataset for null values and duplicates and checked the datatype of attributes. We have observed that there are no null values in the dataset and there are 4 attributes which are numerical, and all others are categorical. We have then checked the dataset for duplicate values, and we found that there are 18,078 records which are duplicates. But after analyzing the dataset we went ahead without removing the duplicates as we are assuming that there can be people with the exact same statistics. Our target variable is the column HeartDisease. In the dataset, we have seen that the percentage of people having a heart disease is about 8.6%. This means that the dataset is imbalanced, and we will have to deal with the bias using some other techniques.

Exploratory Data Analysis

We have carried out the exploratory data analysis, which helps us determine the key factors affecting heart disease and understand the dataset in depth. We have made graphs for all the categorical and numerical attributes. Here, we are explaining about some of the graphs that we found intuitive for categorical variables.

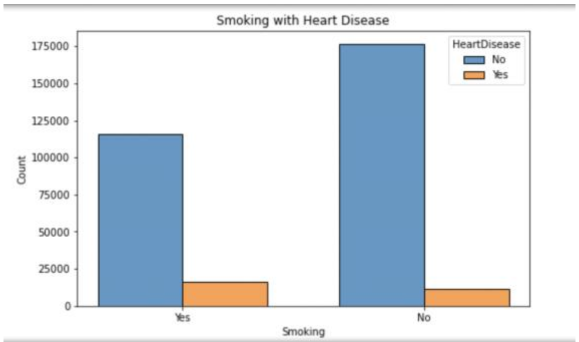


Figure 1: People having
heart disease in the smoking category

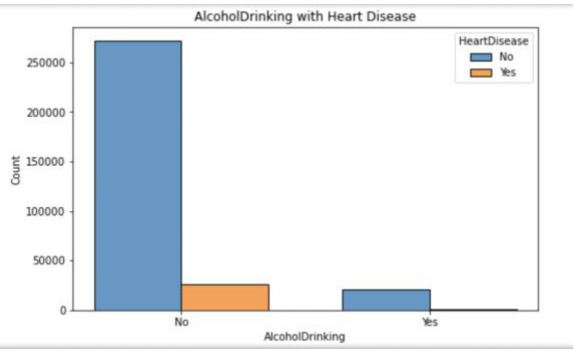
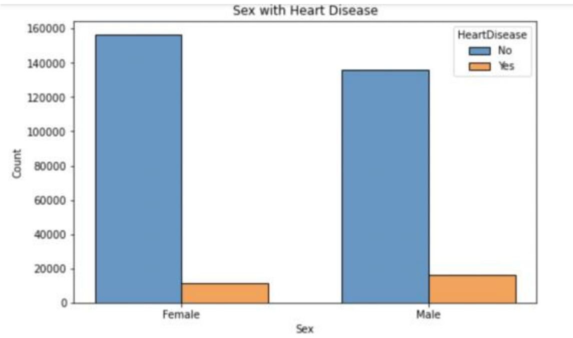


Figure 2: People having heart disease in the
alcohol drinking category



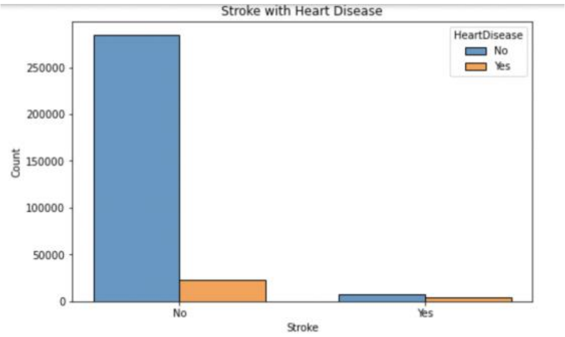


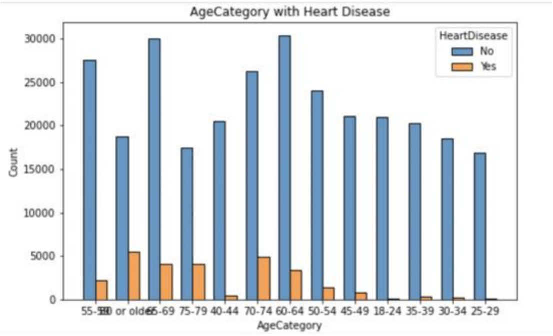
Figure 3: People having

heart disease in the gender category

Figure 4: People having heart disease in the

alcohol drinking category

From these graphs we can say the following: The proportion of people having heart disease is proportionally more in people who smoke, consume alcohol, and have had a stroke in the past. We can also see that the number of heart attacks in females is much more compared to males.



