M.Sc. Data Science and Analytics

**Interim Progress Report**

7COM1039-0509-2021- Advanced Computer Science Masters Project

**Student Number: 19060085**

**Student Name: Mallidi Laxmi Kiran Reddy**

**Supervised by: Chidinma Chiejina**

**Project Title**

**Building a Machine Learning Application for Heart Risk Prediction**

**Table of Contents**

[**Introduction** 3](#_Toc108792149)

[Aim 3](#_Toc108792150)

[Research Questions 4](#_Toc108792151)

[Objective Questions 4](#_Toc108792152)

[**Literature Review** 4](#_Toc108792153)

[Relevant Work 5](#_Toc108792154)

[Comparison of previous work 6](#_Toc108792155)

[**Project Progress** 7](#_Toc108792156)

[Data Mining 7](#_Toc108792157)

[Attributes of data 9](#_Toc108792158)

[**Application architecture** 10](#_Toc108792159)

[**Social, Professional, Legal and Ethical Issues** 11](#_Toc108792160)

[Social Issues 11](#_Toc108792161)

[Legal and Ethical Issues 11](#_Toc108792162)

[Professional Issues 11](#_Toc108792163)

[**Tools, technologies, and algorithms** 11](#_Toc108792164)

[Tools Required 11](#_Toc108792165)

[Algorithms 11](#_Toc108792166)

[Technologies Required 12](#_Toc108792167)

[Project Planning 12](#_Toc108792168)

[**Appendices** 12](#_Toc108792169)

[Dataset URL 12](#_Toc108792170)

[**References** 14](#_Toc108792171)

# **Introduction**

Heart is the most important and sensitive entity to all living organisms for survival. Primary function of the heart is to pump the blood to all parts of the human body. It is essential to protect heart with proper care which includes proper diet, adequate exercise daily etc. Any interruption to heart function can result in severe consequences like heart attack (Cardiac arrest). There are different types of heart diseases present; coronary artery disease (CAD), and heart failure (HF) are the most common heart diseases that are present (Srinivas & Rahul, 2020).

Heart diseases have been rapidly growing for last few years and effecting wide range of age groups. Various symptoms can be observed in the heart disease patients which varies from person to person. Heart disease is very critical and can cost lives. So, there is a firm need of identifying this in a efficient way and start treatment as soon as possible. Heart disease risk factors are divides into two categories. The first category includes the patient family history like sex and age which cannot be changed; second category includes factors like high cholesterol level, smoking, high blood pressure and inadequate physical activities (L. Ali, 2019). However, there are different ways to treat heart disease; the most popular and used way by the physicians is Angiography technique (Srinivas & Rahul, 2020). But it is expensive and tedious method, as the physicians must go through all the medical parameters manually and detect the heart disease. This may also lead to human errors which might panic the patient. So, there is need for heart disease detection to be automated.

Machine Learning has been evolving in the medical field for various problems for years. It can be an efficient solution for the chosen problem as well. We can build an application using machine learning to predict the heart disease in the patient using various medical features.

In this project, machine learning application will be built to be able identify heart disease using the various patient medical features. Project will be driven purely based on the data science life cycle that includes five phases: Exploratory Data Analysis (EDA), Feature Engineering, Feature Selection, Model Building and Model Deployment. In EDA phase, all medical features will be explored in detail against target feature. Furthermore, data will be cleaned during the feature engineering phase. Then, will find the most important features in the data for the model training using correlation technique. Followed by which, model will be trained with processed data and performance will be interpreted using various techniques like accuracy, confusion matrix, f1-score, precision, roc and AUC curve and AUC score. Then different supervised, unsupervised, and deep networks models performance will be compared; and then best model will be deployed into web application using flask framework.

## Aim

The project aim is to build a machine learning application that can identifies the risk of heart attack using a patient medical history. Heart disease may occur due to many reasons that includes decrement of blood flow in the body, alcohol consumption, smoking, cholesterol in body, having over fatty foods, not having enough proper diet and lack of physical activities (Samreen, 2021). In this project, a web application will be developed using flask framework integrated with HTML, where user can enter his/her required medical details and then machine learning algorithm in the backend will be able to produce the full detailed report to the user though the front-end web application.

