Heart Disease Prediction Model

***Abstract*—Coronary heart disease remains the leading cause of mortality in the United States. Various factors may contribute to heart disease such as sex, age, ethnicity, and race, all of which are important to be aware of to help reduce, control, and prevent heart disease. Throughout this report, we will explore different supervised machine learning algorithms and build predictive models to accurately identify the presence of heart disease based on the given data of a patient. Our method involves extracting key attributes and utilizing them to train a Bagged Decision Tree, Random Forest, and Logistic Regression model. An accurate and timely diagnosis can aid doctors in employing early intervention techniques, which have a higher success rate, effectively reducing the mortality rate. We evaluated the performance of the three models and found that our Logistic Regression model had an accuracy of 77.56%, precision of 79.79%, recall of 73.53%, and an F1 Score of 76.53%. On the other hand, our Bagged Decision model had an accuracy of 98.54%, precision of 97.14%, recall of 100%, and an F1 Score of 98.55%. Finally, Random Forest has an accuracy of 98.54%, precision of 97.14%, recall of 100%, and an F1 score of 98.55%. We found that the Bagged Decision Tree model yielded superior results in heart disease prediction when compared to the other two types of models.**

Keywords: machine learning, heart disease prediction, classification, logistic regression, bagged decision tree, random forest

I. Introduction

In the United States, there were about 695,000 deaths in 2021 due to coronary heart disease, which is 1 in every 5 deaths [1]. Early detection and prevention of such conditions are crucial in lowering the mortality rate in the United States as well as decreasing the overall costs of health services related to heart disease. When it comes to heart disease, there are three key risk factors, high blood pressure, high cholesterol, and diabetes, with 47% of all Americans having one of the three [2]. Machine learning provides a promising approach to enhancing the diagnostic process by efficiently analyzing datasets and predicting disease outcomes. This report aims to utilize various machine learning techniques to predict the presence of heart disease based on a range of parameters.

Specifically, we plan to implement Bagged Decision Tree, Logistic Regression, and Random Forest models. Logistic Regression will serve as a baseline for our predictions, given its efficiency in binary classification tasks. On the other hand, Bagged Decision Tree and Random forest will allow us to leverage an ensemble learning approach to ideally improve prediction accuracy by addressing variance and bias in the data, thereby providing a more robust model capable of handling the complexities of the dataset used.

Our exploratory data analysis (EDA) reveals significant insights into the dataset, which consists of several features such as 'age', 'sex', 'cholesterol levels', 'resting blood pressure', and many more features. Key findings from the analysis indicate diverse distributions across different variables, with some, such as maximum heart rate and serum cholesterol, showing potential as predictors of heart disease due to the variance found between individuals with and without heart disease.

A. Significance of the Study

The significance of this study includes improving healthcare outcomes through better predictive models and reducing the economic burden associated with cardiovascular diseases. By increasing the predictive accuracy of heart disease diagnoses, healthcare professionals can offer more targeted interventions sooner, potentially saving lives and resources. Additionally, the study's findings could contribute to the broader field of medical research by providing methodologies that can be applied to other areas of healthcare. The development of more sophisticated predictive models can serve as a foundation for tackling various other dangerous diseases. Therefore, we are able to provide an intervention before it arises more severely.

B. Organization of the Paper

The remainder of the paper is organized as follows: Section II provides a comprehensive literature review. Section III describes the dataset and the exploratory data analysis in detail. Section IV discusses the methodology, including the machine learning models employed and their configurations. Section V presents the results and evaluations of the models, and Section VI concludes with a discussion of the findings and future directions for this research.

II. Literature Review

In the article, Heart Disease Prediction using Machine Learning Techniques by Devansh Shah, Samir Patel, and Santosh Kumar Bharti, several machine learning algorithms have been explored for predicting heart disease. Each of them shows their strengths and weaknesses.

A. Naïve Bayes Classifier

The simple probabilistic classifier is based on Bayes' theorem with independence assumptions between features. It is easy to implement and efficiently handle large datasets with non-linear relationships. However, its downside is loss of accuracy as the assumption of independence between attributes. Studies report that the Naive Bayes Classifier achieved 84.16% using a subset of important predictors.

B. Decision Trees

Decision trees create a model that predicts the value of a target variable based on predefined decision rules from data features. They are interpretable and handle both categorical and numerical data well. However, they can be prone to overfitting. This report shows accuracies range from 42.89% to 71.43% depending on the dataset.

C. K-Nearest Neighbor (K-NN)

This algorithm classifies instances based on the majority class among the k-nearest neighbors. It is effective for small datasets but can be computationally expensive for large datasets. This algorithm is versatile and commonly used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy. the report shows that it has 83.16% which demonstrates its potential for high accuracy with proper parameter tuning.

