Prediction of Cardiovascular Disease using Machine Learning Algorithm

PROJECT PROPOSAL

UNIVERSITY OF NORTH TEXAS

DEPARTMENT OF COMPUTER SCIENCE

SOFTWARE DEVEPMENT FOR ARTIFICIAL INTELLIGENCE

Prediction of Cardiovascular Disease using Machine Learning Algorithm

by

Gopireddy Lakshmi Keerthi (Student ID: 11653876)
LakshmiKeerthiGopireddy@my.unt.edu
Major: Artificial Intelligence
Role: Python programming, Project Documentation,
Domain Understanding & Feature Engineering,
Exploratory Data Analysis, Model Training, and Model Evaluation

Indhu Sri Krishnaraj (Student ID: 11661746)
IndhuSriKrishnaraj@my.unt.edu
Major: Artificial Intelligence
Role: Python programming, Data Visualization,
Exploratory Data Analysis, Data Pre-processing,
Parameter tuning and Model Training

Vishal Rachuri (Student ID: 11658243)
vishalrachuri@my.unt.edu
Major: Artificial Intelligence
Role: Python programming,
Dataset Selection, Parameter tuning and
Model Prediction and Evaluation

Jyothika Vollireddy (Student ID:)
jyothikavollireddy@my.unt.edu
Major: Cyber Security
Role: Python programming, Data Visualization,
Data Pre-processing, Feature Engineering,
Algorithm Selection and Model Prediction

Pavani Jaya Keerthana Chittedi (Student ID:)
pavanijayakeerthanchittedi@my.unt.edu
Major: : Cyber Security
Role: Python programming, Dataset selection,
Domain Understanding & Feature Engineering,
Parameter tuning and Model training

Meeting Schedule and Collaboration

We are a group of 5 individuals from multiple departments, having different class schedules. Considering the need for team collaboration, we have decided to either meet virtually with the help of Microsoft Teams/Google Meet or meet in person at Willis Library depending on the availability at the below-mentioned timings:

Monday: 6:00 PM to 7:00 PM

Thursday: 10:00 AM to 12:00 PM

Friday: 11:00 AM to 1:00 PM

We are using these timings to discuss the progress of individual tasks and to understand the dependency of one task on another. We ensured to share the roles and responsibilities in such a way that each task is handled by 2 or more individuals. In this way, we can brainstorm multiple ideas and techniques which will yield a higher learning curve and exposure.

We are using the below-mentioned platforms for smooth communication and secure information storage:

Outlook Mailing List: sdforai_cardiovascular@groups.students.untsystem.edu

SharePoint: https://myunt.sharepoint.com/:w:/r/sites/SDforAI-Project1

Please note that the links have limited access. Drop an email to any team member for access.

Individual Roles

We have divided the project into multiple tasks and below is the summary of the owner-wise task allocation:

Task	Keerthi	Indhu	Vishal	Jyothika	Keerthana
Python Programming	Yes	Yes	Yes	Yes	Yes
Understanding the problem statement	Yes	Yes	Yes	Yes	Yes
Dataset Selection	-	-	Yes	-	Yes
Domain Understanding	Yes	-	-	-	Yes
Feature selection and Feature engineering	Yes	-	-	Yes	Yes
Exploratory Data Analysis	Yes	Yes	-	-	-
Data Visualization	-	Yes	-	Yes	-
Data Pre-processing	-	Yes	-	Yes	-
ML Model/Algorithm selection	Yes	-	-	Yes	-
Model Training	-	Yes	-	-	Yes
Model Prediction	-	-	Yes	Yes	-
Model Evaluation	Yes	-	Yes	-	-
Model Re-training/ Parameter tuning	-	Yes	Yes	-	Yes
Boosting Techniques	Yes	Yes	Yes	Yes	Yes
Model Evaluation and Final Model Saving	Yes	-	Yes	-	-
Project Documentation	Yes	Yes	-	-	Yes

Details of roles of each individual:

Gopireddy Lakshmi Keerthi is responsible for:

a. *Python programming:* Focusing on programming involving Scikit-learn
b. *Project Documentation:* Solely responsible for documenting the work done by the team on a regular basis
c. *Domain Understanding:* Focusing on understanding the problem statement in deeper terms to understand what each feature refers to.
d. *Feature selection & feature engineering:* With enough domain knowledge, responsible for selecting features manually depending on multiple criteria. Select features using feature selection techniques like Recursive Feature Elimination if necessary.
e. *Exploratory Data Analysis:* Perform univariate and bivariate analysis of the data to analyze the patterns of data. Share the knowledge extensively with feature engineering task members.
f. *Model Training:* Focusing on training data on tree-based machine learning models.
g. *Model Evaluation:* Depending on the sensitivity of the project, decide on the metrics which should be used to evaluate the performance of the model. Create necessary output visualizations for easy understanding.

Indhu Sri Krishnaraj is responsible for:

a. *Python Programming:* Focusing on model training using the Scikit library.
b. *Understanding the problem statement:* Understanding the problem and focusing on the solution to solve it.
c. *Exploratory Data Analysis:* Will be analyzing datasets with the help of data visualization.
d. *Data Pre-processing:* Preparing the raw data and making it suitable for the ML model.
e. *Data Visualization:* Focusing on different visualization methods to make the dataset easier for understanding.
f. *Model Training:* Focusing on training data sets using support vector machines and Naïve Bayes classifier.
g. *Parameter tuning:* Manual parameter tuning and understanding of each parameter after selecting the algorithm.

Vishal Rachuri is responsible for:

a. *Python Programming:* Focusing on programming involving Numpy and Pandas.
b. *Dataset Selection:* Considering project requirements, select an accurate dataset from all available datasets and evaluate the data quality.
c. *Data Visualization:* Researching available data visualization tools and suggesting the appropriate data visualization tool for heart disease prediction
d. *Parameter tuning:* Defining the performance metric, preparing the data, and splitting the data into training and testing sets.
e. *Model Evaluation:* Focusing on understanding evaluation metrics, comparing different models' accuracies after testing, and choosing the best among them.

Jyothika Vollireddy is responsible for:

a. *Python Programming:* To achieve this, we will have to import various modules in Python. We will be using Jupyter notebook for execution.
b. *Understanding the problem statement:* Concentrating on deep understanding of region-based features and their relationship with the output.
c. *Feature selection and Feature engineering:* Region-based feature selection depending on domain knowledge.
d. *Data Visualization:* Different visualization techniques are to be used to understand the relationship between features.
e. *Data Pre-processing:* Learn to use WEKA, an open-source software tool that consists of an accumulation of machine learning algorithms for data mining and data pre-processing.
f. *ML Model/Algorithm selection*: Work on understanding Machine Learning algorithms like Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbour, and Naive Bayes and use it for model training.
g. *Model Prediction:* Use the trained ML model to predict the output for test data.

Pavani Jaya Keerthana Chittedi is responsible for:

a. *Python Programming:* Perform data analysis, cleaning, and pre-processing using Pandas and NumPy libraries.
b. *Understanding the problem statement*: Understanding the problem and focusing on the solution using Machine Learning. Feature-wise domain understanding.
c. *Dataset Selection:* We have to collect datasets from Kaggle and GitHub.
d. *Domain Understanding:* Concentrate on the project statement in detail and analyze the requirements to understand and derive the solution.
e. *Feature selection and Feature engineering*: Selection of features using domain knowledge and exploratory data analysis
f. *Model Training:* Focusing on splitting the dataset into training and testing. Creating a flow for training the model with the training dataset.
g. *Parameter tuning:* Work on parameter tuning for necessary algorithms

