The problem we aim to address with this project is to create a ML model for the early detection of heart disease using machine learning techniques. Heart disease is a widespread health issue affecting a significant portion of the population globally. It is important to understand the demographics affected, leading causes, diagnostic methods, and treatments to fight this condition effectively.

Before, heart disease diagnosis depended on clinical assessments, medical history and diagnostic tests such as electrocardiograms and blood tests. Now this approach may have limitations in early detection and may not use available data for accurate predictions.

Solving this problem with machine learning is really important so that it can provide more accurate and timely predictions, enabling early intervention and personalized treatment plans. This particular solution will be helping healthcare professionals for patient management and improving total patient outcomes. Our team's approach is making a machine learning model which is getting trained on the patient data to predict the likelihood of heart disease. Here, our team's contribution to this project will be the development of an accurate predictive model that can help healthcare professionals in making decisions and improving the care of patients.

Then, the summary of our approach involves using some machine learning techniques to make a predictive model for early detection of any heart disease. We will preprocess the dataset, and perform feature scaling as necessary. Then, we will explore various machine learning algorithms such as logistic regression, k-nearest neighbors, perceptron and random forest to train and evaluate models. We will first try to make a model using default parameters and see the accuracy that we attained with it. Then select hyperparameters for each and every model using cross-validation. Now, using RandomSearchCV and GridSearchCV, best hyperparameters are to be found which shows better accuracy. These trained models are applied to the test set. Our major contribution will be the development of an accurate(nearly) predictive model that can help healthcare professionals check individuals at risk of heart disease.

Data Exploration

The dataset used for this project is the Heart Disease Prediction dataset, collected from Kaggle in 1988 and comprising four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, with experiments focusing on a subset of 14 attributes. The target variable indicates the presence (1) or absence (0) of heart disease, that basically means if the patient is diagnosed or not.

The dataset includes some features such as age, sex, chest pain type, blood pressure, cholesterol levels, and electrocardiographic results, among others. Interestingly, the dataset's attributes cover various aspects of cardiovascular health, providing information for predictive modeling.

For the classification task, the distribution of classes (presence and absence of heart disease) in the dataset needs to be examined to ensure balance and watch any of the class imbalance issues. Additionally, understanding the dataset's size and demographics represented is crucial for training our model and evaluation.

The distribution of classes in the dataset for the heart disease prediction classification task is as follows:

- Class 0 (No disease): 499 records
- Class 1 (Disease): 526 records

This indicates a relatively balanced distribution between the two classes, which is beneficial for training machine learning models without the risk of significant class imbalance issues.

The dataset contains 1025 records, each of them represents an individual's health data. Features in the dataset include age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise, the slope of the peak exercise ST segment, total number of major vessels colored by fluoroscopy, and thalassemia.

One interesting aspect of the dataset that we found is its origin from 1988, comprising four databases from different locations: Cleveland, Hungary, Switzerland, and Long Beach V. Despite its age, this dataset remains relevant for heart disease prediction research, indicating the much importance of early detection and intervention.

Metrics for Model Evaluation:

For classification tasks:

- We evaluated our model using Recall (if we missed heart disease it would have severe consequences) and Precision. False Negatives should be minimum. We ve also taken out F1 score and Accuracy. But Recall is most crucial for heart disease prediction.

Baseline:

For classification tasks: 526 target variables are positive.

- 51% is in the positive class.

Methodology:

- Data Splitting: We split the data into training and test sets using the train_test_split function from sklearn.model_selection. 20% is allocated in test size and the remaining 80% is allocated in 80%.
- Cross-Validation: Yes, we used K-fold cross-validation with 5 folds (cv=5) during hyperparameter tuning.
- Hyperparameter Tuning: We tuned the hyperparameters of our models using grid search (GridSearchCV) or randomized search (RandomizedSearchCV) techniques.

- Machine Learning Algorithms: We tried Logistic Regression, Random Forest, KNN (K-Nearest Neighbors), and Perceptron.
- Best Hyperparameters: We determined the best hyperparameters for each model based on their performance metrics.

Results:

The best-performing machine learning algorithms on the test set and their respective performance metrics (precision, recall, and F1 score) compared to the baseline are summarized in the table below:

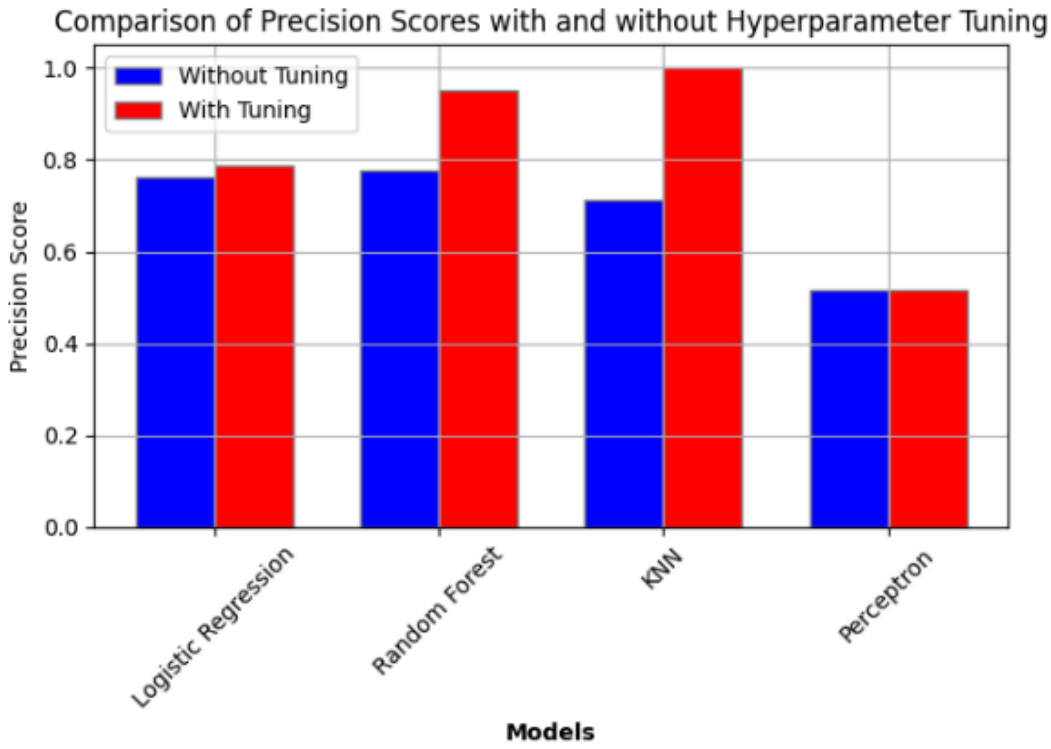
	Model	Precision	Recall	F1 Score
0	Logistic Regression	0.785714	0.942857	0.857143
1	Random Forest	0.952830	0.961905	0.957346
2	KNN	1.000000	0.971429	0.985507
3	Perceptron	0.517241	1.000000	0.681818

KNN has the highest precision,recall and F1-score from all the three tested above so it is best for us.

The combination of metrics such as Precision, Recall and F1 Score is used to determine the effectiveness of the methodology employed for heart disease detection in Patients. We are taking the combination of these metrics instead of just one such that meaningful interpretations are derived easily from looking at the overall results

Discussion:

- Hyperparameter tuning improves the performance of our models by optimizing them for the given task.



- Experimenting with different algorithms and tuning their hyperparameters helped in identifying the most suitable model for the task.

Conclusion:

- Through this project, we learned the importance of thorough experimentation, cross validation and hyperparameter tuning in building different effective machine learning models.
- In the future, we aim to further explore advanced techniques like the XGBoost algorithm to improve model performance and provide interface for practical usability.

Literature Review.

Title: Heart Disease Prediction using Hybrid Machine Learning Model

Authors: M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, R. Sai Suraj

https://ieeexplore.ieee.org/abstract/document/9358597

Summary:

The authors propose an effective machine learning approach to predict heart disease in this paper. Cleveland heart disease dataset is used for the model. The authors resorted to regression and classification techniques, further implementing the Random Forest, Decision Tree, and a Hybrid model. The results at the end of this study showed an accuracy of 88.7% with the hybrid model (Kavitha et al., 2021). It is evident from the results that the hybrid model is outperforming individual algorithms. The interface which is user friendly and convenient provides great application in practical usability, showcasing the potential of machine learning in early detection of heart disease (M.Kavitha et al., 2021).

Title: Heart Disease Prediction Using Machine Learning

Authors: Chaimaa Boukhatem, Heba Yahia Youssef, Ali Bou Nassif

https://ieeexplore.ieee.org/abstract/document/9734880

Summary:

The authors in this paper explore various machine learning approaches for predicting cardiovascular disease based on electronic health data. Four classification methods were used for building out the model by authors in this paper namely Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB). The authors have resorted to pre-processing the electronic health data before modeling to enhance the ease of implementing Machine learning methods. Feature selection is conducted before model building. The Support Vector Machine (SVM) model in the paper achieved the highest accuracy of 91.67%, demonstrating its effectiveness in heart disease prediction over Multilayer Perceptron (MLP), Random Forest (RF) and Naïve Bayes (NB) (C. Boukhatem et al., 2022).

Title: Heart Disease Prediction using Machine Learning Techniques

Authors: Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta

https://ieeexplore.ieee.org/abstract/document/9362842

Summary:

The authors in this paper use machine learning techniques to predict heart-related diseases early in patients. The dataset used for the research is taken from UCI which consists of multiple machine learning algorithms such as Random Forest, Support Vector Machine (SVM), Naive Bayes, and Decision Tree.The authors have derived various metrics such as accuracy, precision and many other parameters to compare the significance of all the machine learning algorithms. The results indicated that the Random Forest algorithm produced accurate and better predictions than the other machine learning models. Authors also highlight the importance of such models in the context of global healthcare.

Title: Heart Disease Prediction using Machine Learning Techniques

Authors: Devansh Shah, Samir Patel, Santosh Kumar Bharti

https://link.springer.com/article/10.1007/s42979-020-00365-y

Summary:

The paper aims to predict heart disease in patients using machine learning techniques. The authors use the Cleveland database from the UCI repository in this research. The data consists of 303 records with 76 features, out of which the authors select 14 important features for testing. The authors resort to using Naïve Bayes, Decision Tree, K-nearest neighbor, and Random Forest algorithms for model training and prediction. The results of the K-nearest neighbor model seems promising when compared with all others.

We have reviewed a huge amount of existing literature on the research conducted by various academic professionals post-2013 using machine learning for heart disease prediction. These studies outlined various methodologies such as logistic regression, decision trees, support vector machines, and neural networks. It is evident that a good number of research projects have achieved promising results in terms of accuracy and predictive performance.

However, one aspect not adequately addressed in these papers is the effect of integrating advanced feature selection techniques to improve model interpretability and generalization. This aspect is about using advanced methods to choose which information is most important for the predictions. This selection process can make the models easier to understand and more accurate in predicting who might get heart disease.