D. Random Forest

This is a method that constructs multiple decision trees during training and outputs the mode of the classes. It is efficient in robustness to overfitting and handles missing values. Studies report that Random Forest has the highest accuracies overall, it ranges from 91.6% to 97% depending on different types of forests or datasets.

Finally, inspired by the findings of our preliminary research, we implemented and evaluated linear regression and bagged decision tree models to determine their accuracies on our dataset. Our results indicate that the bagged decision tree outperformed logistic regression. The random forest model also demonstrated competitive performance since they are both ensemble methods of constructing multiple decision trees to enhance prediction stability and accuracy.

III. Dataset Description and Exploratory Data Analysis

A. Dataset Description

For our report, we will be utilizing a University of California, Irvine Heart Disease Dataset which contains values from four databases: Cleveland, Hungary, Switzerland, and Veteran Affairs Long Beach. The dataset is composed of 1025 instances, each with 14 attributes, including a target feature of either 0 or 1, with 0 indicating no presence of heart disease and 1 indicating the presence of heart disease. The 14 attributes include:

1) **age**: the patient's age (in years)
2) **sex**: sex of the patient (1 = male; 0 = female)
3) **cp**: type of chest pain experienced (0 = asymptomatic; 1 = atypical angina; 2 = non-anginal pain; 3 = typical angina)
4) **trestbps**: resting blood pressure (in mm Hg (Mercury) on admission to hospital)
5) **chol**: patient's serum cholesterol level (in mg/dl)
6) **fbs**: patient's fasting blood sugar level (¿ 120 mg/dl, 1 = true; 0 = false)
7) **restecg**: resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricle hypertrophy)
8) **thalach**: patient's maximum heart rate achieved
9) **exang**: exercise-induced angina (1 = yes; 0 = no)
10) **oldpeak**: ST depression induced by exercise relative to rest
11) **slope**: the slope of the peak exercise ST segment (0 = up-sloping; 1 = flat; 2 = down-sloping)
12) **ca**: number of major blood vessels colored my fluoroscopy
13) **thal**: blood disorder Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
14) **target**: presence of heart disease (0 = no; 1 = yes)

B. Data Observations

When building a predictive model, understanding both the nature of the data we are using, the distributions of said data, and the population being studied in the dataset, can provide insight and help identify any potential issues that could negatively impact the performance of our models.

Figure 1 shows both the age and sex distributions for the patients in the dataset. We observed that the average age for the patients was 54 years, with the total range of ages being from 22 to 77 years. Since our age range is quite large, this diversity will help our models generalize better for unseen data that may contain different age groups. The American Heart Association (AHA) reports that about 40% of people aged 40-59, and 75% of people aged 60-79 have heart disease, and 86% of people over 80 have heart disease [3]. The average age of our dataset aligns with the common age groups that are at higher risk of heart disease, which assures us that our dataset is useful for clinical applications. There is a slight gender imbalance as 70% of the patients in the dataset are male, which is something to consider when evaluating the performance and generalizability of our models.

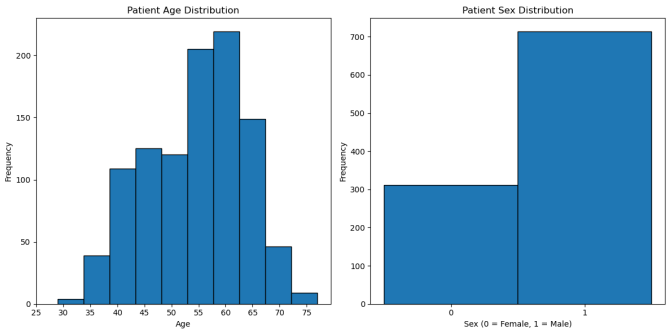

Fig. 1. Patient Age and Sex Distributions

Overall heart disease distribution in the dataset was found to be 51.3% No Heart Disease, and 48.7% Heart Disease.

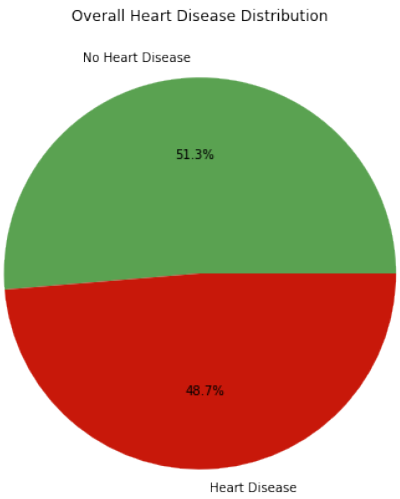

Fig. 2. Overall Heart Disease Distribution. This figure encapsulates the distribution amongst both males and females. Overall, 48.7 Percent of individuals did have heart disease. And 51.3 Percent of people didn't have heart disease.

