We declare that we have completed this assignment completely and entirely on our own, without any consultation with others. We have read the UAB Academic Honor Code and understand that any breach of the Honor Code may result in severe penalties.

We also declare that the following percentage distribution ***faithfully*** represents individual group members' contributions to the completion of the assignment

Name	Overall Contribution (%)	Major work items completed by me	Signature or initials	Date
Seshagiri rao Mallena	20	Data preprocessing SVM Implementation and tuning	Giri M	20-April-2022
Hrishikesh Vikram	20	Naive Bayes Implementation and paper	HV	20-April-2022
Jyothi Prakash Rasineni	20	Decision Tree Implementation and data preprocessing	R Jyothi Prakash	20-April-2022
Sainadh Reddy Sandi	20	Logistic Regression and data preprocessing	Sainadh	20-April-2022
Kiran teja Ruthala	20	Random Forest Techniques and PPT	Kiran R	20-April-2022

Prediction of Heart Disease Using Classification Techniques

ABSTRACT

The main cause of death in the United States, according to the CDC (Center for Disease Control), is heart disease. Finding out if you have a cardiac ailment can be difficult in and of itself because certain diseases are notorious for remaining silent. BMI, alcohol use, age, inherited characteristics, smoking, and a variety of other factors all have a role in heart disease.

The major purpose of this work was to employ classification algorithms and supervised learning models to find trends and maybe forecast the likelihood of developing heart disease in people with certain health problems [4].

1 Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is a public health phone survey that gathers data on public health issues such as behavior, long-term medical disorders, medication usage, and preventative treatment. The data, which is gathered from all fifty states in the United States, can be seen on the CDC website[4].

2. Data

While the CDC dataset is large, it was not the best fit for our investigation. The primary dataset for this study was a preprocessed version of the same dataset that is accessible on Kaggle.

The features of the dataset are listed below along with what was asked from the participants for the following feature.[1]

Table 1: Features of Heart Disease Dataset

Feature	Explanation
HeartDisease	Respondents that have ever reported having coronary heart disease
BMI	Body mass Index
Smoking	Have you smoked at least 100 cigarettes in your entire life?
Alcohol Drinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
Stroke	(Ever told) (you had) a stroke?
PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30
MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good?
DiffWalking	Do you have serious difficulty walking or climbing stairs?
Sex	Male or Female?
AgeCategory	Fourteen-level age category
Race	Ethnicity
Diabetic	Are you diabetic?
PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
GenHealth	Opinion on personal General health
SleepTime	How many hours on

	average sleep time you get in 24 hour period?
Asthma	Do you have Asthma?
KidneyDisease	Were you ever told you had kidney disease? Not including kidney stones, bladder infection or incontinence
SkinCancer	Do you have skin cancer?

3 Data Preprocessing

Pandas dataframe was used to load the input data, which has 18 features and over 300,000 entries. We used pandas dataframe methods like info and describe to do preliminary statistics on the data. After that, we used exploratory data analysis to learn more about the dataset. There are 18 mixed-type features in the entire dataset.

To cope with nominal data, we employed encoding techniques from the sklearn package and substituted boolean Yes and no with 1 and 0 values. For numeric data, we also employed normalization. HeartDisease is the predictor variable that we must pick. We are trying to predict whether the person has Heart Disease or not. When we try to figure out how many records show the individual is suffering from heart illness out of all the records, we find that just 8% of the total records reveal the person is suffering from heart disease. So, before fitting the model, we wanted to balance the dataset, so we utilized the imblearn package and tried to oversample the data to make it balanced. We also tried to undersample to check the run time of some algorithms like Support vector Machine to determine the accuracies.

