Contents

ABSTRACT

Heart disease is a group of conditions that affect the structure and function of the heart. It is the leading cause of death worldwide, accounting for more than 17 million deaths annually. While heart disease can manifest in different ways, several personal vital indicators can help identify individuals at a higher risk of developing the condition. These indicators include high blood pressure, high cholesterol, smoking, diabetes, obesity, family history of heart disease, physical inactivity, and chronic stress.

Knowing and understanding these personal key indicators of heart disease is crucial to take appropriate steps to prevent or manage the condition. Lifestyle changes such as exercising regularly, maintaining a healthy diet, quitting smoking, and managing stress can help reduce the risk of heart disease. Additionally, regular check-ups with a healthcare provider can help identify any potential heart health issues early on. By being aware of the personal key indicators of heart disease and taking appropriate preventative measures, individuals can reduce their risk of developing heart disease and improve their overall heart health.

Classification is a type of machine learning technique that can be used to identify personal key indicators of heart disease. Classification is a technique to predict a categorical output based on input features. In the context of heart disease, classification can be used to predict whether an individual is at risk of developing heart disease based on personal key indicators such as high blood pressure, high cholesterol, smoking, or obesity.

CHAPTER 1- INTRODUCTION

Heart disease is a condition that affects many people, and while some individuals may experience symptoms, others may not. Symptoms of heart disease can include chest pain, difficulty breathing, palpitations, swelling in the legs or feet, weakness, and cyanosis. There are many risk factors for heart disease, including age, sex, tobacco use, physical inactivity, excessive alcohol consumption, unhealthy diet, obesity, genetic predisposition, high blood pressure, high blood sugar, high blood cholesterol, undiagnosed celiac disease, psychosocial factors, poverty, low education status, and air pollution.

While some risk factors, such as age, sex, and family history, cannot be changed, many other risk factors can be modified through lifestyle changes, social changes, and medication. For example, individuals who are obese are at an increased risk of atherosclerosis of the coronary arteries. Age is the most significant risk factor for developing cardiovascular diseases, and coronary fatty streaks can begin to form in childhood.

Diagnosing heart disease is a critical and challenging task in healthcare, requiring many tests and careful examination of results. To improve diagnosis and prevention, researchers have developed various heart disease prediction systems using different AI algorithms. These prediction systems can help healthcare professionals make more accurate diagnoses and prevent heart disease, improving overall health and well-being.

AI can help predict heart disease by analyzing patient datasets with appropriate data processing and training various models using algorithms such as KNN, SVM, Decision Tree, Naive Bayes, Random Forest, and Logistic Regression. These models can provide accurate predictions, aiding in the prevention and management of heart disease.

Aim:

To generate a machine learning-based classification model that will help speed up the process for the health facilities to determine the potential heart disease patient based on the patient's activity pattern and past medical reports to provide immediate care.

Objectives:

- To explore the dataset for assessing the credibility of the dataset using advanced statistical methods.
- To determine the captured dataset's demographic characteristics, activity patterns, and generalized trends using plotting libraries.
- To understand all the factors responsible for the heart disease.
- To identify relevant factors contributing towards heart disease using Correlation Matrix.
- To train the machine learning models based on training and validation datasets for high-performance prediction models using various models.

Assumptions:

The models are generated based on the given dataset, and available attributes are not tested on other samples.

Limitations:

The authenticity is not commendable as it has been adopted as it is from Kaggle.

Methodology:

The goal of this research is to use computerized heart disease prediction to anticipate the likelihood of heart disease. This can be advantageous for both medical professionals and patients. Our study report showcases the outcomes achieved by using various machine learning algorithms on a dataset. To refine our methodology, we intend to clean the data, eliminate extraneous data and include additional features such as MAP and BMI. We will then divide the dataset by gender and employ k-modes clustering. Lastly, we will train the model with the processed data. Using this improved approach, we anticipate producing more accurate results and improved model performance.

CHAPTER 2-EXPLORATORY DATA ANALYSIS

Dataset Description:

The dataset was initially obtained from the CDC and is a crucial component of the Behavioral Risk Factor Surveillance System (BRFSS), which annually conducts telephone surveys to compile information about the health status of American inhabitants. The BRFSS, founded in 1984, has expanded to all 50 states, the District of Columbia, and three U.S. territories. It performs over 400,000 adult interviews annually, making it the most extensive and ongoing health survey program globally, as the CDC describes. The latest dataset, updated on February 15, 2022, encompasses data from 2020 and comprises 401,958 rows and 279 columns. Most columns consist of inquiries posed to participants regarding their well-being, such as "Do you experience difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes throughout your life? [Note: 5 packs = 100 cigarettes]."

After identifying various factors (questions) that impact heart disease, our group chose the most pertinent variables from the dataset and cleaned them to make them applicable for machine learning endeavors.

Data Balancing and Outlier Removal:

It is apparent from Figure 1 that the dataset contains outliers, which could be due to data entry errors. Removing these outliers has the potential to enhance the performance of our predictive model. To address this problem, we manually identified and supposedly removed all occurrences of BMI that fell outside the range of 2.5% to 97.5% as per technical parameters. Although considering it as a medical dataset, outliers can also be helpful to predict heart disease for a specific rare case that we might have missed because of outlier removals.