Looking at sex distribution with respect to heart disease specifically, we can find heart disease rates by gender.

For heart disease distribution in males, we find that 57.03% of males do not have heart disease and 42.97% of males have heart disease.

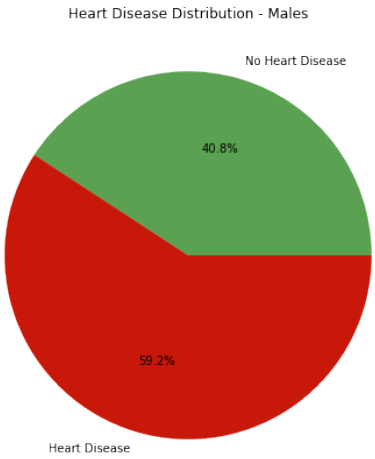

Fig. 3. Male Heart Disease Distribution. Overall, 59.2 Percent of males did have heart disease. And 40.8 Percent of males didn't have heart disease.

For heart disease distribution in females, we find that 82.77% of females do not have heart disease and 17.23% of males have heart disease.

By plotting out these results, we find that that males in the dataset have more heart disease than females.

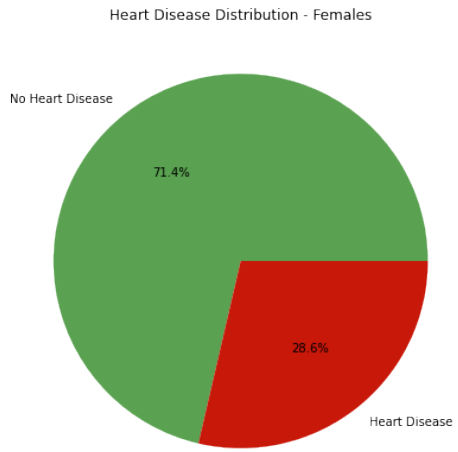

Fig. 4. Female Heart Disease Distribution. Less than male.

C. Other Feature Observations

Figures 5 and *6* show the histograms for chest pain and resting blood pressure. With this, we can visualize the distributions for each, and we can note that the most common type of chest pain (cp) is asymptomatic. For resting blood pressure (trestbps), there was a range of 94-200 mm Hg, with an average of about 131 mm Hg. Analyzing the distributions of the rest of the features, serum cholesterol (chol) levels ranged 126 to 564 mg/dl, with an average of 246 mg/dl. About 15% of patients had a fasting blood sugar (fbs) level that was greater than 120 mg/dl. Most patients had normal resting electrocardiographic results (restecg). Maximum heart rate achieved (thalach) had a range of 71 to 202, with an average of 149. Around 34% of patients experienced exercise-induced angina (exang). The average ST depression induced by exercise (oldpeak) was 1.07, with the range of values being from 0 to 6.2. Most patients saw a down-sloping peak exercise ST segment (slope). The average number of major blood vessels colored by fluoroscopy (ca) for patients was 0. Finally, 51.32% of patients in the dataset do not have heart disease while the other 48.68% do have heart disease.

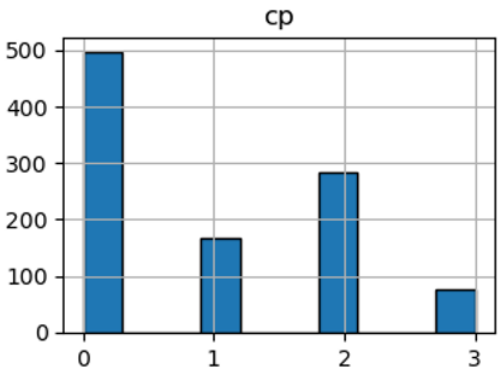

Fig. 5. Histograms for Chest Pain (cp)

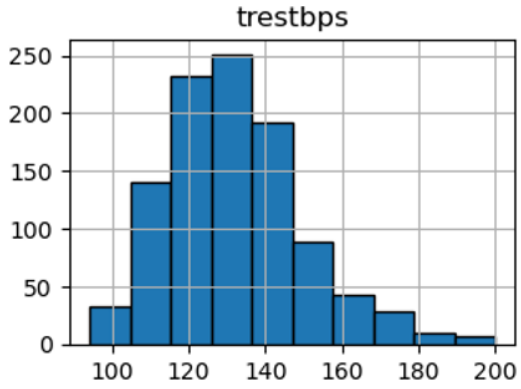

Fig. 6. Histograms for Resting Blood Pressure (trestbps) (The plot resembles a normal distribution curve, that is slightly skewed as well)

D. Correcting Target Data

When examining the original Kaggle dataset and plotting the target value with respect to features, we found unconventional results. This is highlighted in *Figure 7*, where it was shown that younger patients had a higher incidence of heart disease. Likewise, older patients had less heart disease.

