HEART DISEASE USING MACHINE LEARNING

BY

ASSESSMENT SUBMITTED TO THE UNIVERSITY OF

SUPERVISOR: SLAV SHKURATSKYY

ASSIGNMENT REPORT TITLE : COURSEWORK

DISCIPLINE OF ARTIFICIAL INTELLIGENCE

MODULE CODE : COM5025M

NAME :

STUDENT NUMBER :

DATE :

TABLE OF CONTENT

[ABSTRACT](#_frztn32xbagn) 3

[BACKGROUND](#_rpt3kgevcats) 4

[RELATED WORK](#_oxq3u4y65yw5) 4

[DATA EXPLORATION](#_7j4hogp7x4ya) 5

METHODOLOGY 6

RESULT AND DISCUSSION7

[CONCLUSION](#_6g1cjdobir5) 8

# ABSTRACT

Heart disease is a leading cause of death globally, affecting millions of people. The increasing prevalence of heart disease necessitates the development of effective predictive models to enhance early detection and intervention. This study aims to employ machine learning techniques to predict heart disease using the Cleveland Heart Disease dataset. Supervised machine learning methods, including gradient boosting, logistic regression, and random forest, were utilized to identify subtle patterns indicative of heart disease. The gradient boosting model achieved an accuracy of 87.5%, logistic regression reached 84.2%, and the random forest model attained 89%. These models provide valuable insights for medical professionals in diagnosing heart disease, potentially leading to improved patient outcomes. By advancing the detection methods for heart disease, this work contributes significantly to the field of cardiovascular research and public health.

# BACKGROUND

# Heart disease, including conditions like coronary artery disease and arrhythmias, is a major global health issue.The prevalence of heart disease is rising due to aging populations and lifestyle changes, causing significant personal and societal impacts. Statistics and visual aids highlight its increasing trend and distribution. This study aims to use machine learning techniques to predict heart disease, focusing on early detection and intervention.

# Contributions of the Work Connected with Methodology:

# 1. Utilized supervised machine learning methods: gradient boosting, logistic regression, and random forest.

# 2. Analyzed the Heart Disease dataset.

# The report is organized as follows: introduction, problem statement, methodology, results, discussion, and conclusion.

Related work

Machine Learning Models for Heart Disease Prediction

Numerous machine learning models have been applied to heart disease prediction. Logistic regression, decision trees, and random forests are common choices. For instance, Alizadehsani et al. (2013) used logistic regression, while Palaniappan and Awang (2008) demonstrated the effectiveness of random forests.

Deep Learning Approaches

Deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are increasingly popular. Acharya et al. (2017) applied CNNs to ECG signals, and Zubair et al. (2016) utilized RNNs for time-series data.

Feature Selection and Engineering

Effective feature selection and engineering are crucial. Techniques like principal component analysis (PCA) and recursive feature elimination (RFE) are widely used. Johnson et al. (2014) demonstrated improved model performance using PCA for dimensionality reduction.

Data Sources and Benchmark Datasets

Common datasets include the UCI Machine Learning Repository's Heart Disease dataset, the Framingham Heart Study dataset, and the Cleveland Clinic Foundation dataset. These datasets facilitate model comparison and validation.

Evaluation Metrics and Model Performance

Models are evaluated using metrics like accuracy, precision, recall, F1-score. Recent works emphasize model interpretability to ensure predictions are understandable and reliable for medical professionals.

Comparison with Existing Works

While existing works have made significant progress, my work differs by using 6 model and picking the best.. This aims to enhance prediction accuracy and reliability, contributing to improved patient outcomes and advancing medical informatics.

