Course: IST 707 Applied Machine Learning
Title: Heart Disease Prediction

Project Description:

Objective of the project is to understand the fundamental components that are the major causes of a heart attack which could help lower the risk of heart disorders. The goal is to use machine learning to predict if a person would have heart disease based on dataset's major key performance indicators (KPIs) like smoking, alcohol consumption, sex, stroke, age category, sleep time, BMI, mental health, physical health, etc.

Dataset Description:

The dataset for this study was taken from Kaggle in the form of a csv file. The dataset originally was from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS). This was the most recent dataset (as of February 2022) at the time which included data from 2020. It consisted of 401,958 rows and 279 columns. The dataset was reduced to about 18 attributes and 319,795 transactions.

Data Exploration:

We observed that there were no null values in the dataset and there were 4 attributes which were numerical, and all others were categorical. We then checked the dataset for duplicate values, and we found that there are 18,078 records which were duplicates. But after analyzing the dataset we went ahead without removing the duplicates as we assumed that there can be people with the exact same statistics. Our target variable was the column HeartDisease. In the dataset, we saw that the percentage of people having a heart disease was about 8.6%. This implied that the dataset was imbalanced and planned on dealing with the bias using some sampling techniques.

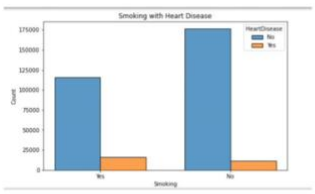

Figure 1: People having heart disease in the smoking category

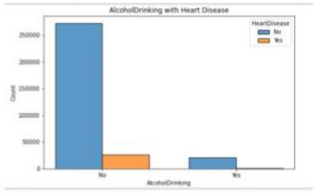

Figure 2: People having heart disease in the alcohol drinking category

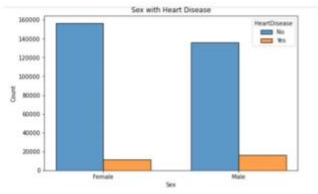

Figure 3: People having heart disease in the gender category

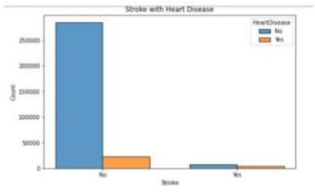

Figure 4: People having heart disease in the alcohol drinking category

We observed that, the proportion of people having heart disease is proportionally more in people who smoke, consume alcohol, and have had a stroke in the past. Also that the number of heart attacks in females was much more compared to males.

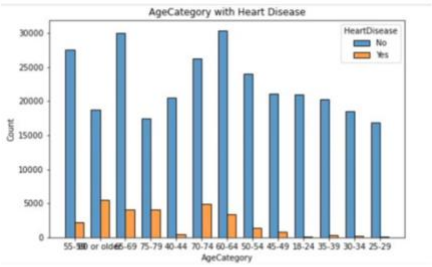

Figure 5: People having heart disease in different age categories

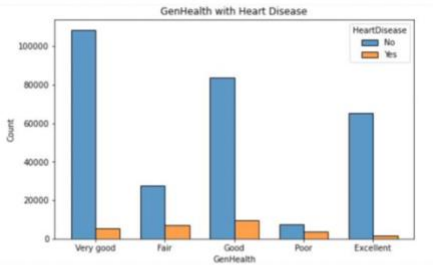

Figure 6: People having heart disease in the different general health categories

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

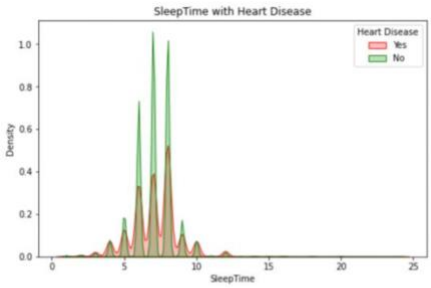

Figure 7: SleepTime density comparison for our target variable HeartDisease

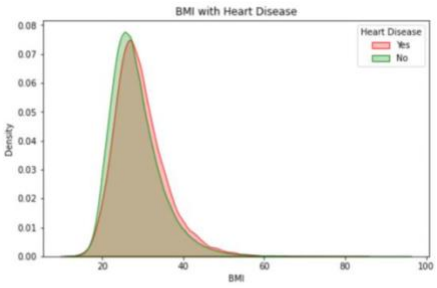

Figure 8: BMI density comparison for our target variable HeartDisease

Here we observed that people who slept less than the average number of hours of sleep had chances of developing or having a heart disease. People having higher BMI were prone to having a heart disease.