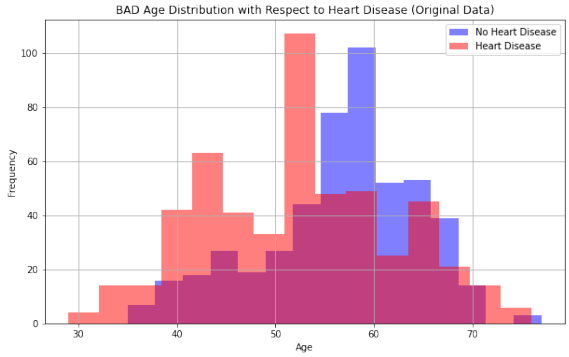

Fig. 7. Original Incorrect Dataset (Younger Patients Had More Heart Disease)

These results contradict conventional scientific wisdom, that older patients should have a higher incidence of heart disease. To determine if the dataset was still viable, we conducted further research. In a paper by Brandon Simmons II, published in May 2021, it was concluded that "the target variable coding for the Kaggle data should be reversed." After implementing these changes to the dataset, we found more expected and consistent results when plotting our age distribution, as shown in *Figure 8*.

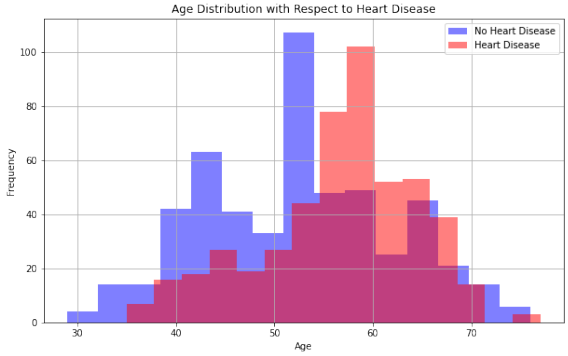

Fig. 8. Correct Dataset Age Distribution Results

After inverting the target value for the dataset, the dataset became corrected as shown by the corrected age distribution values.

E. Removing 'thal' From Dataset

While analyzing the features in the dataset, we found unclear descriptions for the 'thal' value. It was found that in the Kaggle dataset, 'thal' has values of 0,1,2,3. However, in the Cleveland dataset, 'thal' has values of 3,6,7.

The Kaggle dataset didn't describe the 'thal' value properly. It gives descriptions for 'thal' values 1, 2, and 3, but no description for the 0 value. The 0 'thal' value had a count of 7 in the dataset.

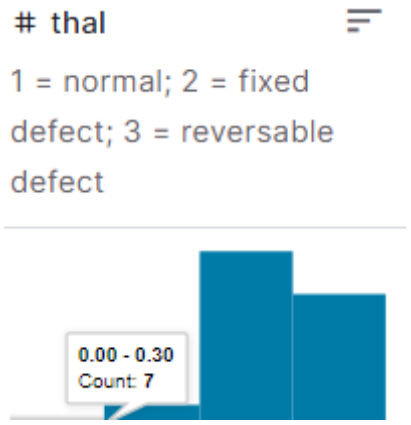

Fig. 9. 'thal' in Kaggle Dataset. Dataset doesn't describe the 0 value.

While exploring this discrepancy for the 'thal' value, we encountered research by Simmons addressing this issue in the dataset. Simmons writes the following about the 'thal' value in the Kaggle Dataset, explaining that "because each feature is categorical... It appears that each of their missing values was replaced in the Kaggle data set". Furthermore, he explains that "since these two features do not appear in the dataset", he "decided to drop it" for his predictive model. Dropping the 'thal' feature is a "minor difference" because of the data size [8].

Taking into account Simmon's research, and the implications towards our own model. We felt that it was necessary to drop the feature entirely to improve the overall efficiency of our model and reduce any errors which may arise from keeping bad values in our dataset.

Therefore, the 'thal' feature was not used for building our predictive models.