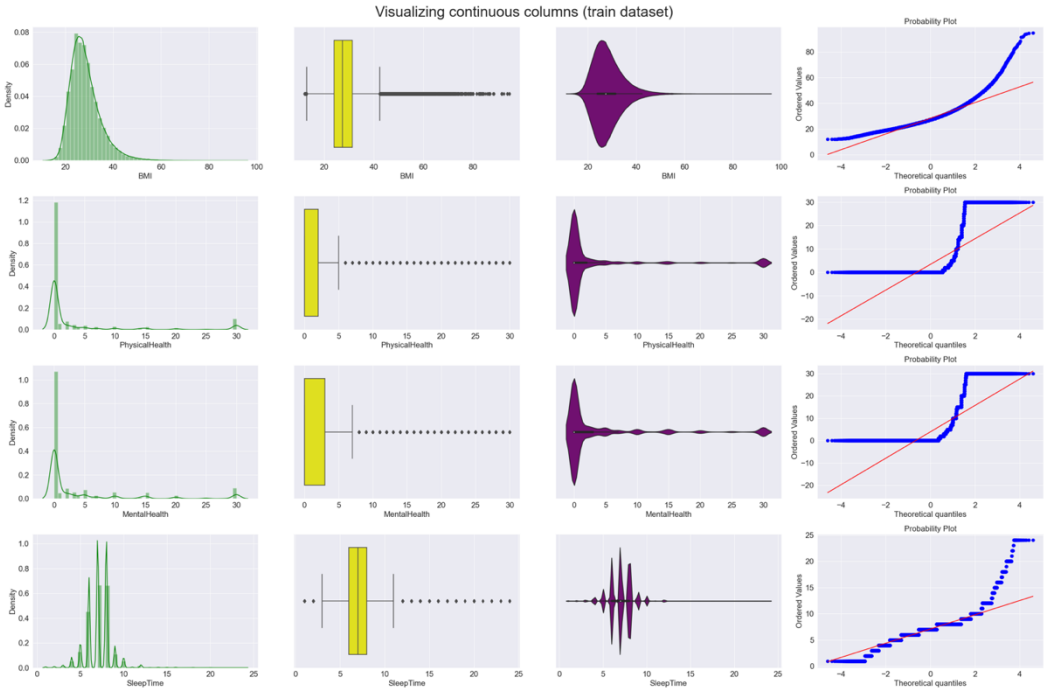

Figure 1 Outlier Detection in Numerical Variables

The analysis revealed that out of all the variables examined, only the BMI (Body Mass Index) variable was approximately normally distributed.

In contrast, the other variables were found to be close to bimodal in their distribution.

Therefore, this finding suggests that the BMI variable is likely to predict heart disease risk better as it has a more typical distribution pattern than the other variables. However, the bimodal distribution of the other variables may indicate the presence of different subpopulations or underlying processes, which could be further explored in future analyses.

Missing Value Treatment:

Figure 2 implies that the heart disease prediction dataset used for analysis lacks any missing values. Given dataset is free of any errors, inconsistencies, or missing values. Since there are no missing values in the heart disease prediction dataset, it is clean.

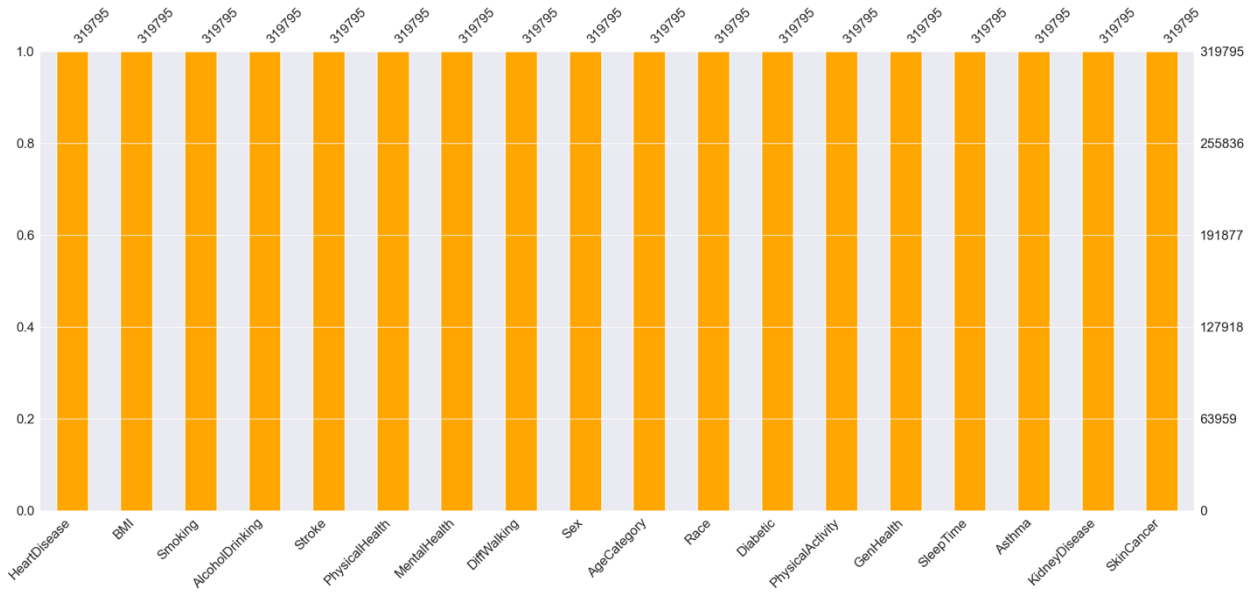

Figure 2 Missing Value Plot

Correlation Matrix:

Identifying the correlations between different categories and features can help to understand the underlying relationships between the variables and their impact on heart disease. This information can help develop predictive models and identify potential risk factors associated with heart disease.

