**Name:Somesh Nakka**

**Student Id: 23006089**

**Project Title:**

Developing a Machine Learning Framework for Early Prediction of Heart Disease: A Comparative Study of Models and Feature Importance

**Summary and Background**

Heart disease is one of the top causes of death worldwide, accounting for a large number of deaths each year. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) account for roughly 32% of all global fatalities, making early detection and prevention an important part of public health. Timely intervention, whether through lifestyle modifications, drugs, or medical procedures, has the potential to drastically cut mortality rates and improve patient outcomes.In recent years, machine learning (ML) and artificial intelligence (AI) have emerged as effective techniques for predicting cardiac disease by recognizing patterns in patient demographics, clinical data, and lifestyle factors. Traditional diagnostic methods, such as electrocardiograms (ECG), stress tests, and blood tests, while beneficial, typically need specialised knowledge, expensive equipment, and lengthy wait times for results.

**Research Question**

How effectively can machine learning algorithms predict heart disease based on patient demographic and clinical data, and which features and models contribute most significantly to accurate predictions?

**Project Objectives**

1. **Data Preprocessing and Feature Engineering: EDA** is performed on the data for data quality verification, treatment of missing values, normalizing numeric features, and encoding of categorical variables.
2. **Dataset Suitability:** Ensure the dataset that will be used is balanced and representative. In cases of imbalanced datasets, any class imbalance has to be preprocessed accordingly through techniques like SMOTE or undersampling.
3. **Hyperparameter Tuning:** The aim is to avoid overfitting and enhance the performance by tuning the models using Grid Search or Randomized Search.
4. **Model Performance Evaluation:** The performance of the models is going to be tested for accuracy, precision, recall, F1-score, AUC-ROC, and further analyzed using the confusion matrix.
5. **Feature Importance Analysis:** SHAP, permutation importance, and model-based techniques shall be employed to understand significant features of heart disease prediction.

**References:**

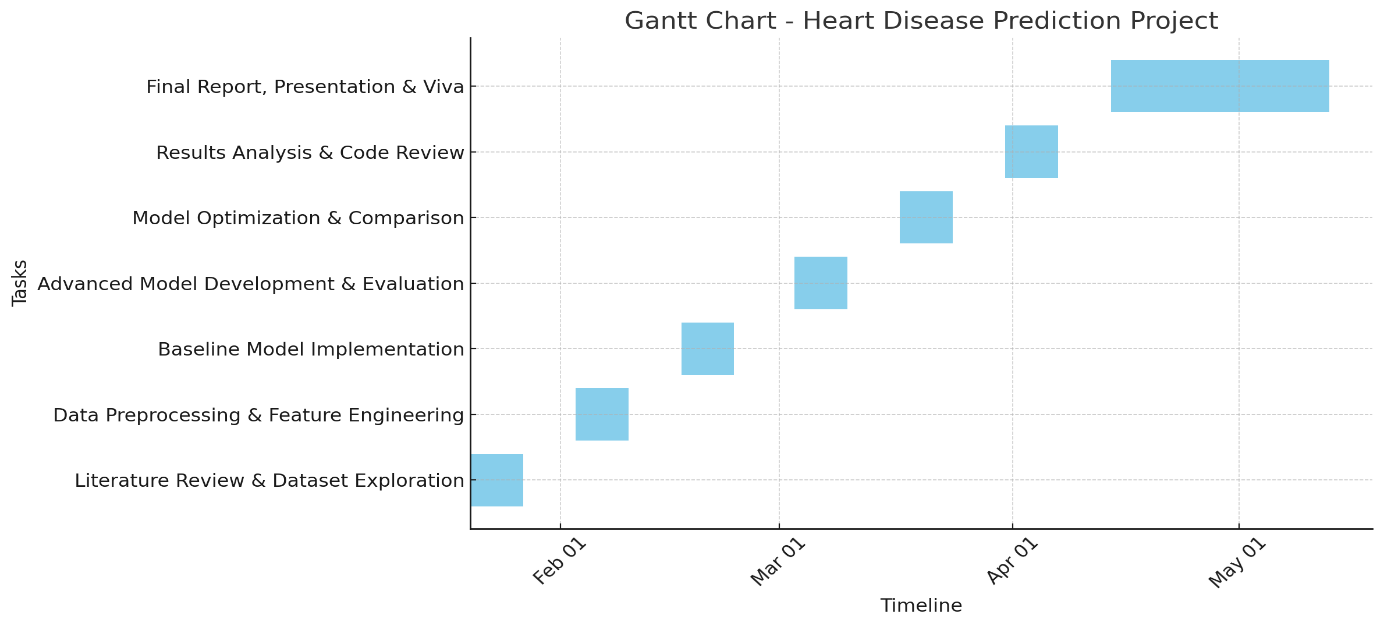
1. M. R. Ahmed, S. M. Hasan Mahmud, M. A. Hossin, H. Jahan and S. R. Haider Noori, "A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms," *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2018, pp. 1951-1955, doi: 10.1109/CompComm.2018.8781022. URL: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8781022&isnumber=8780575>
2. V. M B, S. S. R, K. U and K. Y, "Exploratory Data Analysis of Heart Disease Prediction using Machine Learning Techniques-RS Algorithm," *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India, 2024, pp. 209-216, doi: 10.1109/ICoICI62503.2024.10696414. URL: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10696414&isnumber=10695948>