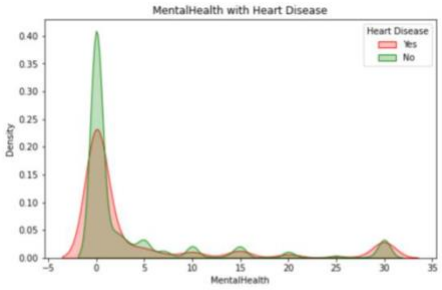

Figure 9: MentalHealth density comparison for our target variable HeartDisease

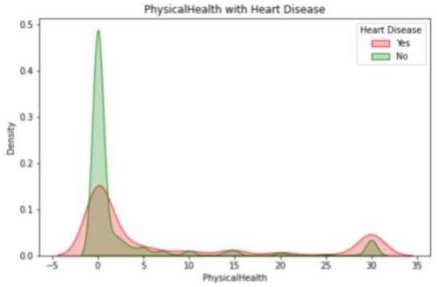

Figure 10: PhysicalHealth density comparison for our target variable HeartDisease

Here we observed that, people with physical and mental health problems showed similar characteristics. Greater the number of days a person had these problems, more was the possibility of them developing or having a heart disease.

		Heart Disease		Percentage
		No	Yes	
Race	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%

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

	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
AlcoholDrinking	-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
PhysicalActivity	-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

Here we observed that as the age of a person increased, the probability of getting the heart disease also increased. People with better physical health were less prone to getting a heart disease. People having a kidney disease, stroke and diabetes were more prone to getting a heart disease. More number of days a person had difficulty walking, more were the chances of getting a heart disease.

Data Modelling:

Data Pre-Processing:

All categorical variables were encoded using Label Encoder into numerical values so that they can be in machine-readable form. We 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 earlier, we used 3 sampling techniques to overcome bias: Undersampling, Oversampling and SMOTE Oversampling.

- For Undersampling, we reduced the size of our majority class such as it is twice the size of our minority class.
- For Oversampling, we increased the size of our minority class to match the size of our majority class.
- For SMOTE oversampling, just like oversampling, we increased the size of our minority class to match the size of our majority class.

Models used:

We used 6 algorithms to build our models: K-Nearest Neighbors, Decision Tree, Naïve Bayes, Random Forest, XGBoost Classifier and Logistic Regression

We then used 3 evaluation metrics namely accuracy, precision and recall to compare these models to find our best fit model.

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

We did not entirely rely on accuracy as there was bias in the dataset. We considered Recall as our main evaluation metric. Comparing all models' performance, we deduced that Tree based models performed better for this data. Our best models out of all the models we ran were Decision Tree and Random Forest.

Insights:

- Exploratory analysis gave us some of the attributes which were used as factors.
- Correlation heat map helped us find the relation between the attributes affecting the target variable.
- To overcome bias, we had used sampling techniques.
- Random Forest and Decision Tree gave us the best results.

- After Hyperparameter tuning our best model – Decision Tree for all types of sampled data: undersampled, oversampled and SMOTE oversampled data, the following 2 decision models were 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

Specific contribution to the project:

- Cleaning data and recognizing useful features for our analysis.
- Sampling
- Machine learning modelling in the project.

Learning outcomes from the project:

- Better understanding of sampling techniques and how to implement them.
- Use of various models in our project and finalizing our best model with the use of evaluation metrics.
- Relevant hyper-parameter tuning of our best fit model to get more accurate results.