**Phase-1 Submission**

**Student Name:** DINESH KUMAR N

**Register Number:** 712523106004

**Institution:** PPG INSTITUTE OF TECHNOLOGY

**Department:** ELECTRONICS AND COMMUNICATION ENGINEERING

**Date of Submission:** [26/04/2025]

**1. Problem Statement**

*The goal of this study is to estimate the total risk of heart disease using Logistic Regression and Machine Learning. Heart attacks account for about four out of every five CVD-related deaths. Identifying those at risk of CVD is critical for implementing preventative / strategies and giving timely therapies.*

**2. Objectives of the Project**

* *Design and train a machine learning model to predict the likelihood of heart disease based on various risk factors.*
* *Analyze and identify the most significant risk factors contributing to heart disease.*
* *Enable early detection of heart disease risk, allowing for preventive measures.*
* *Create informative visualizations to communicate findings and insights.*

**3. Scope of the Project**

* *The project assumes that the developed system will be adopted and utilized by healthcare professionals and patients.*
* *Extracting and creating relevant features from the dataset to improve model performance.*
* *The project assumes that the developed model will achieve a satisfactory level of accuracy.*

**4. Data Sources**

* *National Health and Nutrition Examination Survey (NHANES).*
* *Contains data on Medicare beneficiaries, including those with heart disease.*
* *A database containing data on patients admitted to intensive care units.*
* *Electronic health records (EHRs) from Stanford Health Care.*

**5. High-Level Methodology**

* **Data Collection** *– Download publicly available datasets (e.g., Kaggle, UCI) in CSV format containing patient health features and heart disease status.*
* **Data Cleaning** *– Address missing values (impute or remove), duplicates (remove), and inconsistent formats (standardize, encode categoricals). Handle outliers as needed*.
* **Exploratory Data Analysis (EDA)** *– Use summary statistics and visualizations (histograms, box plots, scatter plots, correlation heatmaps) to understand data distributions, relationships between features, and patterns related to heart disease.*
* **Feature Engineering** – *Explore creating interaction terms, binning numerical features, applying transformations, and scaling features to potentially improve model performance*.
* **Model Building** – *Experiment with classification algorithms like Logistic Regression, SVM, Decision Trees, Random Forests, Gradient Boosting Machines (e.g., XGBoost), KNN, and potentially Neural Networks, chosen for their suitability for binary classification and ability to capture different data patterns.*
* **Model Evaluation** – *Measure model performance using metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrices. Employ train-test split and cross-validation for robust evaluation.*
* **Visualization & Interpretation** – *Present key findings and predictions using charts (e.g., performance comparisons, ROC curves, feature importance plots) and potentially a summary dashboard to communicate insights about heart disease risk factors and model performance.*
* **Deployment** – *Depending on the project goals, consider deploying as a web application (e.g., using Flask or Django), a dashboard (e.g., Tableau, Power BI, Streamlit), or as a shareable notebook or script.*

**6. Tools and Technologies**

* **Programming Language** – *Python*
* **Notebook/IDE** – *Jupyter Notebook*
* **Libraries** – *pandas, NumPy, Matplotlib, Seaborn, scikit-learn (potential TensorFlow/Keras for advanced modeling)*
* **Optional Tools for Deployment** – *Streamlit, Flask, Gradio, FastAPI (depending on needs)*

**7. Team Members and Roles**

|  |  |  |
| --- | --- | --- |
| **MEMBERS** | **ROLE** | **DESCRIPTION** |
| HARI A | *Data**Collection**&**Preprocessing* | *Responsible for acquiring the dataset, cleaning the data, and preparing it for analysis* |
| DINESH KUMAR N | *Exploratory Data Analysis & Feature Engineering* | *In charge of performing EDA, visualizing insights, and engineering features to enhance model performance* |
| *JONEYABIRAHAM Y* | *Model Building & Evaluation* | *Develops and trains the deep learning model (CNN), tunes hyperparameters, and evaluates performance metrics.* |
| *MARI RAMESH M* | *Visualization & Deployment* | *Handles interpretation of results, builds the user interface (if applicable), and manages optional deployment using tools like Streamlit or Gradio.* |