F. Feature Correlation

Before implementing ML algorithms, let us first analyze our dataset to discover whether there are any relations between features or the target variable. *Figure 10* shows the visualization of the correlation matrix for our dataset. From the figure we can observe that chest pain (cp) has the highest correlation to the target variable, followed by max heart rate (thalach), and the slope of peak exercise ST segment (slope), showing that these features are strong predictors of heart disease.

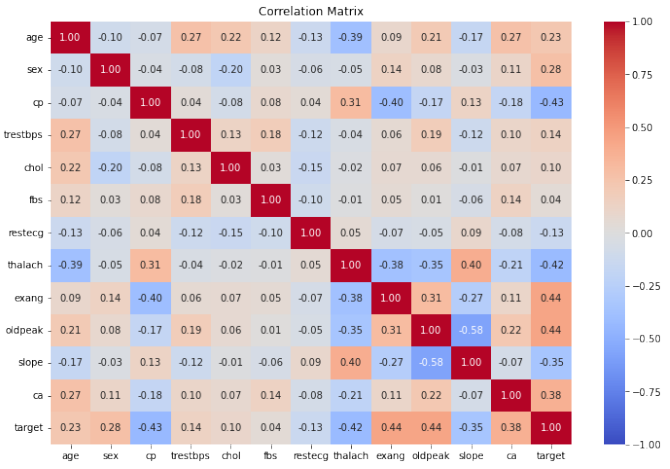

Fig. 10. Feature Correlation Matrix with Heatmap. Maximum Heart Rate achieved (thalach) and Chest Pain (cp) were more correlated with Heart Disease.

G. Outlier Detection

Checking our dataset for any outliers is an important step of EDA as they can reduce the overall accuracy of our model. We calculated the number of outliers by using the first and third quartiles of each feature and using those to compute the Inter-Quartile Range. We then flagged every data point that either fell below the minimum ($min = Q1 - 1.5 * IQR$) or above the maximum ($max = Q3 + 1.5 * IQR$) as outliers. *Figure 11* shows the box plots for each feature in our data set, which includes a visualization of the various outliers present (indicated by a circle). In total, we discovered 256 outliers.

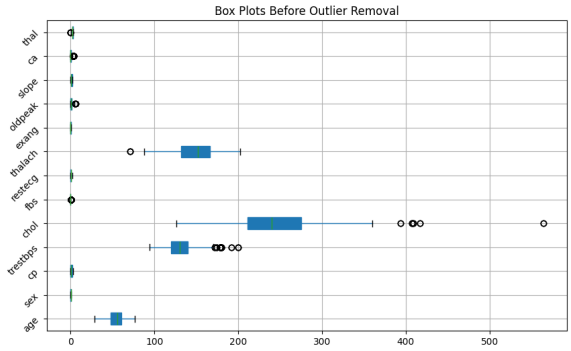

Fig. 11. Box plots for all features in the dataset. Cholesterol (chol) and Resting Blood Pressure (trestbps) were found to have most outliers.

Below is a closer look at the outliers for Cholesterol (chol) data and Max Heart Rate (thalach) data.

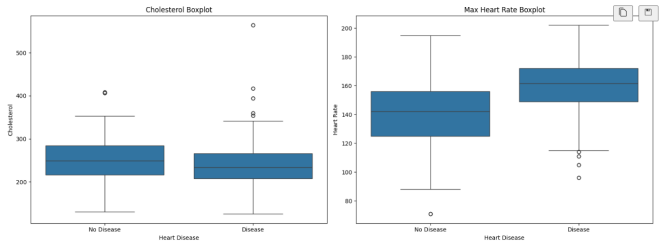

Fig. 12. Box plot for Cholesterol and Max Heart Rate Values. Both of these features were shown to have outliers.

IV. Proposed Methodology

The project that we have built employs a multitude of machine learning models. The models are Logistic Regression, Bagged Decision Tree, and Random Forest Tree. With these particular models in place, we have developed a robust web-based application capable of predicting heart disease based on inputted clinical attributes. The goal of this report is to compare the different predictive accuracies of different models and to identify the most significant features that contribute to the prediction of whether or not a particular 'user' has heart disease or not. When selecting the important features, our dataset was pretty simple. So we felt that it was not important to exclude any features since it didn't contribute to making the model(s) perform in a noticeably different way.

A. Model Selection

When selecting the models to develop, our team decided to pick the following three models. The Bagged Decision Tree model, the Random Forest Model, and the Logistic Regression model. We compare the 'accuracy' of these three models amongst each other, and then evaluate which one is the more suited for utilization. The model chosen is aimed to be a binary classifier. In the root word, binary, you either do, or you don't.