## Research Questions

Below research questions have been prepared to achieve the project aim and to conduct the project smoothly and successfully.

1. What are the significant features and their contributes to the model training?
2. Is it required to handle the imbalanced data, if so, how different techniques effect the results?
3. What is the best machine learning model that makes prediction effective with good accuracy and AUC score?

## Objective Questions

The following objectives have been made to fulfil aim successfully.

1. To do background research on the usage of machine learning for the heart disease prediction.
2. To acquire the good heart disease data with relevant variables and adequate size.
3. To study the gathered data and explore more to understand each variable and the role it plays.
4. To find the variables that are most important for the training using various statistical techniques.
5. To clean and transform the data into a form that can used to train machine learning algorithms.
6. To study and understand in depth of each classification machine learning algorithm which includes supervised, unsupervised, and deep learning.
7. To evaluate the algorithms performance using various metrics like accuracy, roc curve, AUC score, confusion matrix, f1 score and precision.
8. Once the model selected, them will deploy the model into web API using flask framework integrated with HTML.

# **Literature Review**

Many researchers have done the research and published their ideas on identifying the heart disease using machine learning application. There are many approaches for heart disease prediction. But the accuracy of the model if manly dependent on the quality of that data in terms of relevancy. In this section, some of the significant research papers will be reviewed and discussed the challenges that they have faced during the implementation; and the way they overcome.

## Relevant Work

Narendra Mohan et al. (2022) have used various machine learning supervised algorithms for the efficient way of automatic identification of heart disease in a human being using different medical parameters. The main objective of this research was to boost the accuracy of supervised models for heart disease prediction. The heart disease data has been taken from the Kaggle website. Author faced main problem with uncertainty in the data which they have overcome through processing and cleaning. They have used K Nearest Neighbours (KNN), Naïve Bayes, Random Forest classifier and Logistic Regressor algorithms. Finally, they have obtained 90% accuracy using Logistic Regression.

Rethna,J et al. (2021) have also approached this problem using supervised machine learning algorithms. The data set has been taken from UCI repository (Cleveland data). In this research, total four supervised algorithms have been employed for training for heart disease identification. The problem they faced might be imbalanced data. Hence, they have analysed the performance of the model’s using area under curve (AUC) instead the accuracy. AUC scores as follows; Decision tree: 0.80, Logistic Regression: 0.82, Naïve Bayes: 0.89 and SVC: 0.80.

Yadav, D.P et al. (2021) have gathered heart disease data from UCI repository. They have used four machine learning algorithms for heart disease prediction. Their main objective is to boost the accuracy of the existing models by adding a new step called feature optimization which has been applied after the model performance analysed. The initial challenges they had were accuracy same as existing models. So, then they have used the feature optimization, with which they could obtain 0.97 accuracy for Naïve Bayes model which was 0.87 before the feature optimization. However, the rest three model’s accuracy remained same or not much effected.

Kavitha, et al. (2021) have obtained data from UCI. The initial challenge that researchers found that no improvement or no impressive accuracy from individual models. Then they decided to combine two classification model (Decision tree + Random Forest) to build hybrid model. It is all about training the second model with probabilities acquired from first model. So, the model would be more robust. As a result, they have acquired 0.88 accuracy.