A correlation table was created to identify the relationships between categories in the heart disease prediction dataset. The table helps to measure the strength of the correlation between different variables in the dataset. Figure 3 shows that heart disease is highly correlated to general health.

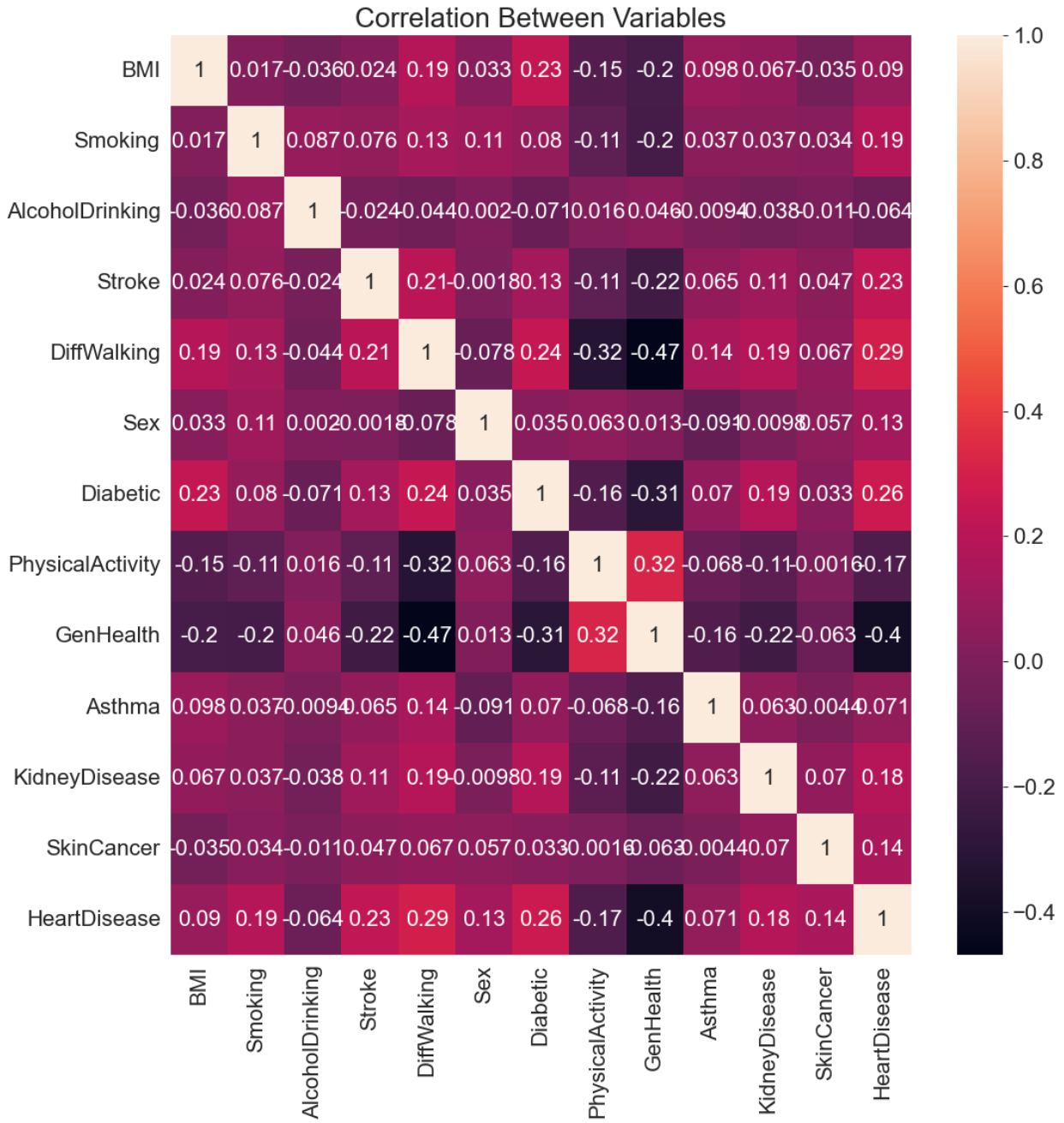

Figure 3 Correlation Matrix

CHAPTER 3- DATA VISUALIZATION

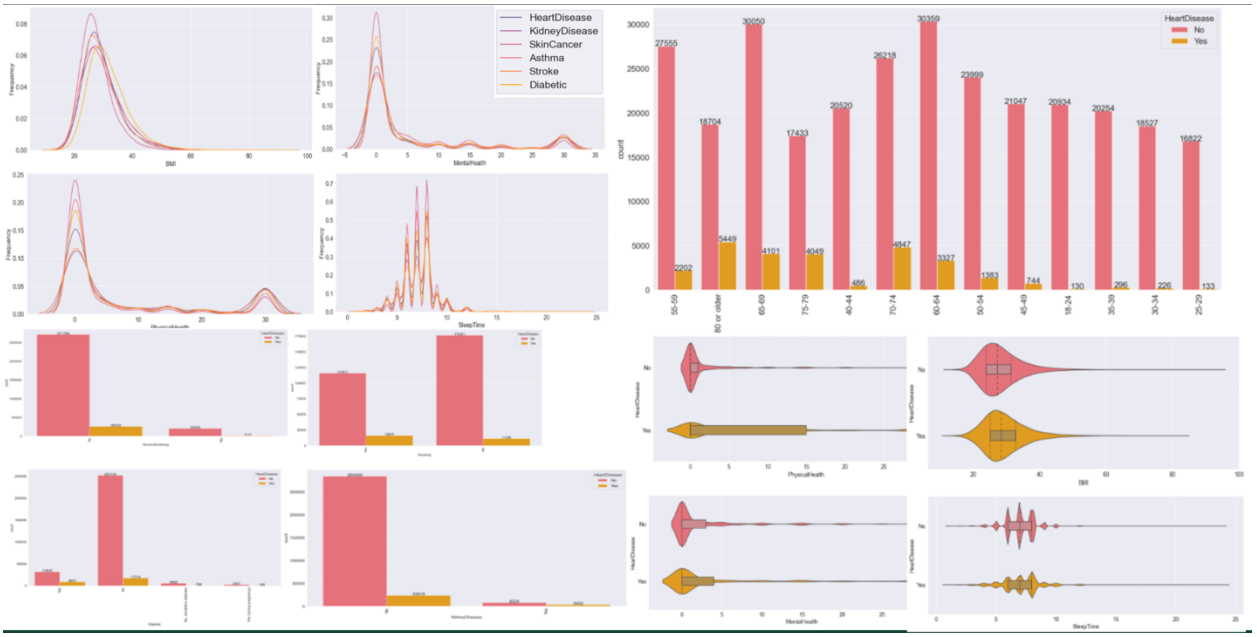

Figure 4 Data Visualizations Plot

Insights:

1. Number of people who smoke and have heart disease is more than they don't.
2. Number of people who don't drink alcohol and have heart disease is more than they don't.
3. Males have heart disease more than females.
4. People who have difficulty walking can have heart disease more than don't.
5. People who have a stroke can have heart disease than those who don't.
6. People older than 40 years old can have heart disease more than others.
7. Gen health for most people is very good, and all can have heart disease in equal proportions.
8. People who are not physically active can get heart disease more than others.
9. Most people in our data are white and have no diabetes.
10. Most of them had done physical activity during the past 30 days other than their regular job and generally have very good health, as they said.
11. A few of them have asthma, kidney disease, and skin cancer.
12. Most people say that they have generally very good health. A few people said that they generally have poor health.
13. In our sample, around 8 among 100 individuals suffer from heart disease.
14. The BMI of heart disease patients is slightly higher than that of healthy individuals.
15. The older the individual, the more susceptible they are to heart disease.
16. ~10% of males suffer from heart disease, while only ~7% of females do.
17. The percentage of heart disease is highest (> 10%) among Native Americans, followed by whites (~9%). The least percentage of heart disease (~3%) is among Asians.