Figure 1 : Preprocess Steps

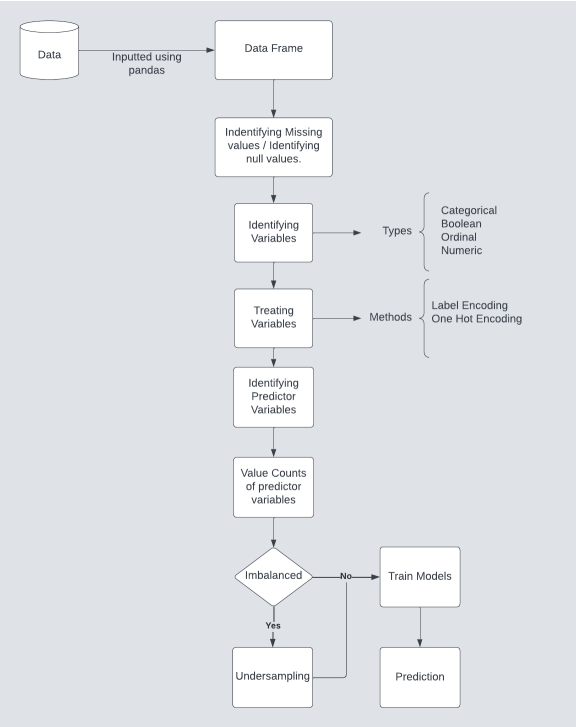


Figure 2 Plot of the Correlation Table

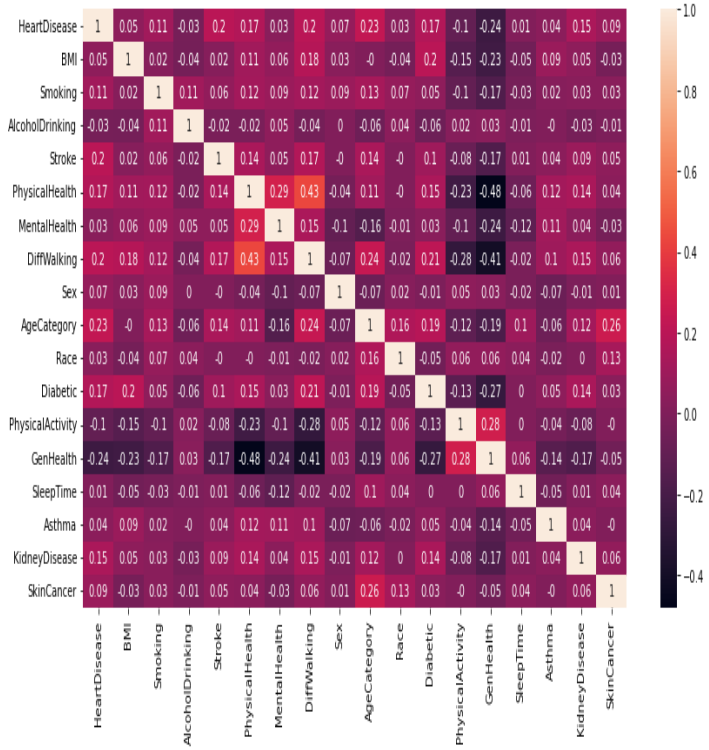


Figure 3 People with smoking as a habit and effected by heart disease:

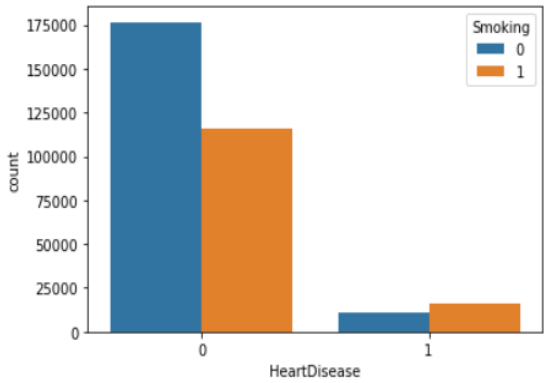


Figure 4 People with Alcohol Drinking as a habit and affected with Heart Disease:

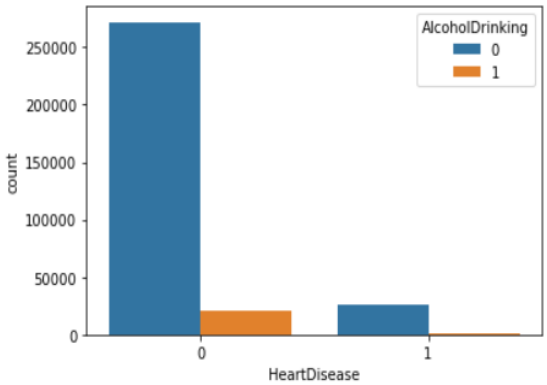


Figure 5 People with Heart Disease basing on Race:

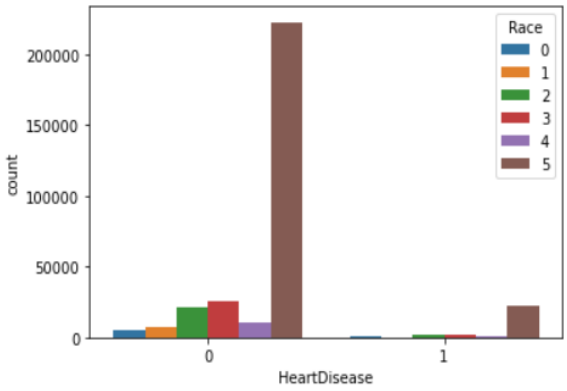


Figure 6 People with Diabetics and affected by Heart Disease:

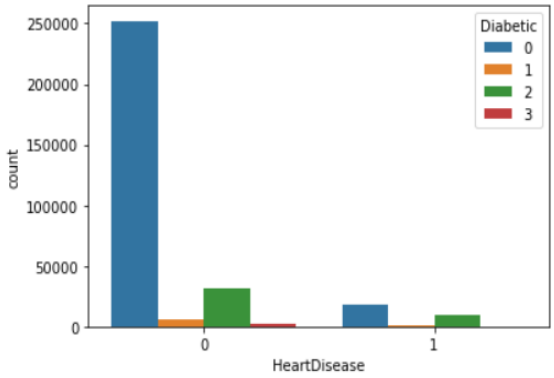


Figure 7 People with Heart Disease basing on their General Health:

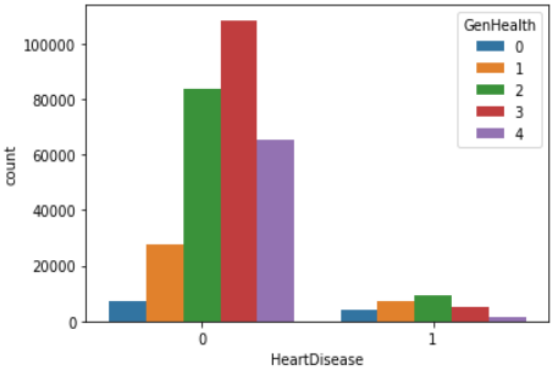


Figure 8 Frequency:

[BMI]