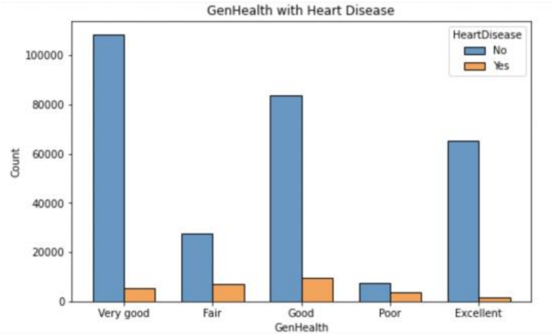


Figure 5: People having

heart disease in different age categories

Figure 6: People having heart disease in the

different general health categories

We can also observe that the age group of 60 and older and 70-74 have the highest proportion of heart diseases and heart diseases have occurred proportionally more in people having poor general health.

Similarly, we have plotted density graphs for numerical variables.

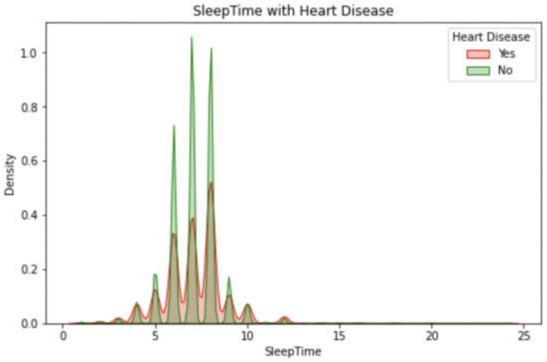


Figure 7: SleepTime density comparison

for our target variable HeartDisease

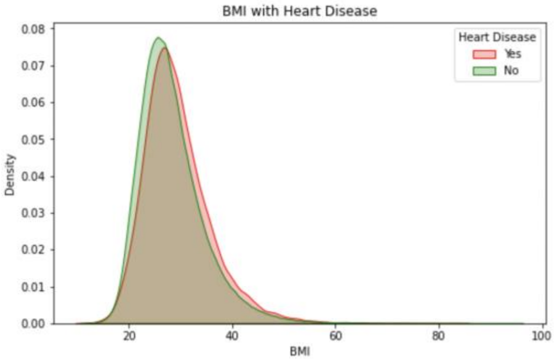


Figure 8: BMI density comparison

for our target variable HeartDisease

From these graphs we can observe that, people who sleep less than the average number of hours of sleep have chances of developing or having a heart disease. People having higher BMI are prone to having a heart disease.