## Comparison of previous work

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Authors** | **Algorithm Used** | **Data** | **Accuracy** | **DOI** | **Summary** |
| Narendra Mohan et al.(2022) | KNN, Naïve Bayes, Random Forest. | Heart Disease Data from Kaggle. | 0.9 | 10.1109/ISCON52037.2021.9702314 | Author has faced problems uncertainity in the data which he has overcome by processing data. Accuracy 90% has been obtained through various models. |
| J. Rethna Virgil Jeny et al. (2021) | Decision Tree, Naïve Bayes, Logistic Regression and Support Vector Classifier | Heart Disease data from UCI (Cleveland). | 0.89 | 10.1109/ISPCC53510.2021.9609468 | In this system, four supervised ML algorithms have been used. AUC scores obtained as follows; Decision tree : 0.80, LogisticRegression : 0.82, Naïve bayes : 0.89 and SVC: 0.80. |
| D.P. Yadav, et al. (2021) | Random forest, KNN, SVM and Naïve Bayes. | Heart Disease data from UCI. | 0.96 | 10.1109/ISCON52037.2021.9702410 | Four ML algorithms have been used by the authors in this research. Their objective is to improve the accuracy of existing models by adding a new step called feature optimization. They have obtained 0.97 accuracy after oipotimization which was 0.87 before. However, rest 3 algorithms accuracy reamined same. |
| Kavitha, et al. (2021) | Hybrid model(RandomForest + Decision tree) | Heart Disease data from UCI. | 0.88 | 10.1109/ICICT50816.2021.9358597 | As researcher found out that individual models not giving effective accuracy. They have combined two classifiers (RandomForest + Decision Tree) and built hybrid model. They could able to obtain 0.88 accuracy using hybrid model. |

# **Project Progress**

As the aim of the project is to build a machine learning application to identify the heart disease. It has been divided into two parts; primary one is training a different model and pick it up a model with a good accuracy than the existing models, whereas in second part web application needs to be built which runs on the machine learning algorithm at the backend. As of now few phases in the first part have been finished (Data gathering, Data exploration and data cleaning). Model building has been started recently and currently in progress.

## Data Mining

Data has been collected from the Kaggle website. As per the information in the Kaggle page this selected dataset is a combined version of data from four databases. They are Cleveland, Hungary, Switzerland, and Long Beach V (Kaggle, 1988).

The original data has seventy-six attributes, but the sample data on Kaggle contains only 14 attributes. The dependent feature has been named as “target” which is binary column, contains only “1” or “0”. Dataset contains 1025 instances and 14 attributes in total.

Hungary

Long Beach V

Switcher land

Cleveland

**Combined Data**

A screenshot of a computer

Description automatically generated with medium confidence

**Fig 1: Raw data**

A screenshot of a computer

Description automatically generated with medium confidence

**Fig 2: cleaned Data**

## Attributes of data

The attributes of the data are discussed as follows:

* **Age**: The age feature indicates the age of individual patient in the dataset.
* **sex**: The sex attribute indicates the gender of the patient that takes the heart disease detection test.
* **cp**: This feature indicates the type of chest pain that has been encountered by the patient at the time of observation.
* **trtbps**: This attribute indicates the blood pressure of the patient during the observation.
* **chol**: This is the numerical features in the heart disease dataset that indicates the value of cholesterol in the patient at the time of observation.
* **fbs**: This is the numerical features in the heart disease dataset that indicates the value of blood sugar in the patient at the time of observation.
* **restecg**: This attribute of the heart disease dataset represents the electrocardiographic results while the patients were resting.
* **thalachh**: This feature in the dataset indicates the maximum heart rate of the patient while they were resting.
* **exng**: This Column of the dataset indicates the value of agina while at the time of exercise by the patients.
* **oldpeak**: The quality of the coronary illness dataset shows the ST despondency level that has been prompted in the body of the patient at the hour of activity.
* **slope**: The characteristic of the coronary illness dataset that demonstrates the depression level of the patient who has been taken into perception to recognize coronary illness.
* **ca**: The characteristic of the coronary illness dataset shows the count of the significant vessels of the patients who were under the test for the assurance of cardiovascular failure.
* **thal**: The characteristic of the coronary illness dataset shows the kind of deformity that has been found in the core of the patients who has been taken into perception to distinguish coronary illness.
* **Target**: This is the reliant characteristic of the information and will be utilized as the objective information to foresee whether the patient has coronary illness. The property contains two names that are marked as 1 for coronary illness and 0 for a sound heart.