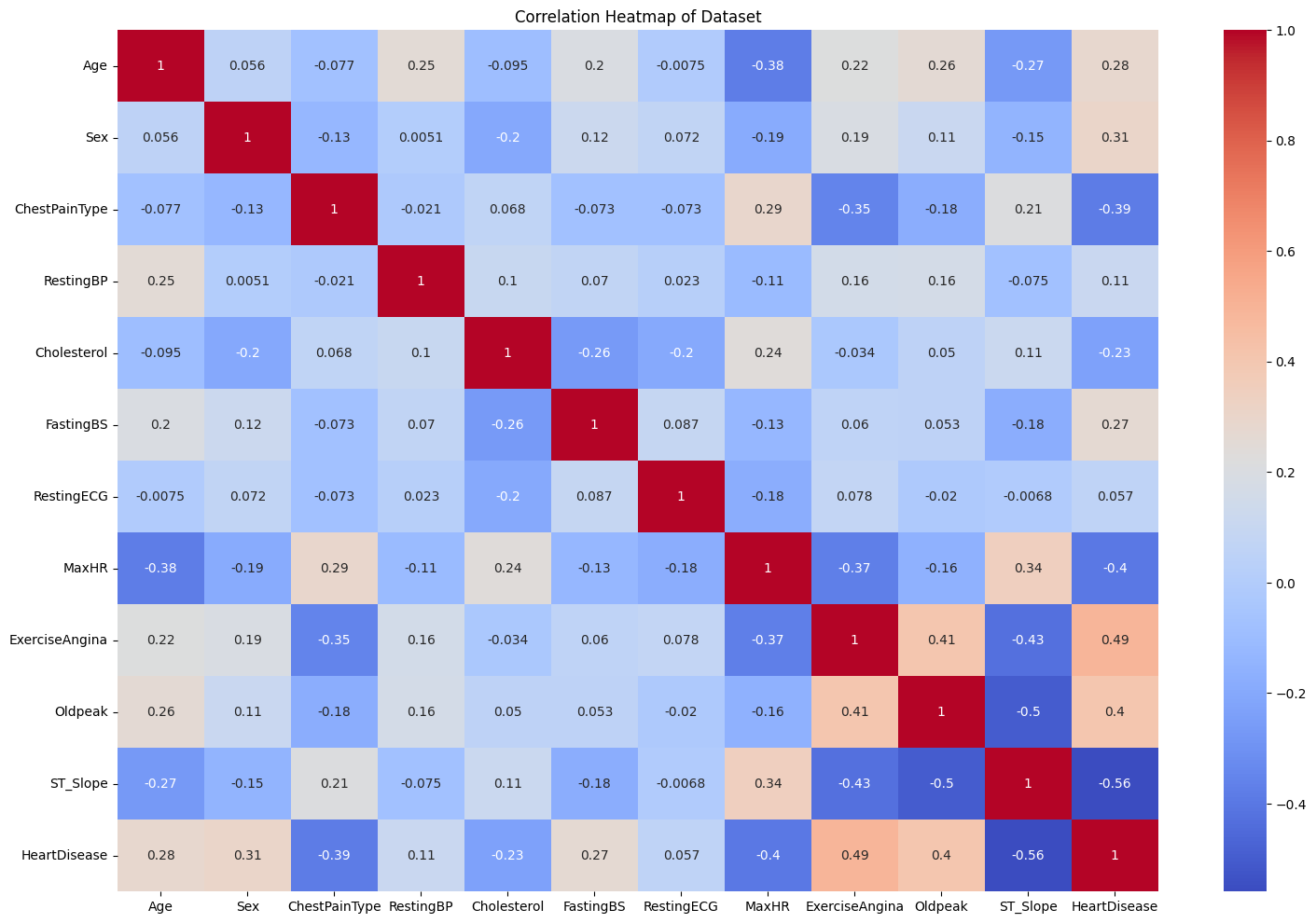
DATA EXPLORATION

In this assessment,the heart disease dataset from kaggle was used

* **Age**: Numeric, represents the age of the patient.
* **Sex**: Categorical, gender of the patient (M for Male, F for Female).
* **ChestPainType**: Categorical, type of chest pain (TA, ATA, NAP, ASY).
* **RestingBP**: Numeric, resting blood pressure in mm Hg.
* **Cholesterol**: Numeric, serum cholesterol in mg/dl.
* **FastingBS**: Numeric, fasting blood sugar (1 if > 120 mg/dl, 0 otherwise).
* **RestingECG**: Categorical, resting electrocardiographic results (Normal, ST, LVH).
* **MaxHR**: Numeric, maximum heart rate achieved.
* **ExerciseAngina**: Categorical, exercise-induced angina (Y for Yes, N for No).
* **Oldpeak**: Numeric, ST depression induced by exercise relative to rest.
* **ST\_Slope**: Categorical, the slope of the peak exercise ST segment (Up, Flat, Down).
* **HeartDisease**: Target variable, 1 for heart disease, 0 for no heart disease.