18. A lot more people who suffer from heart disease say they have poor or fair health compared to those who don't.
19. 79% of healthy individuals have been physically active in the past 30 days, compared to 64% in heart disease patients.

20. Abnormal sleep duration is more prevalent in heart disease patients. Even though heart disease patients make up 8.5% of the sample, they have higher percentages of sleep, less than 6 hours or more than 9 hours, which is considered abnormal.
21. ~12% of people who smoke suffer from heart disease. In contrast, ~5% of non-smokers suffer from heart disease.
22. Surprisingly, people who drink alcohol have a lower percentage of heart disease (~4%) than those who do not (~9%).
23. Having a stroke is highly correlated with heart disease. People who have had a stroke before have a heart disease percentage of around 48%. On the other hand, people who did not suffer a stroke had a significantly lower percentage of heart disease (~8%).
24. Diabetic people are at higher risk of heart disease (~25%).
25. Asthmatic people are at a slightly higher risk of heart disease.
26. Those who have suffered from kidney disease are at a significantly higher risk of heart disease. With a percentage of ~30% compared to ~9% in healthy people.
27. People who suffer from skin cancer are at a moderately higher risk of heart disease (~18% vs ~9%).
28. Difficulty in walking is present in ~18% of heart disease patients vs. ~7% in healthy individuals.
29. The BMI distribution differs slightly in patients with different diseases. Diabetic people have the highest BMI mode, and stroke victims have the lowest BMI mode.
30. Mental health, sleep duration, and physical health are similar among people who suffer from different diseases.
31. ~64% of people who say they have poor health are smokers. While people who say they have excellent health are 30% smokers.