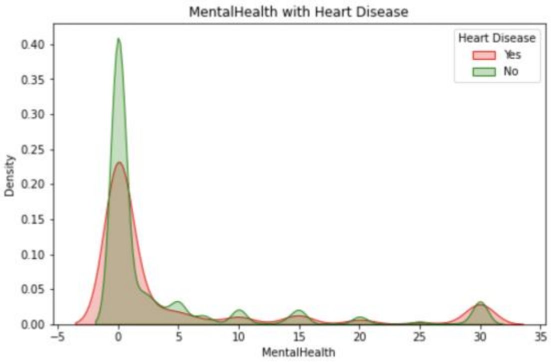


Figure 9: MentalHealth density comparison

for our target variable HeartDisease

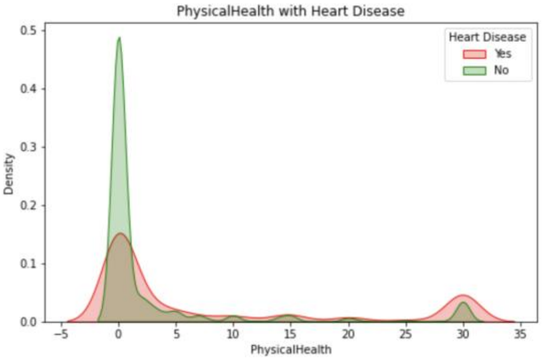


Figure 10: PhysicalHealth density comparison

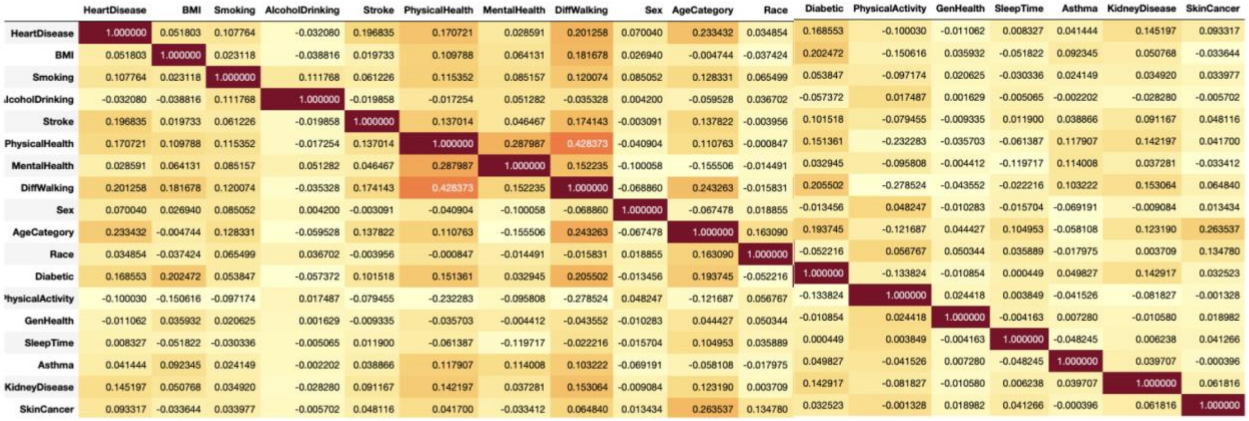
for our target variable HeartDisease

From these graphs we can observe that, people with physical and mental health problems show similar characteristics. Greater the number of days a person has these problems, more is the possibility of them developing or having a heart disease.

We have also observed that the greater number of heart diseases can be seen in American Indian/ Alaskan Native followed by the white race in the dataset.

		Heart Disease		Percentage
Race		No	Yes	
	American Indian/Alaskan Native	4660	542	10.42%
	Asian	7802	266	3.30%
	Black	21210	1729	7.54%
	Hispanic	26003	1443	5.26%
	Other	10042	886	8.11%
	White	222705	22507	9.18%