In terms of this model's aim, you either have heart disease, or you don't. The section below essentially outlines the rationale for each selection of each of the according models. In the following section(s), we also assess how the model does by testing it with a validation set. (The validation set is often referred to as the testing set in Machine Learning terminology, the two terms are synonymous).

1) Bagged Decision Tree: The bagged decision tree is known as an ensemble tree. The ensemble tree is known as a general method to increase the stability and accuracy of ML algorithms by combining multiple decision trees together with the respective computed values that they have. This is where the term 'forest' comes from. Utilizing all of these decision trees unanimously allows us to reduce variance and prevent overfitting of the model. With this methodology, our model is ideal for various types of data that would come in.

2) Random Forest Model: This then leads us to the random forest model. (It's known as an extension of the bagged decision tree). 'Randomness' is introduced into the model. Each 'tree' essentially takes into account a randomness factor which then leads to each tree being built from a random sample of features. This method introduces even more accuracy than before since it causes the variance to increase. It also gives us very important information regarding the predictors of heart disease.

3) Logistic Regression Model: This is the simplest model out of the three models. This is essentially known as the baseline model for any ML task. As a baseline model, logistic regression is efficient in binary classification problems. It provides a probabilistic framework for modeling the presence or lack of presence of heart disease. It is a classification model that is extremely intuitive, and it achieves good performance. It distinguishes things into linearly separable classes. It is one of the more popular models used in industry due to its simplicity. At the foundation of it, you get a value between 0 and 1 of whether or not an instance belongs to a particular class or not. A basic understanding of statistics and probability would suffice in understanding the model.

V. Experimental Results and Evaluation

In our evaluation, we used several metrics to measure the performance of our models. These metrics provide different insights into how well the predictions of our models are. These metrics include Accuracy, Precision, Recall, and F1 Score.

1) *Accuracy* measures the proportion of correct predictions out of the total number of predictions made.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{1}$$

Where TN is True Negative or the instances where the model correctly predicts the negative class (the patient has no heart disease), TP is True Positive or the instances where the model correctly predicts the positive class (the patient has heart disease), FP is False Positive or the instances where the model incorrectly predicts the positive class, and FN is False Negative which is when the model incorrectly predicts the negative class.

2) *Precision* measures the proportion of true positive predictions out of all predictions made by the model.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

3) *Recall*, which is also known as sensitivity or the true positive rate, measures the proportion of actual positive cases that were predicted by the model.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

4) *F1 Score* takes both *Precision* and *Recall* into account, and is useful when we need to balance the importance of the two. It is often a better overall measure than accuracy. For the purposes of this report, we will be focusing on the F1 Scores of each model since accuracy may not be the best metric, such as the case where a person with heart disease is classified as being healthy. An error like this would be detrimental in our case, as we are predicting heart disease which could be fatal without the proper intervention techniques.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

A. Logistic Regression

Logistic Regression uses the logistic function (sigmoid) to map the output of a linear regression model (z) to a probability value between 0 and 1, where $S(x)$ is the sigmoid function and e is Euler's number.

$$S(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

We implemented a Logistic Regression model with 1000 iterations and 80-20 train-test splits. This model shows an accuracy of 77.56%, a precision of 79.79%, a recall of 73.52%, and an F1 Score of 76.53%, which indicates that it is a decent model to use for prediction but not the best. The loss of accuracy can be impacted by noise and outliers in the dataset. For example, when we observed the Cholesterol and Max Heart Rate values, we noticed some outliers with large variances that could lead to overfitting. *Figure 13* shows the confusion matrix for our Logistic Regression model, which displays 27 instances of a False Negative, and 19 instances of a False Positive. In our case, both False Negative and False Positives are detrimental, as we are aiming to predict heart disease, which if untreated can lead to serious complications and even be fatal.

B. Bagging Decision Tree

Bagging Decision Trees, also known as Bootstrap Aggregating or Bagging, is a supervised classification algorithm where we combine the predictions of multiple decision trees to produce a more accurate model. The idea behind bagging decision trees is to improve stability and accuracy while reducing variance and mitigating overfitting. This is done by