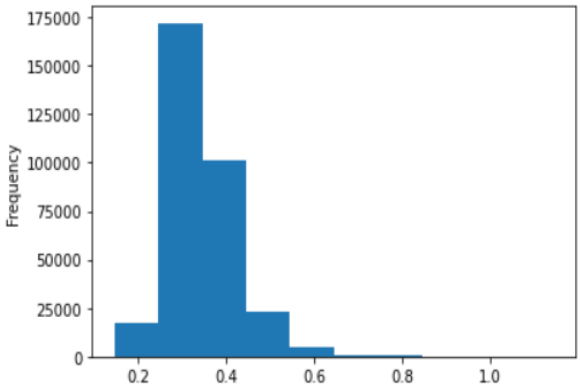


Figure 9 MentalHealth plot

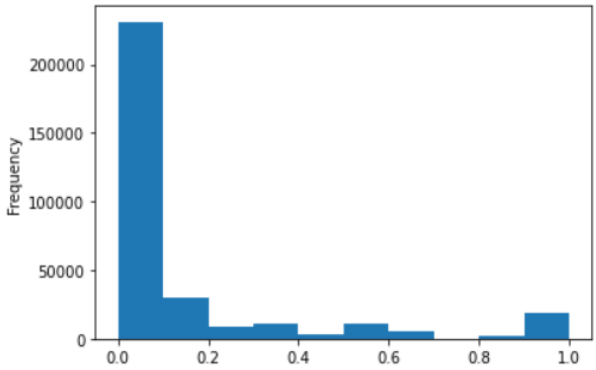
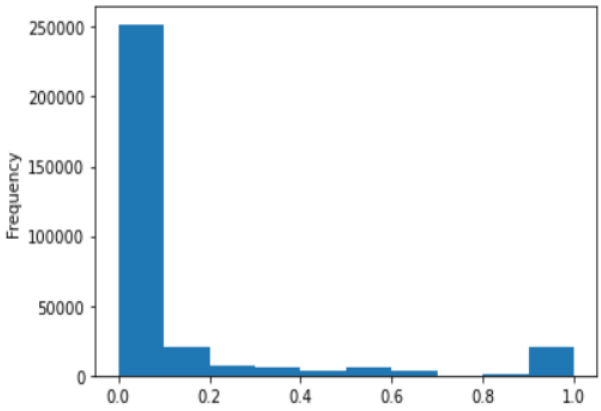


Figure 10 Physical health plot



4 Predictor Model and Results:

To train the prediction model, we compared and contrasted several different classification methods. These methods include Logistic Regression, Decision trees, Support Vector Machine Gaussian Naïve Bayes, Random Forest and also used bagging classifiers.

The prediction models were first run on the original dataset without any additions but failed to return strong results. This is mostly because of the imbalance present in the dataset. For example, the Support Vector Machine algorithm returned an accuracy of 90% yet the precision and recall values for the Yes category were zero. Another instance being the f1 measures also being zero, while the the records for the No category is 95%. .

Figure 11 Heart Disease data

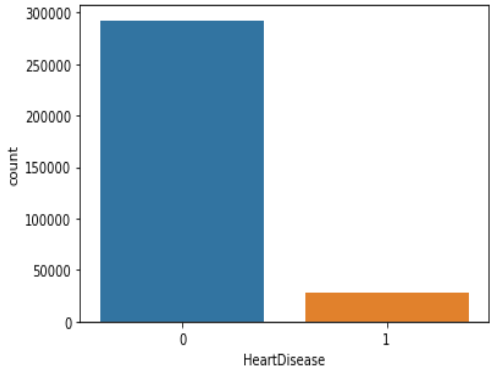


Table 2: The Accuracy, Precision and Recall of the original dataset

Model	Accuracy	Precision for No (0)	Precision for Yes (1)	Recall for No (0)	Recall for Yes(1)
Logistic regression	0.91	0.92	0.52	0.99	0.1
Decision tree	0.91	0.92	0.55	0.99	0.06
Support Vector Machine	0.92	0.92	0.64	1	0.02
Naive Bayes	0.84	0.95	0.26	0.88	0.47
Random Forest	0.91	0.92	0.36	0.98	0.12

When we observed the precision and recall values of the original dataset the results don't look promising so we decided to run the classification algorithms on modified dataset and yield good precision and recall scores.