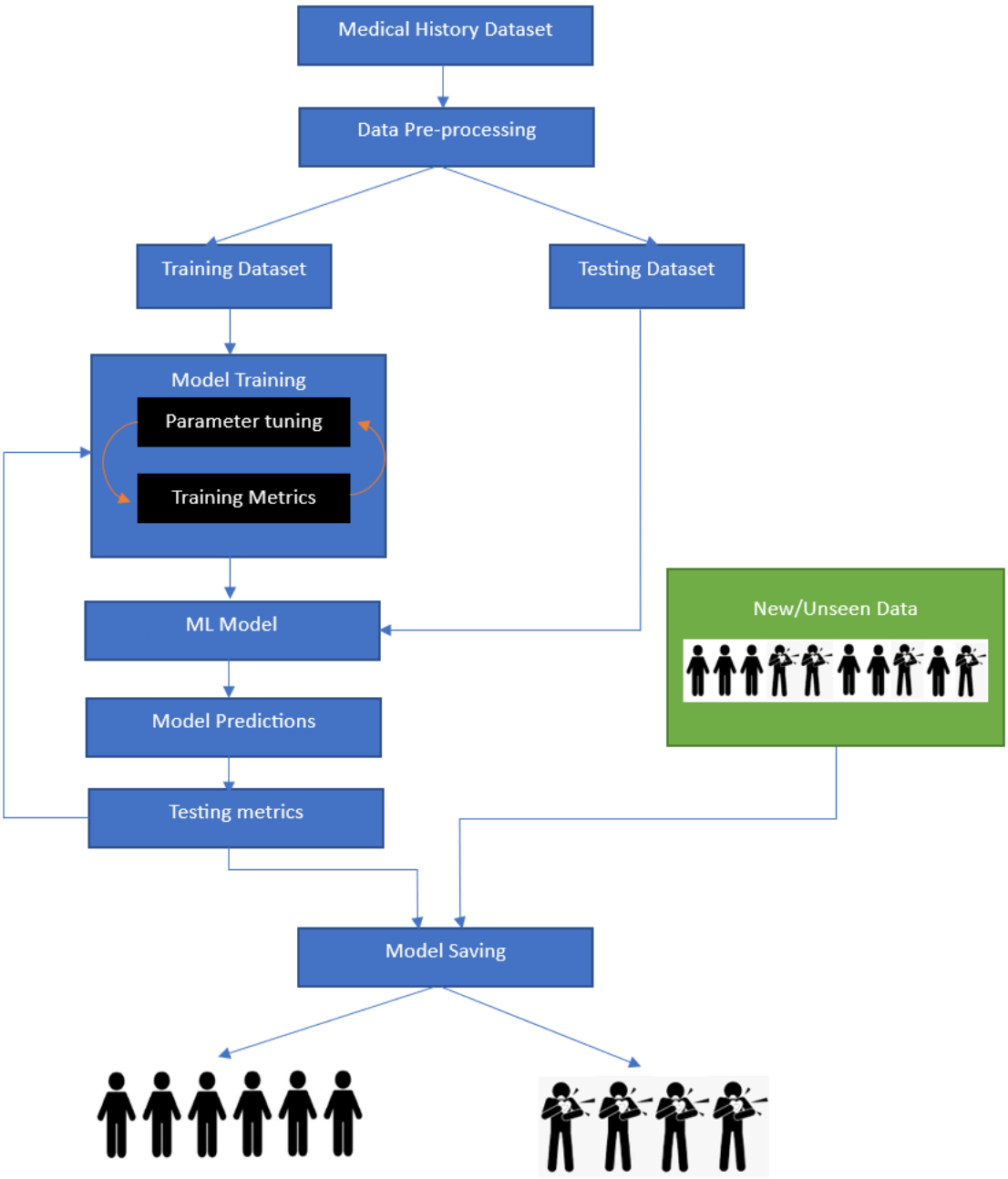
Project Abstract

Did you know that one person dies every 34 seconds in the United States of America due to cardiovascular disease? Yes, according to the statistics provided by the centers for disease control and prevention, about 697,000 people in the United States died from heart disease in 2020; that is 1 in every 5 deaths. With the continuous changes in lifestyle and food habits, it is apparent to envision an increase in the number of persons affected with cardiovascular disease over the next years. Early detection of cardiovascular disease can be helpful in timely treatment and reduce the chances of death. Several machine-learning techniques are being used in the medical field these days for the early prediction of disease. Our study aims to build a machine-learning model for the prediction of cardiovascular disease.