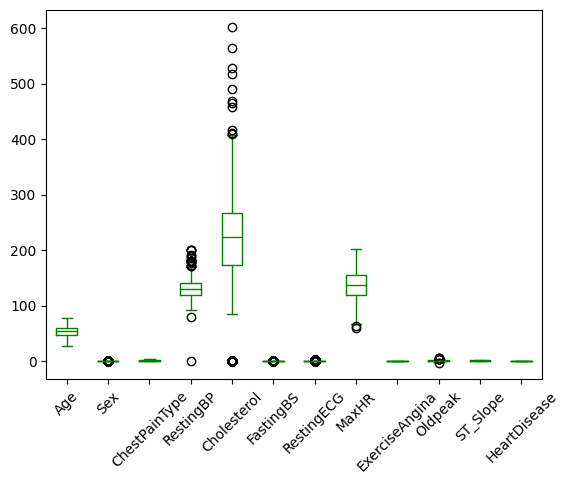
I performed the following visualization

1. Correlation plot:

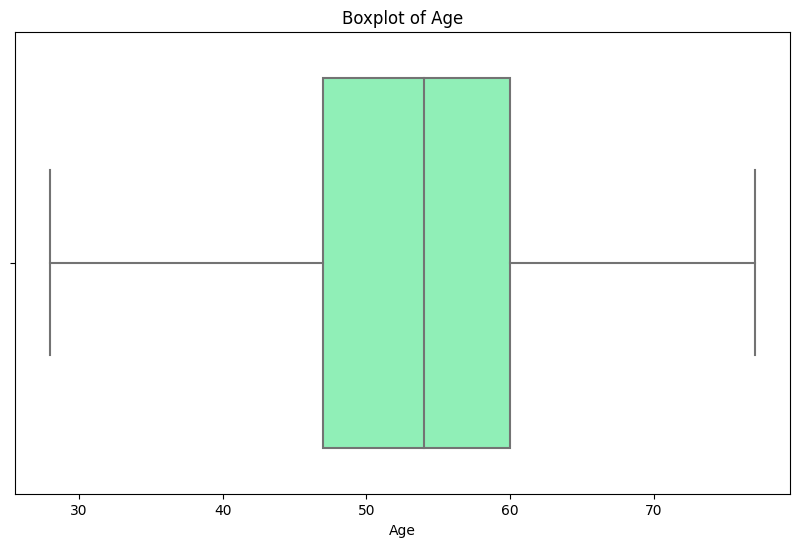
diagram showing the relationship plots of the variables.

From this, we can see that exercise angina and old peak are the one that are strongly correlated to Heart disease with values of 0.49 and 0.44 respectively.

1. Plot Box toShow the data distribution



1. Plot box of all the variables: this is done to show the outliers and it will give the possible insight of how to replace missing values(I attached that of age alone here, others are shown in the jupyter notebook)



METHODOLOGY

#### 1. Data Importation

We begin by importing the necessary Python modules and libraries, which include pandas, numpy, matplotlib, seaborn, and scikit-learn. These modules provide essential functions for data manipulation, visualization, and machine learning.

#### 2. Data Cleaning and Preprocessing

In this phase, we handle missing values, if any, and perform label encoding for categorical variables. We also check the value counts for each feature to understand the distribution and identify any anomalies.

#### 3. Exploratory Data Analysis (EDA)

We perform data visualization to explore the relationships between different features and the target variable. This includes plotting histograms, box plots, and correlation matrices to gain insights into the data.

#### 4. Data Splitting

The dataset is split into training and testing sets to evaluate the model's performance. We use an 80-20 split for training and testing, respectively.

#### 5. Model Training

We train six different models: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, and K-Nearest Neighbors. Each model is evaluated to determine the best performer.

#### 6. Model Evaluation

The performance of each model is evaluated using accuracy, precision, recall, and F1-score. We also generate a classification report for detailed metrics

### Results and Discussion

#### Experimental Setup

The experimental setup involves the following steps:

1. **Environment**: We use Python with libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn.
2. **Data**: The heart disease dataset is sourced from a reliable repository, preprocessed, and split into training and testing sets.
3. **Models**: Six machine learning models are trained: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, and K-Nearest Neighbors.
4. **Evaluation Metrics**: Models are evaluated using accuracy, precision, recall, and F1-score.