Table 3:

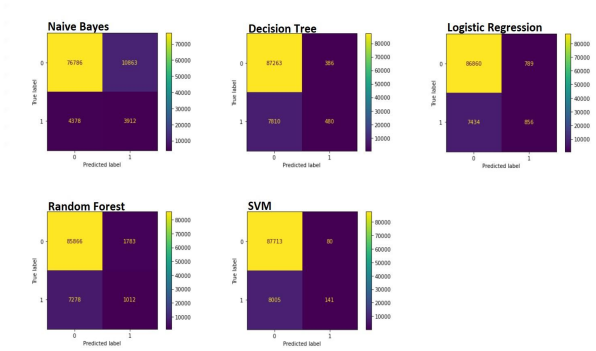
The Accuracy, Precision and Recall of the Modified dataset:

Model	Accuracy	Precision No	Precision Yes	Recall No	Recall Yes
Logistic regression	0.76	0.77	0.76	0.75	0.78
Decision tree	0.75	0.76	0.74	0.73	0.77
Support Vector Machine	0.77	0.79	0.76	0.75	0.80
Naive Bayes	0.71	0.67	0.77	0.82	0.60
Random Forest	0.96	1	0.93	0.93	1

We also generated the confusion matrix to identify True Positive and True Negative values.

Figure 12 Confusion matrix

[a]



The four cells from top to bottom, left to right are true negatives, false negatives, false positives and true positives.

Figure 13 Confusion Matrix Results of Processed data set

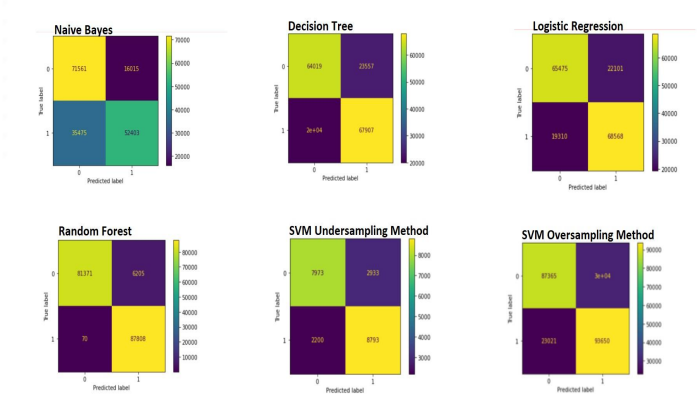


Figure 14 Bagging Classifier results of Random Forest

	precision	recall	f1-score	support
0	1.00	0.98	0.99	87576
1	0.98	1.00	0.99	87878
accuracy			0.99	175454
macro avg	0.99	0.99	0.99	175454
weighted avg	0.99	0.99	0.99	175454

Figure 15 Hyperparameter Tuning for SVM

```
print(grid.best_params_)


print(grid.best_estimator_)
```

```
{'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}
SVC(C=10, gamma=0.01)
```

5 Discussions and Conclusions

After evaluating the results generated from running the algorithms, we are confident about the accuracy of the models. Attempts to improve the accuracy such as hyperparameter tuning using GridSearchCV for support vector machine, bagging classifiers for random forest were used. Based on the results, we can believe that these models can be used to predict whether

a person could suffer from heart disease or not, considering the input data is reliable.

6 Future Work:

Further preprocessing can be done using various techniques apart from the ones used in this study. The models can be trained using deep learning methods [3] to further improve the classification accuracy of the results. Principal component analysis can be employed on the original dataset from CDC [4] to train the models to determine and compare the accuracy.

Acknowledgements:
We drew inspiration from a project on Kaggle called Personal Key Indicators for Heart Disease by Kamil Pytlak [1]. Kamil did a Exploratory analysis of the CDC 2020 data. We believed we could relate and extend this work and build sophisticated models to predict the risk of heart disease.

References:

[1] Pytlak, K. (2022, February 15). *Personal key indicators of heart disease*. Kaggle. Retrieved April 20, 2022, from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[2] https://scikit-learn.org/stable/supervised_learning.html

[3] Goodfellow, Ian, et al. *Deep Learning*. The MIT Press, 2017.

[4] "CDC - 2020 BRFSS Survey Data and Documentation." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 27 Aug. 2021, https://www.cdc.gov/brfss/annual_data/annual_2020.html.