Previously collected demographic and medical information of patients acts as the main source to analyze patterns and predict the disease. There are multiple machine learning algorithms like Random forests, K-Nearest Neighbour (KNN), Support Vector Machines, and Naïve Bayes, which will use the dataset containing demographic information and medical history to build a predictive model which will project the probability of an individual having cardiovascular disease. The model should be trained on a diverse dataset and should be evaluated by considering different evaluation metrics considering the importance of the medical field and its direct relation to human life.

This study is to demonstrate that machine learning can be effectively used for predicting the risk of cardiovascular heart disease and can be a valuable tool for healthcare providers in making informed decisions for the prevention and treatment of the disease.

Visual Demonstration - Prediction of Cardiovascular Disease



Project Design and Technology Stack

The technology stack for implementing a cardiovascular disease prediction system can include the following:

1. **Programming languages:** Python and R are commonly used programming languages in the field of machine learning and data science. They have a large number of libraries and modules available for building machine learning models.
2. **Feature Selection/Engineering**: We use modules in the Scikit Learn library for feature selection which gives us the feature importance of each feature.
3. **Data Pre-processing/Analysis**: Using NumPy, Pandas, and SciPy libraries for Data Pre-processing and Exploratory data analysis.
4. **Machine Learning Algorithms/Frameworks**: As we are using a tabular dataset to create the model for this classification problem, it is ideal to go with machine learning models like Naive Bayes classifier, support vector machines, tree-based classifiers – XGBoost, LightGBM, etc. For implementing these algorithms, we will be using the Scikit-Learn library considering the ease.
5. **Data Visualization libraries**: We are using Plotly, Matplotlib, and Seaborn for Data Visualization.
6. **Client-side hardware/software:** The heart disease prediction system can be accessed by medical professionals and patients through a web interface. The client-side hardware and software can include a web browser and a device with Internet access, such as a desktop computer, laptop, or mobile device.

Additionally, various design methods and techniques can be used in heart disease prediction, such as principal component analysis, cross-validation, and model interpretability. These methods help to improve the accuracy and reliability of the predictive models.

Workflow

Considering multiple levels of developing a machine learning model, we have come up with **10** milestones that will help us to evaluate the progress of the project:

1. Understanding the problem statement
2. Dataset Selection
3. Domain Understanding
4. Feature selection and Feature engineering
5. Data Preparation:
    a. Exploratory Data Analysis
    b. Data Visualization
    c. Data Pre-processing
6. Machine Learning Model Implementation:
    a. Model/Algorithm selection
    b. Model Training
    c. Model Prediction
7. Model Evaluation
8. Model Re-training:
    a. Parameter tuning
    b. Boosting Techniques
9. Model Evaluation and Final Model Saving
10. Project Documentation

References

Datasets:
https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?resource=download
https://github.com/Syed-Owais-Noor/ML_Project_on_Heart_Disease_Dataset
https://github.com/Bakar31/Heart-Disease/blob/master/data/heart.csv
https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data
https://www.kaggle.com/datasets/bhadaneeraj/cardio-vascular-disease-detection
https://archive.ics.uci.edu/ml/datasets/heart+disease
https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

Research Papers:
https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf
https://www.researchgate.net/publication/341870785_Heart_Disease_Prediction_using_Machine_Learning
https://www.ijeat.org/wp-content/uploads/papers/v9i3/B3986129219.pdf
https://ieeexplore.ieee.org/abstract/document/9333574

Algorithm Papers:
https://eudl.eu/pdf/10.4108/eai.13-7-2017.2270596
https://www.jmlr.org/papers/volume2/tong01a/tong01a.pdf
https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
https://www.cs.toronto.edu/~rgrosse/courses/csc2515_2019/tutorials/tut4/tut4_boost.pdf