Table 1: Percentage of heart disease in different races



	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
HeartDisease	1.000000	0.051803	0.107764	-0.032080	0.196835	0.170721	0.028591	0.201258	0.070040	0.233432	0.034854	0.168553	-0.100030	-0.011062	0.008327	0.041444	0.145197	0.093317
BMI	0.051803	1.000000	0.023118	-0.038816	0.019733	0.109788	0.064131	0.181678	0.026940	-0.004744	-0.037424	0.202472	-0.150616	0.035932	-0.051822	0.092345	0.050768	-0.033644
Smoking	0.107764	0.023118	1.000000	0.111768	0.061226	0.115352	0.085157	0.120074	0.085052	0.128331	0.065499	0.053847	-0.097174	0.020625	-0.030336	0.024149	0.034920	0.033977
lcoholDrinking	-0.032080	-0.038816	0.111768	1.000000	-0.019858	-0.017254	0.051282	-0.035328	0.004200	-0.059528	0.036702	-0.057372	0.017487	0.001629	-0.005065	-0.002202	-0.028280	-0.005702
Stroke	0.196835	0.019733	0.061226	-0.019858	1.000000	0.137014	0.046467	0.174143	-0.003091	0.137822	-0.003956	0.101518	-0.079455	-0.009335	0.011900	0.038866	0.091167	0.048116
PhysicalHealth	0.170721	0.109788	0.115352	-0.017254	0.137014	1.000000	0.287987	0.428373	-0.040904	0.110763	-0.000847	0.151361	-0.232283	-0.035703	-0.061387	0.117907	0.142197	0.041700
MentalHealth	0.028591	0.064131	0.085157	0.051282	0.046467	0.287987	1.000000	0.152235	-0.100058	-0.155506	-0.014491	0.032945	-0.095808	-0.004412	-0.119717	0.114008	0.037281	-0.033412
DiffWalking	0.201258	0.181678	0.120074	-0.035328	0.174143	0.428373	0.152235	1.000000	-0.068860	0.243263	-0.015831	0.205502	-0.278524	-0.043552	-0.022216	0.103222	0.153064	0.064840
Sex	0.070040	0.026940	0.085052	0.004200	-0.003091	-0.040904	-0.100058	-0.068860	1.000000	-0.067478	0.018855	-0.013456	0.048247	-0.010283	-0.015704	-0.069191	-0.009084	0.013434
AgeCategory	0.233432	-0.004744	0.128331	-0.059528	0.137822	0.110763	-0.155506	0.243263	-0.067478	1.000000	0.163090	0.193745	-0.121687	0.044427	0.104953	-0.058108	0.123190	0.263537
Race	0.034854	-0.037424	0.065499	0.036702	-0.003956	-0.000847	-0.014491	-0.015831	0.018855	0.163090	1.000000	-0.052216	0.056767	0.050344	0.035889	-0.017975	0.003709	0.134780
Diabetic	0.168553	0.202472	0.053847	-0.057372	0.101518	0.151361	0.032945	0.205502	-0.013456	0.193745	-0.052216	1.000000	-0.133824	-0.010854	0.000449	0.049827	0.142917	0.032523
hysicalActivity	-0.100030	-0.150616	-0.097174	0.017487	-0.079455	-0.232283	-0.095808	-0.278524	0.048247	-0.121687	0.056767	-0.133824	1.000000	0.024418	0.003849	-0.041526	-0.081827	-0.001328
GenHealth	-0.011062	0.035932	0.020625	0.001629	-0.009335	-0.035703	-0.004412	-0.043552	-0.010283	0.044427	0.050344	-0.010854	0.024418	1.000000	-0.004163	0.007280	-0.010580	0.018982
SleepTime	0.008327	-0.051822	-0.030336	-0.005065	0.011900	-0.061387	-0.119717	-0.022216	-0.015704	0.104953	0.035889	0.000449	0.003849	-0.004163	1.000000	-0.048245	0.006238	0.041266
Asthma	0.041444	0.092345	0.024149	-0.002202	0.038866	0.117907	0.114008	0.103222	-0.069191	-0.058108	-0.017975	0.049827	-0.041526	0.007280	-0.048245	1.000000	0.039707	-0.000396
KidneyDisease	0.145197	0.050768	0.034920	-0.028280	0.091167	0.142197	0.037281	0.153064	-0.009084	0.123190	0.003709	0.142917	-0.081827	-0.010580	0.006238	0.039707	1.000000	0.061816
SkinCancer	0.093317	-0.033644	0.033977	-0.005702	0.048116	0.041700	-0.033412	0.064840	0.013434	0.263537	0.134780	0.032523	-0.001328	0.018982	0.041266	-0.000396	0.061816	1.000000

Figure 11: Correlation heat map

From the correlation heat map, we can observe that closer the correlation value is to 1, stronger is the correlation between the attributes. Similarly closer the value to 0, weaker is the correlation. Hence, we see that, as the age of a person increases, the probability of getting the heart disease also increases. People with better physical health are less prone to getting a heart disease. People having a kidney disease, stroke and diabetes are more prone to getting a heart disease. The more number of days a person has difficulty walking, the more are the chances of getting a heart disease.