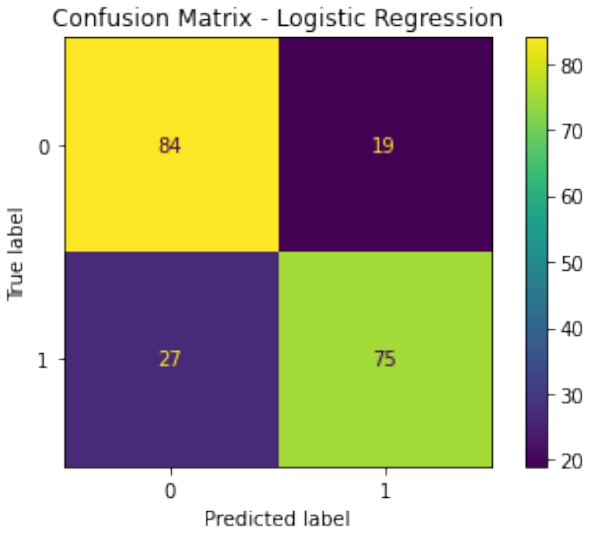

Fig. 13. Confusion Matrix for Logistic Regression. This model was less accurate than ensemble models (Bagging/Random Forest)

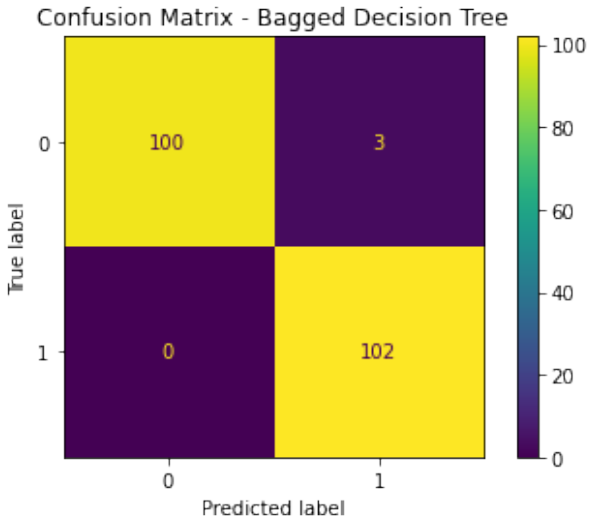

Fig. 14. Confusion Matrix for Bagged Decision Tree and Random Forest Tree. Compared to the Logistic Regression model, the Bagged Decision Tree model had less false negatives and was more accurate overall.

creating a certain number of bootstrap samples by sampling with replacement, training a decision tree on a specified number of sample observations, and using the majority vote of the resulting decision trees for classification.

We implemented a Bagging Decision Tree model with *DecisionTreeClassifier()* set as the base estimator of our model and *'n_estimators'* set to 50. This means that each model in our ensemble will be a decision tree, and we will create 50 decision trees in total, where each tree will be trained on a bootstrapped sample of the data [5].

Training and testing our model on our uncleaned dataset resulted in an Accuracy of 98.54%, Precision of 97.14%, Recall of 100%, and an F1 Score of 98.55%. *Figure 14* shows the confusion matrix for our Bagged Decision Tree Model. Compared to the Logistic Regression model, the Bagged Decision Tree model had less false negatives and was more accurate overall.

C. Random Forest

Random Forest is an algorithm similar to Bagged Decision Trees, but instead, creates decision trees where each is independently trained using a random subset of data and features. This randomness reduces overfitting by making each individual subtree unique, lowering the correlation between them. Each tree in the forest individually predicts a class label and the final overall prediction is determined by taking the average of the predictions of all the trees. Our model showed that Random Forest has an accuracy of 98.54%, a precision of 97.14%, a recall of 100%, and an F1 score of 98.55%. Random Forest Tree model scores were the same as Bagged Decision Tree model scores.

In Random Forest, feature importance is a measure of how much a feature contributes to the overall prediction of the model. *Figure 15* shows the visualization of feature importance scores, where the higher the score, the more important the feature. The importance of each feature is computed using Gini Impurity, which is used to determine how the features of a dataset should go about splitting nodes to form the tree [4]. Looking at the importance of each feature, we see that chest pain (cp) and the number of major blood vessels colored by fluoroscopy (ca) have the highest values, followed by maximum heart rate (thalach). This ranking is to be expected as chest pain and maximum heart rate achieved were the features most correlated with our target.

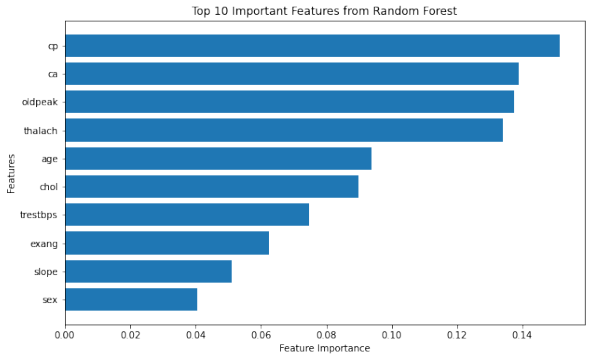

Fig. 15. Random Forest Feature Importance. Top 10 features were selected to retrain the models.