# **Application architecture**

Go through previous works

Data Gathering

Data Exploration

Feature Engineering

Feature Selection

1. Null values
2. Outliers
3. Data types
4. Clean data

Select important features

Split the data

Model Deployment

**Algorithm Training**

1 Supervised classifiers

2 Unsupervised

3 deep networks

Handle Imbalanced data

Metric Evalution

Handle Imbalanced data

Web Application

Flask framework & HTML

Is accuracy good?

YES No

# **Social, Professional, Legal and Ethical Issues**

## Social Issues

* No one personal information has been or will be used in the process of conducting this project.
* This is being designed to benefit society with early detection of heart disease.

## Legal and Ethical Issues

* Dataset has been taken from the public source (Kaggle) that can be accessed by anyone for free.
* No permission or payments required to obtain data.
* Data that has been gathered will not be misused.

## Professional Issues

* This project work is solely being conducted by me; no other person is involved.
* All the work will be done as per the university requirements and guidelines.
* Anything that will be taken from other sources will be cited.

# **Tools, technologies, and algorithms**

## Tools Required

In this project, python version: 3.9.12 has been chosen due to high flexibility to analyse and build a machine learning application.

Secondly, might need a VS code to deploy the trained model into flask integrated with HTML for web application.

## Algorithms

* Logistic Regression
* Decision Tree
* Naïve Bayes
* Random forest
* XGBoost
* ADAboost
* SVM
* KNN
* ANN

## Technologies Required

* Data analysis
* Statistical analysis
* Data Visualization
* Machine Learning
* Deep learning
* Web development

# Project Planning

Project panning with be show below using the Gannt chart.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Day 1-12** | **Day 13-21** | **Day**  **22-28** | **Day**  **29-40** | **Day**  **41-45** | **Day**  **45-65** | **Day**  **66-70** | **Day 71-80** |
| **Introduction** |  |  |  |  |  |  |  |  |
| **Go through relevant works done** |  |  |  |  |  |  |  |  |
| **Gather data and analyse** |  |  |  |  |  |  |  |  |
| **Select machine learning classifiers (algorithms)** |  |  |  |  |  |  |  |  |
| **Training the different algorithms and evaluating the test performance using various metrics. Compare the result with other models.** |  |  |  |  |  |  |  |  |
| **Build the Web API and integrate with HTML and best model that gets picked based on their performance.** |  |  |  |  |  |  |  |  |
| **Showcase the detailed result of heart disease detection** |  |  |  |  |  |  |  |  |
| **FPR and Concluding the research** |  |  |  |  |  |  |  |  |

# **Appendices**

Few insights from the data analysis will be discussed in this section. The data exploration part and cleaning part have been finished.

Some insights data analyses insights will be shown below.

## Dataset URL

Dataset has been taken from Kaggle website and URL will be given below:

**URL**: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Table

Description automatically generated Chart, histogram

Description automatically generated

**Fig1: NULL value report Fig2: Distribution of Age**

**after cleaning**

Timeline

Description automatically generated

**Fig3:Heatmap (Used to show glimpse of correlation between features)**

Chart, bar chart

Description automatically generated

**Fig4: Target count against gender**

# **References**

L. A., 2019. An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure. *IEEE Access,* Volume 7, pp. 54007-54014.

Samreen, 2021. A Classification Approach for Heart Disease Diagnosis using Machine Learning. *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC).*

Srinivas, P. & Rahul, K., 2020. Predicting Heart Disease at Early Stages using Machine Learning: A Survey. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC),* pp. 302-305.

N. Mohan, V. Jain and G. Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-3.

J. R. V. Jeny, N. S. Reddy, P. Aishwarya and Samreen, "A Classification Approach for Heart Disease Diagnosis using Machine Learning," 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021, pp. 456-459.

D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-5.

M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333.

Kaggle, 1988. Heart Disease Dataset. [Online] Available at: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset [Accessed 2022]