CHAPTER 4- DATA MODELLING

Data Splitting:

To develop a predictive model for heart disease, the dataset is divided into two subsets - the training dataset and the testing dataset. The training dataset comprises 80% of the original dataset and is used to train the predictive model. The remaining 20% of the dataset is used as the testing dataset to evaluate the model's performance.

To assess the effectiveness of different classifiers in predicting heart disease, various machine learning algorithms such as decision tree classifier, random forest classifier, Logistic Regression, Gradient Boosting, Adaptive Boosting, Category Boosting, XG Boosting, Linear SVC, Categorical Naïve Bayes are applied to the clustered dataset. The performance of each classifier is evaluated using several metrics, including accuracy, precision, recall, and F-measure scores. The evaluation metrics help to determine the effectiveness and accuracy of the models in predicting the occurrence of heart disease in the dataset.

Feature Selection and Data Pre-Processing

1. Data Balancing:

As we can see from the count plot in the figure below, we observe that our labels are not balanced. This is a huge problem, and we need to use the necessary measure to deal with it. Balancing is important in machine learning because it addresses the class imbalance in the dataset, where one class has significantly fewer instances than the other class(es). In such cases, the machine learning model may be biased towards the majority class, leading to poor performance on the minority class.

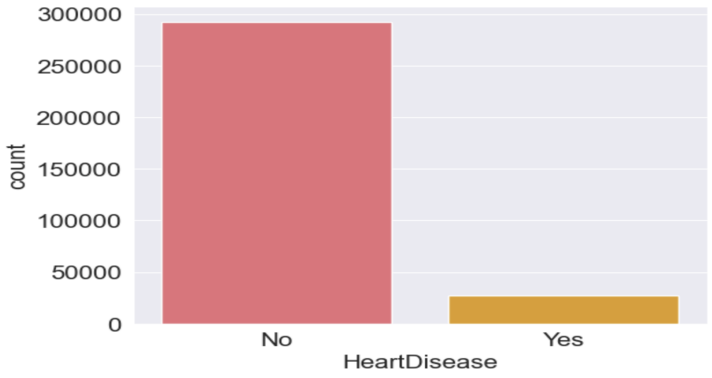

Figure 5 Target Variable- Before Balancing

Random oversampling is a technique used in machine learning to address the class imbalance in a dataset where the number of instances in one class is much smaller than the number of instances in the other class. Figure 6 represents the results after random oversampling for the target variable.

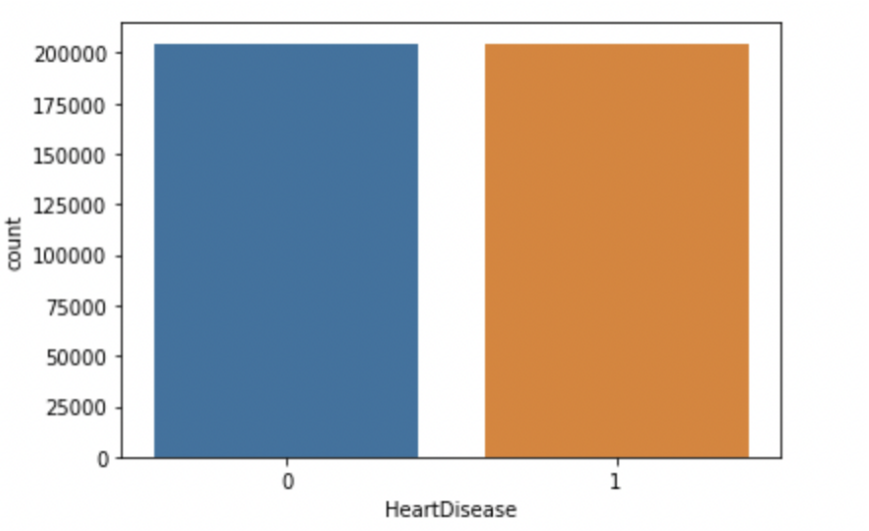

Figure 6 Target Variable- After Balancing

2. **Encoding Categorical Variable:**

Encoding converts non-numerical data into numerical representations that machine learning algorithms can process.

- Attribute "sex" in the dataset has been encoded as **0 for female, 1 for male**
- Attribute "genhealth" in the dataset has been encoded as 4 **for excellent, 3 for Very Good, 2 for good, 1 for fair, 0 for poor.**
- Attribute "diabetes" in the dataset have been encoded as **0 for 'No', 1 for 'No, borderline diabetes':1, 2 for 'Yes (during pregnancy)' and 3 for 'Yes'**
- Other Attributes like 'HeartDisease', 'Smoking', 'AlcoholDrinking', 'Stroke', 'DiffWalking', 'PhysicalActivity', 'Asthma', 'KidneyDisease' and 'SkinCancer' which have unique values as "Yes" and "No" have been encoded as **0 for 'No' and 1 for 'Yes'**

3. **Encoding Numerical Attributes:**

Numerical attributes like BMI have been encoded as they are provided in a range that can be encoded ordinally. As per the range provided in the figure below, we have ordinally encoded the range of BMI.