Data Pre-Processing

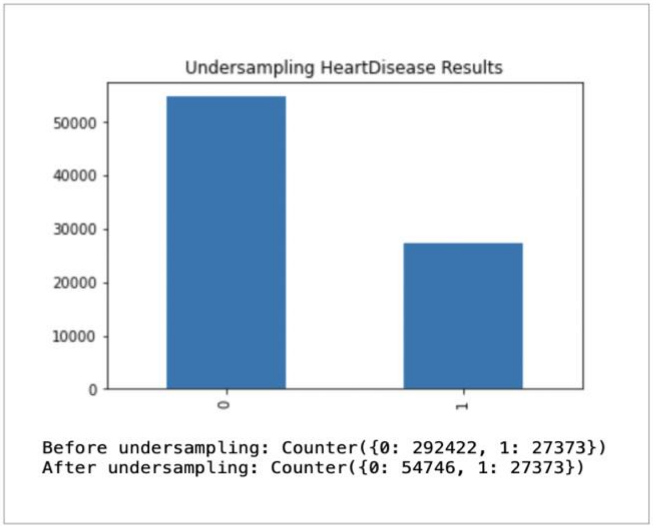
All categorical variables are encoded using Label Encoder into numerical values so that they can be in machine-readable form. Machine learning algorithms make better predictions in the label encoded form.

We have then split the dataset into train and test sets with a train size of 70 percent and a test size of 30 percent.

As mentioned previously, we had to deal with the problem of bias in the dataset, as the percentage of the people having a heart disease is comparatively much lower to the percentage of people not having a heart disease, we performed sampling techniques to overcome the bias.

Sampling

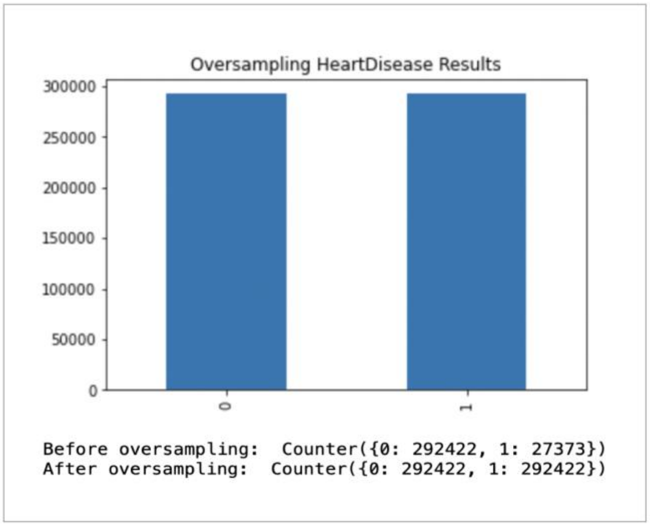
We have used 3 sampling techniques on all our models namely Undersampling, Oversampling and SMOTE Oversampling. We have compared our sampled datasets results with each other along with our original imbalanced dataset. We have then used evaluation metrics to compare these models to find our best fit model.



```
Before undersampling: Counter({0: 292422, 1: 27373})
After undersampling: Counter({0: 54746, 1: 27373})
```

Figure 12: Majority, minority class row count after Undersampling

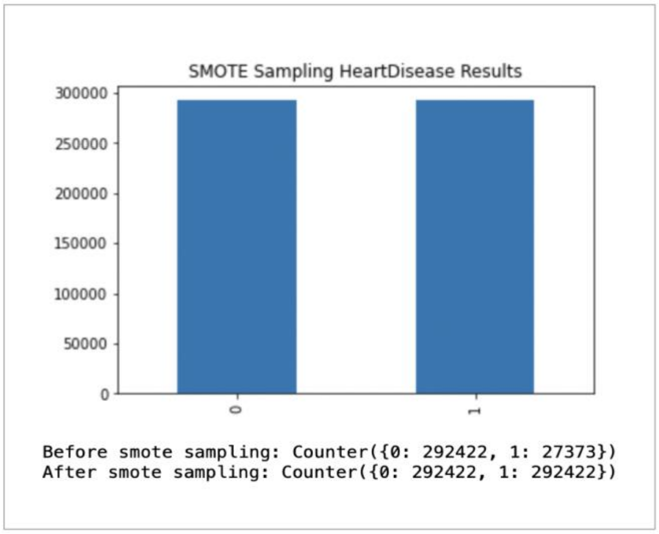
For Undersampling, we reduced the size of our majority class such as it is twice the size of our minority class.



```
Before oversampling:  Counter({0: 292422, 1: 27373})
After oversampling:  Counter({0: 292422, 1: 292422})
```