A. Lakshmi and R. Devi, "Heart Disease Prediction Using Enhanced Whale Optimization Algorithm Based Feature Selection With Machine Learning Techniques," *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2023, pp. 644-648, doi: 10.1109/SMART59791.2023.10428617. URL: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10428617&isnumber=10428128>

**2. Project Plan: Task List and Timeline**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weeks | Start Date | End Date | Task | Task Description |
| 1-2 | 20-Jan | 27-Jan | Literature Review & Dataset Exploration | Literature Review of Papers on Heart Disease Prediction Models and Dataset Overview |
| 3-4 | 03-Feb | 10-Feb | Data Preprocessing & Feature Engineering | Cleaning, Missing Value Handling, Normalization, and Encoding Categorical Variables |
| 5-6 | 17-Feb | 24-Feb | Baseline Model Implementation | Build Baseline Machine Learning Models for : Logistic Regression, Decision Trees, Random Forest |
| 7-8 | 03-Mar | 10-Mar | Advanced Model Development & Evaluation | Model Development and Evaluation: SVM, Neural Networks, Ensemble Learning |
| 9-10 | 17-Mar | 24-Mar | Model Optimization & Comparison | Feature Importance Analysis for Hyperparameter Optimization to Model Improvement |
| 11-12 | 31-Mar | 07-Apr | Results Analysis & Code Review | Analyzing Model Performance and validate code quality |
| 13-15 | 14-Apr | 13-may | Final Report, Presentation & viva |  |
| Prepare the final report draft and presentation slides for the project submission. |

**Timeline (Gantt Chart Representation)**



**3. Data Management Plan**

**Overview of the Dataset**

The dataset utilized in this work is the Heart Disease UCI Dataset on Kaggle. It contains demographic and clinical variables of patients, primarily for predicting the probability of heart disease. In this dataset, a wide range of patient attributes is represented, including age, gender, blood pressure, cholesterol, fasting blood sugar, ECG, maximum heart rate achieved, and types of chest discomfort. These parameters have been widely researched in medical studies as vital pointers of cardiovascular health and are extremely useful in designing predictive models for diagnosing heart disease.

**Dataset Link:** <https://www.kaggle.com/datasets/shantanugarg274/heart-prediction-dataset>

**Data Collection**

The dataset was acquired via Kaggle and is accessible for educational and research purposes. The compilation was derived from many clinical investigations concentrating on cardiovascular health and the identification of risk factors linked to heart disease. The dataset comprises both categorical and numerical variables, essential for conducting predictive analytics and constructing machine learning models.

**Metadata**

* **Dataset Source:** Kaggle (Heart Disease UCI Dataset)
* **Format:** CSV file ( 1000 samples)
* **Size:**40KB
* **Features:** Age, gender, cholesterol, resting blood pressure, ECG results, etc.
* **Handling Missing Data:** Imputation techniques.
* **Ethical Considerations:** GDPR compliance.

**Version Control and Code Administration**

All project advancements will be monitored effectively by GitHub via methodical weekly commits.

**Github Link:**  
**ReadMe File Declaration**A README file will delineate dataset preparation, dependencies, and model specifications. This facilitates effortless replication for subsequent users.

**Security and Safeguard**

Data and code will be safely stored on GitHub and OneDrive. Moreover, a weekly backup is made to prevent accidental loss of data.

**Ethical Considerations**

The dataset has followed the privacy regulations of the GDPR. Research will follow the ethical norms of research.

**Conclusion**

This experiment shows how machine learning can predict patient outcomes using healthcare data. By comparing models, we discovered the most successful forecasting method and important influential features. Our findings emphasize the necessity of ethical data management, GDPR compliance, and reproducibility.