We experimented with feature selection after evaluating our Random Forest model. We took the top 10 features and reran the models on these features. The models after feature selection had no significant score improvements, as showcased in the figure below.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	77.56	79.79	73.53	76.53
1	Bagged Decision Tree	98.54	97.14	100.00	98.55
2	Random Forest	98.54	97.14	100.00	98.55
3	Logistic Regression with Selected Features	76.59	79.35	71.57	75.26
4	Bagged Decision Tree with Selected Features	97.07	97.06	97.06	97.06
5	Random Forest with Selected Features	98.54	97.14	100.00	98.55

Fig. 16. Results Before vs After Feature Selection. Feature Selection had minimal effect on model performance. Best model is Bagged Decision Tree.

D. Removing Outliers

After obtaining the model scores, we aimed to improve the model accuracy results. We removed 256 outliers from the data and reran each of the models. The result of this gave us perfect model scores for Bagged Decision Tree and Random Forest, as shown in the figure below:

```
Total outliers removed: 256
Bagged Decision Tree Metrics (Cleaned):
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0


Random Forest Metrics (Cleaned):
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0


Logistic Regression Metrics (Cleaned):
Accuracy: 0.8571428571428571
Precision: 0.8469387755102041
Recall: 0.9222222222222223
F1 Score: 0.8829787234042553
```

Fig. 17. Model Scores after Outlier Removal. Accuracies increased, however more overfitting occurred.

However, these high-accuracy results are due to overfitting. Removing too many outliers as shown above would make the Bagged Decision Tree and Random Forest models fit the training data too closely. This would make the models capture capturing all variations and noise, leading to overfitting (albeit giving higher accuracy scores). This overfitting is visualized by the perfect accuracy scores shown in *Figure 17*.

Likewise, perfect scores across the evaluation metrics (accuracy, precision, recall, and F1 score) indicate that the model is memorizing the entire training data, due to the training data becoming smaller with the removal of entries with outliers.

VI. Conclusion and Discussion

Overall, our research concludes various Machine Learning classification models resulting in both the Bagged Decision Tree and Random Forest models outperforming the Logistic Regression model in accurately predicting the presence of heart disease. Specifically, the **Bagged Decision Tree model** achieved the highest performance metrics with an **Accuracy of 98.54%, Precision of 97.14%, Recall of 100%, and an F1 Score of 98.55%**. This superior performance can be attributed to the ensemble technique of the Bagged Decision Tree model, which combines multiple decision trees to reduce variance and prevent overfitting. Meanwhile, the Random Forest model also demonstrated competitive performance with an accuracy of 98.54%, a precision of 97.14%, a recall of 100%, and an F1 score of 98.55%. The Random Forest and Bagged Decision Tree produced the exact same results. Although feature importance was used to retrain models, overall model accuracies didn't increase because the amount of features in the dataset isn't large enough. Both models' effectiveness highlights the importance of using ensemble methods to handle the complexities and variances in medical datasets. The Logistic Regression model showed the least accurate result and lower performance metrics. With an accuracy of 77.56%, precision of 79.78%, a recall of 73.53%, and an F1 Score of 76.53%. The Logistic Regression model was less effective in predicting heart disease, particularly in dealing with the outliers and high variance in relationships within the dataset.

References

[1] "Heart Disease Facts," Centers for Disease Control and Prevention, https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html#:~:text=Heart%20disease%20in%20the%20United%20States&text=One%20person%20https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html#:~:text=Heart%20disease%20in%20the%20United%20States&text=One%20person%20dies%20every%2033,1%20in%20every%205%20deaths.dies%20every%2033,1%20in%20every%205%20deaths.

[2] American Heart Association, (2018), https://www.heart.org/en/impact-map

[3] J. L. Rodgers et al., "Cardiovascular Risks Associated with Gender and Aging," Journal of Cardiovascular Development and Disease, vol. 6, no. 2. MDPI AG, p. 19, Apr. 27, 2019. doi: 10.3390/jcdd6020019.

[4] F. Karabiber, "Gini Impurity," Learn Data Science, https://www.learndatasci.com/glossary/gini-impurity/.

[5] "BaggingClassifier," scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html.

[6] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN computer science, vol. 1, no. 6, pp. 345-, 2020.

[7] Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 9, 1-16. https://doi.org/10.4236/jilsa.2017.91001.

[8] Brandon S, (May 2021) Investigating Heart Disease Datasets and Building Predictive Models, Elizabeth City State University, 1-69. https://libres.uncg.edu/ir/ecsu/f/Brandon_Simmons_Thesis-Final.pdf