Figure 13: Majority, minority class row count after Oversampling

For Oversampling, we increased the size of our minority class to match the size of our majority class.



```
Before smote sampling: Counter({0: 292422, 1: 27373})
After smote sampling: Counter({0: 292422, 1: 292422})
```

Figure 14: Majority, minority class row count after SMOTE Oversampling

For SMOTE oversampling, just like oversampling, we increased the size of our minority class to match the size of our majority class.

Models Used

After the exploratory data analysis and the pre-processing, we have performed several modelling techniques like K-Nearest Neighbors, Decision Tree, Naïve Bayes, Random Forest, XGBoost Classifier and Logistic Regression.

Sampling	Accuracy	Precision	Recall
Unsampled	0.90	0.33	0.88
Undersampled	0.82	0.28	0.66
Oversampled	0.88	0.42	0.99
Oversampled (SMOTE)	0.83	0.33	0.97

Table 2: K-Nearest Neighbors performance metrics

Sampling	Accuracy	Precision	Recall
Unsampled	0.86	0.22	0.25
Undersampled	0.82	0.33	0.99
Oversampled	0.99	0.93	0.99
Oversampled (SMOTE)	0.99	0.98	0.96

Table 3: Decision Tree performance metrics

Sampling	Accuracy	Precision	Recall
Unsampled	0.84	0.27	0.46
Undersampled	0.81	0.24	0.54
Oversampled	0.79	0.23	0.59
Oversampled (SMOTE)	0.71	0.19	0.71

Table 4: Naïve Bayes performance metrics

Sampling	Accuracy	Precision	Recall
Unsampled	0.91	0.50	0.09
Undersampled	0.84	0.29	0.54
Oversampled	0.74	0.21	0.76
Oversampled (SMOTE)	0.70	0.18	0.71

Table 5: Logistic Regression performance metrics

Sampling	Accuracy	Precision	Recall
Unsampled	0.90	0.36	0.11
Undersampled	0.87	0.40	0.99
Oversampled	0.99	0.93	0.99
Oversampled (SMOTE)	0.99	0.96	0.98

Table 6: Random Forest performance metrics

Sampling	Accuracy	Precision	Recall
Unsampled	0.91	0.52	0.09
Undersampled	0.84	0.30	0.66
Oversampled	0.75	0.23	0.83
Oversampled (SMOTE)	0.84	0.26	0.46

Table 7: XGBoost performance metrics

After running all the above models, we get the different values for accuracy, recall and precision. We are not entirely relying on accuracy as there is bias in the dataset. We are considering Recall as our main evaluation metric. Comparing all the models' performance, we can say that Tree based models are performing better for this dataset. That is why our best models out of all the models we ran are Decision Tree and Random Forest.

Hyperparameter Tuning

To improve the performance of our best models, we will fine tune them on all the types of sampled datasets.

Conclusion

According to the analysis we have conducted, we can conclude that:

We can identify key indicators which can influence heart disease. Our exploratory analysis gives us some of the attributes which can be counted as possible factors. The correlation heat map helped us to find the relation between the attributes affecting the target variable. We were able to find the attributes which cause both positive and negative effect on HeartDisease attribute. We have observed that our dataset has bias and to overcome that, we have used sampling techniques. Then we have performed the modelling techniques and we can say that Random Forest and Decision Tree gives the best output. After Hyperparameter tuning on one of our best models – Decision Tree for all imbalanced, undersampled, oversampled and SMOTE oversampled data, the following 2 decision models are the best:

1. SMOTE Oversampled Decision Tree Model with a recall of 0.9461
2. Oversampled Decision Tree Model with a recall of 0.9272

References

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

https://www.hindawi.com/journals/cin/2021/8387680/

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01527-5