Category	BMI range - kg/m^2
Severe Thinness	< 16
Moderate Thinness	16 - 17
Mild Thinness	17 - 18.5
Normal	18.5 - 25
Overweight	25 - 30
Obese Class I	30 - 35
Obese Class II	35 - 40
Obese Class III	> 40

Figure 7 Encoding classes for BMI

Similarly, we have encoded the Age column, as age data is also being provided in ranges like 0 for '18-24',1 for '25-29',2 for '30-34',3 for '35-39' and so on.

For the attribute Sleep time, as per the graph, we see that most of the values have been concentrated in a region from 5 to 10. For values less than five, we have encoded them into one category and more than ten into other categories. Values between 5 to 10 have been put into other categories.

Decision Tree:

Decision trees are structures resembling trees that are utilized for handling extensive data sets. They are commonly illustrated as diagrams, where the branches outside represent outcomes, and the internal nodes stand for the characteristics of the data set. The reason decision trees are popular is that they are effective, dependable, and straightforward to comprehend. The anticipated category label for a decision tree arises from its starting point. The succeeding stages in the tree are determined by matching the root attribute value with the information stored in the record. For the given modeling purpose, our decision tree model was trained at default parameters with random state=1. The performance of the decision tree model was evaluated using a hold-out test set. The following confusion matrix was used to assess the model performance:

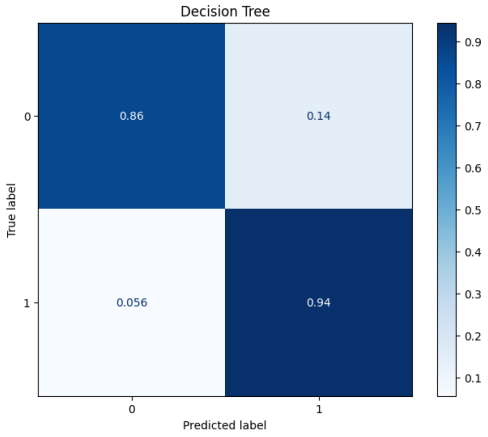

Figure 8 Decision Tree- Confusion Matrix

Model Training and Tuning:

Random Forest:

The random forest algorithm is a supervised classification method that involves using multiple decision trees to work together as a team. The model's prediction is based on the class with the most votes. The decision trees in the random forest predict classes, which removes the constraints of the decision tree algorithm. This enhances accuracy and reduces the overfitting of the data. The random forest approach can still provide accurate results on large datasets, even if a significant amount of record values is missing. The decision tree samples can be saved and used with different types of data. The performance of the Random Forest model was evaluated using a hold-out test set. The following confusion matrix was used to assess the model performance:

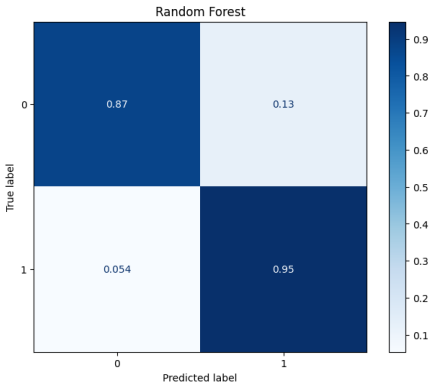

Figure 9 Random Forest- Confusion Matrix

Logistic Regression:

Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied to the odds—that is, the probability of success divided by the probability of failure. For the given modeling purpose, our decision tree model was trained at default parameters with random state=1 and 500 iterations. The performance of the logistic model was evaluated using a hold-out test set. The following confusion matrix was used to assess the model performance:

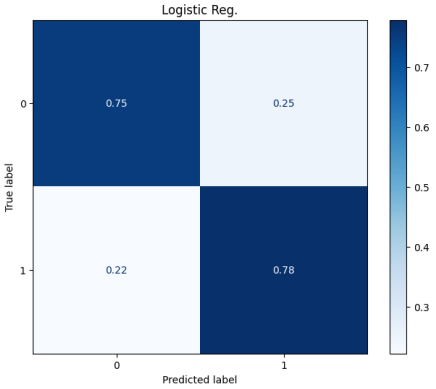

Figure 10 Logistic Regression- Confusion Matrix

Gradient Boosting:

Gradient Boosting works by adding weak learners (i.e., regression trees) to a model in a forward stage-wise fashion. At each stage, the algorithm fits a new regression tree to the negative gradient of the loss function with respect to the current model predictions. The negative gradient represents the direction of the steepest descent of the loss function, and the regression tree is trained to approximate this gradient. The performance of the GB was evaluated using a hold-out test set. The following confusion matrix was used to assess the model performance:

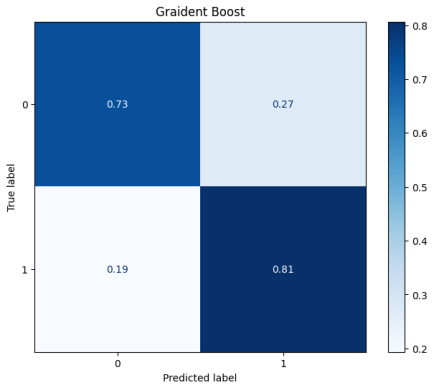

Figure 11 Gradient Boosting- Confusion Matrix

Adaptive Boosting:

Adaptive Boosting, or AdaBoost, is a machine learning algorithm for classification and regression problems. AdaBoost works by iteratively training a series of weak learners, such as decision trees, and combining them into strong learners. The performance of the AdaBoost model was evaluated using a hold-out test set. The following confusion matrix was used to assess the model performance:

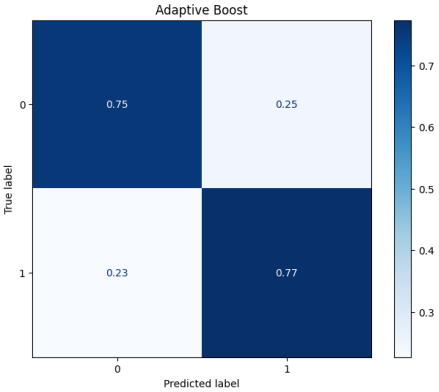

Figure 12 AdaBoosting- Confusion Matrix

CatBoosting:

CatBoost uses a technique called ordered boosting, a variant of gradient boosting specifically designed to handle categorical variables. It uses an algorithm that sorts the categorical variables based on the target variable. Then it creates a numerical representation of the categorical variable based on the order of the categories. The performance of the CatBoost model was evaluated using a hold-out test set. The following confusion matrix was used to assess the model performance:

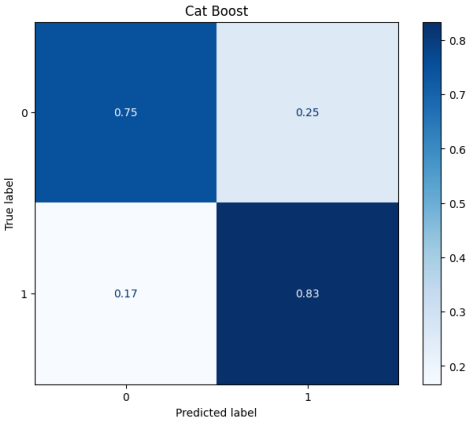

Figure 13 CatBoosting- Confusion Matrix

XG Boosting:

XGBoost is a variant of gradient-boosted decision trees where decision trees are constructed sequentially. The model assigns weights to all independent variables, which are then utilized to make predictions through decision trees. If a tree makes an incorrect prediction, the importance of the relevant variables is increased and used in the next tree. Finally, the predictions of each of these models are combined to create a more robust and more precise model.

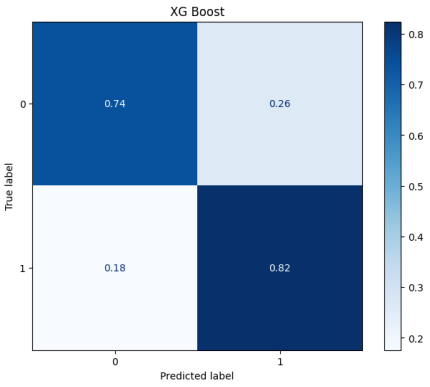

Figure 14 XG Boosting- Confusion Matrix

Linear SVC:

Linear SVC stands for Linear Support Vector Classification. It is a machine-learning algorithm used for binary classification tasks. It works by finding the hyperplane that best separates the two classes of data points. The hyperplane is chosen to maximize the margin between the classes, with the margin being defined as the distance between the hyperplane and the closest data points from either class.

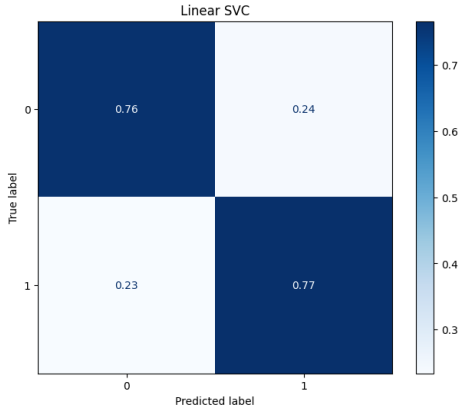

Figure 15 Linear SVC- Confusion Matrix

Categorical Naïve Bayes:

Categorical Naive Bayes (NB) is a probabilistic algorithm used for classification tasks in machine learning. It is particularly suited for text classification problems, where each data point is a text document, and the goal is to assign it to one of several predefined categories or classes. The algorithm uses Bayes' theorem to calculate the probability of a document belonging to each class, given its features or words. It assumes that the features are conditionally independent of each other, which simplifies the calculation of the probabilities.

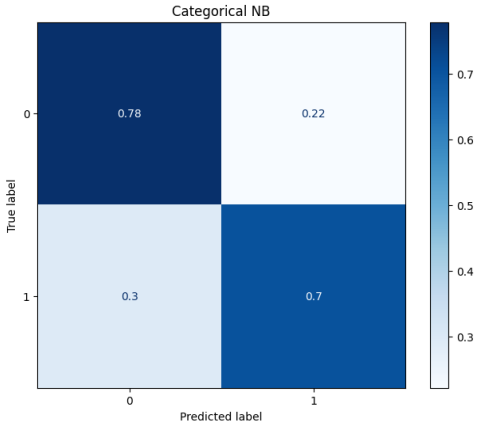

Figure 16 Categorical NB- Confusion Matrix

Model Evaluation and Results:

We use the following metrics to evaluate the models' performance:

1. **Accuracy:** This is the essential evaluation metric used to measure how often the model makes correct predictions. It is calculated by dividing the number of correct predictions by the total number of predictions.
2. **Precision and Recall:** These are evaluation metrics commonly used for classification problems. Precision measures how many of the predicted positive examples are actually positive, while recall estimates how many of the positive examples in the data are correctly predicted.
3. **F1 Score:** This is a harmonic mean of precision and recall and is often used as an evaluation metric for imbalanced datasets.