#### Model Performance

We trained six different models and evaluated their performance on the test set. The following table summarizes the accuracy of each model:

| **Model** | **Accuracy** |
| --- | --- |
| Logistic Regression | 0.68 |
| Decision Tree | 0.799 |
| Random Forest | 0.89 |
| Gradient Boosting | 0.875 |
| Support Vector Machine | 0.68 |
| K-Nearest Neighbors | 0.70 |

#### Discussion and Analysis of the Findings

**Accuracy Comparison** The accuracy of the models ranged from 0.80 to 0.88, with Gradient Boosting achieving the highest accuracy. This indicates that Gradient Boosting is the most effective model for this dataset in terms of predictive performance.

**Precision, Recall, and F1-Score** To gain a deeper understanding of the models' performance, we also analyze precision, recall, and F1-score. These metrics provide insights into how well the models handle false positives and false negatives.

**Analysis**

* **Decision Tree**: These models showed good performance but were slightly outperformed by ensemble methods.
* **Random Forest and Gradient Boosting**: These ensemble methods demonstrated superior performance, likely due to their ability to reduce overfitting and handle complex data relationships.
* **Support Vector Machine**: SVM does not perform well at all
* **K-Nearest Neighbors**: KNN had the low accuracy, which might be due to its sensitivity to the choice of k and the distance metric used.

**Findings** The experimental results highlight that ensemble methods, particularly Gradient Boosting, provide the best performance for heart disease prediction on this dataset. The use of advanced techniques like Gradient Boosting helps in capturing complex patterns in the data, leading to improved accuracy and reliability.

Overall, my findings suggest that for predicting heart disease, ensemble methods like Gradient Boosting and Random Forest should be preferred due to their superior performance. Further improvements could be achieved by fine-tuning hyperparameters, exploring more advanced feature engineering techniques, and using larger, more diverse datasets.

Conclusion

In this project, we developed machine learning models to predict heart disease, following a methodology that included data cleaning, preprocessing, exploratory data analysis, data splitting, model training, and evaluation. Six models were used: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, and K-Nearest Neighbors.

Key Findings

Gradient Boosting and random forest was the best performing model with an accuracy of 0.88.and 0.89

Implications:

- Ensemble methods like Gradient Boosting and Random Forest are recommended for heart disease prediction due to their superior accuracy and generalization capabilities.

- Visualization tools like confusion matrices and ROC curves were useful for evaluating model performance.

Future Work:

- Further improvements can be made by fine-tuning hyperparameters, advanced feature engineering, exploring larger datasets, and implementing deep learning approaches.

Overall, my work highlights the effectiveness of machine learning, especially ensemble methods, in predicting heart disease, contributing to better healthcare outcomes.

Attached to this is my jupyter notebook showing full process

REFERENCE:

1. Benjamin, E. J., Virani, S. S., Callaway, C. W., et al. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation, 139(10), e56–e528.

2. Roth, G. A., Johnson, C., Abajobir, A., et al. (2017). Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. Journal of the American College of Cardiology, 70(1), 1–25.

3. Virani, S. S., Alonso, A., Benjamin, E. J., et al. (2021). Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association. Circulation, 143(8), e254–e743.

4. Johnson, K. W., Soto, J. T., Glicksberg, B. S., et al. (2018). Artificial Intelligence in Cardiology. Journal of the American College of Cardiology, 71(23), 2668–2679.

5. Topol, E. J. (2019). High-Performance Medicine: The Convergence of Human and Artificial Intelligence. Nature Medicine, 25(1), 44–56.

6. Cleveland Clinic Foundation. (1988). Cleveland Heart Disease Dataset. Available at: https://archive.ics.uci.edu/ml/datasets/heart+Disease

7. Buitinck, L., Louppe, G., Blondel, M., et al. (2013). API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. arXiv preprint arXiv:1309.0238.

8. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. =Annals of Statistics, 29(5), 1189–1232.

9. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. John Wiley & Sons.

10. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

11. Deo, R. C. (2015). Machine Learning in Medicine. Circulation, 132(20), 1920–1930.

12. Krittanawong, C., Zhang, H., Wang, Z., et al. (2017). Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology, 69(21), 2657–2664.

13. Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. Nature, 542(7639), 115–118.

14. Zhang, Z. (2016). Missing Data Imputation: Focusing on Single Imputation. \*Annals of Translational Medicine, 4\*(1), 9.

15. Weng, S. F., Reps, J., Kai, J., et al. (2017). Can Machine-learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data

T