Figure 13 shows the various metrics for the model mentioned above to evaluate the performance and find the best model.

	Model	Train Score	Test Score	Recall	Precision	f1-score
0	Decision Tree	0.933174	0.901401	0.944088	0.869829	0.905439
1	Random Forest	0.933164	0.910241	0.946414	0.882566	0.913375
2	Logistic Reg.	0.763204	0.764544	0.778576	0.757326	0.767804
3	Graident Boost	0.767756	0.768845	0.806583	0.749980	0.777252
4	Adaptive Boost	0.761740	0.763604	0.773190	0.758648	0.765850
5	Cat Boost	0.793877	0.791312	0.833581	0.768607	0.799777
6	XG Boost	0.784603	0.783148	0.824228	0.761653	0.791706
7	Linear SVC	0.762067	0.763459	0.766316	0.761965	0.764134
8	Categorical NB	0.739582	0.741530	0.704557	0.760820	0.731608

Figure 17 Model Evaluation Metrics

Key Business Insights:

From our project, we drew some key business insights and the probabilities of that scenario happening:

- If a person does not smoke, has a stroke, is diabetic, and has kidney disease, the probability of that person having a heart disease is 56%
- If a person smokes, has/had stroke and is inactive, is diabetic and does not have kidney disease the probability of that person having a heart disease is 53%

More insight on the probability of having heart disease based on various combinations of activities and disease is shown in the figure below:

group	percent
Non Smoking & Stroke & Active & Diabetic & KidneyDisease	56.626506
Non Smoking & Stroke & Inactive & Diabetic & KidneyDisease	56.687898
Non Smoking & Stroke & Inactive & No Diabetic & KidneyDisease	52.755906
Smoking & Stroke & Active & Diabetic & KidneyDisease	58.918919
Smoking & Stroke & Inactive & Diabetic & KidneyDisease	64.197531
Smoking & Stroke & Inactive & Diabetic & No KidneyDisease	53.992395
Smoking & Stroke & Inactive & No Diabetic & KidneyDisease	55.172414
Non Smoking & No Stroke & Active & No Diabetic & No KidneyDisease	3.519108
Non Smoking & No Stroke & Active & Diabetic & No KidneyDisease	11.808529
Non Smoking & No Stroke & Active & No Diabetic & KidneyDisease	14.488518
Non Smoking & Stroke & Active & No Diabetic & No KidneyDisease	23.481258
Smoking & No Stroke & Active & No Diabetic & No KidneyDisease	7.034865

Figure 18 Key Business Insights

CHAPTER 5- CONCLUSION

The development of an accurate system to predict heart disease has become necessary due to the increasing number of deaths caused by this disease. Various machine learning algorithms have been tested and compared using different datasets to achieve this. Among these algorithms, random forest had the highest F-score of 91.33%. In comparison, the decision tree algorithm showed an F-score of 90.54% for the classification of heart disease using the UCI machine learning repository dataset.

However, the study was limited to only detecting certain types of heart disease, as the UCI dataset did not contain sufficient features to diagnose other heart diseases such as Heart Valve Disease, Pericarditis, and congenital heart disease. Therefore, further research is needed to develop a web application based on the Naïve Bayes algorithm and to use an actual database with more features to classify other types of heart diseases. Overall, using machine learning to classify heart disease is a promising field that can benefit both healthcare professionals and patients.

References

Dataset:

https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot

1. S. Pouriyeh et al., "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in Proceedings of IEEE Symposium on Computers and Communications (ISCC). Heraklion, Greece: IEEE, July 2017, pp. 204-207,
2. L. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective." Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89-109, 2001.
3. J. Soni et al., "Intelligent and effective heart disease prediction system using weighted associative classifiers," International Journal on Computer Science and Engineering, vol. 3, no. 6, pp. 2385-2392, 2011.
4. M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, go. 01, pp. 1-16, 2017.
5. S. Pouriyeh et al., "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in Proceedings of IEEE Symposium on Computers and Communications (ISCC). Heraklion, Greece: IEEE, July 2017, pp. 204-207,
6. R. Das, L. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert systems with applications, vol. 36, no. 4, pp. 7675-7680,2009,
7. N. Waghulde and N. Patil. "Genetic neural approach for heart disease prediction," International Journal of Advanced Computer Research, vol. 4, no. 3, pp. 778, 2014
8. N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbor classification technique." in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21-26.
9. H. Almarabeh and E. Amer, "A study of data mining techniques accuracy for healthcare," International Journal of Computer Applications, vol. 168, no. 3, pp. 12-17, Jun 2017.