**DAYANANDA SAGAR UNIVERSITY**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**SCHOOL OF ENGINEERING**

**DAYANANDA SAGAR UNIVERSITY**

**MAIN CAMPUS(Harohalli)**

**BANGALORE - 562 112**



**MINI PROJECT REPORT**

***ON***

# “HEART ATTACK PREDICTION USING SIMPLE MACHINE LEARNING ALGORTHMS”

**SUBMITTED TO THE 5th SEMESTER**

**ARTIFICIAL INTELLIGENCE AND MACHINE**

**LEARNING(22CS3503)**

**BACHELOR OF TECHNOLOGY**

***IN***

**Computer Science And Engineering 2024-25**

***Submitted by***

**ABHAY** -(ENG22CS0003)

**BILAL AHMED NAUSHAD KALBURGI**-(ENG22CS0032)

**FEROZKHAN H MANTUR**-(ENG22CS0055)

***Under the supervision of***

**DR. T POONGODI**

**Professor, Dept. of CSE, DSU**

# DAYANANDA SAGAR UNIVERSITY

**School of Engineering, Main Campus, Bangalore-560068**



## CERTIFICATE

***This is to certify that Mr.* ABHAY  *bearing USN***

***ENG22CS0003 has satisfactorily completed his Mini Project as prescribed by the University for the V semester B.Tech. programme in COMPUTER SCIENCE AND ENGINEERING during the year 2024-25 at the School of Engineering, Dayananda Sagar University., Bangalore.***

Date:

Signature of the faculty incharge

|  |  |
| --- | --- |
| Max Marks | Marks Obtained |
|  |  |

Signature of Chairman

Department of Computer Science and Engineering

# DAYANANDA SAGAR UNIVERSITY

**School of Engineering, Main Campus, Bangalore-560068**



## CERTIFICATE

***This is to certify that Mr.* BILAL AHMED N KALBURGI  *bearing USN***

***ENG22CS0032 has satisfactorily completed his Mini Project as prescribed by the University for the V semester B.Tech. programme in COMPUTER SCIENCE AND ENGINEERING during the year 2024-25 at the School of Engineering, Dayananda Sagar University., Bangalore.***

Date:

Signature of the faculty incharge

|  |  |
| --- | --- |
| Max Marks | Marks Obtained |
|  |  |

Signature of Chairman

Department of Computer Science and Engineering

# DAYANANDA SAGAR UNIVERSITY

**School of Engineering, Main Campus, Bangalore-560068**



## CERTIFICATE

***This is to certify that Mr.* FEROZKHAN H MANTUR  *bearing USN***

***ENG22CS0055 has satisfactorily completed his Mini Project as prescribed by the University for the V semester B.Tech. programme in COMPUTER SCIENCE AND ENGINEERING during the year 2024-25 at the School of Engineering, Dayananda Sagar University., Bangalore.***

Date:

Signature of the faculty incharge

|  |  |
| --- | --- |
| Max Marks | Marks Obtained |
|  |  |

Signature of Chairman

Department of Computer Science and Engineering

### DECLARATION

We hereby declare that the work presented in this mini project entitled - “**HEART ATTACK PREDICTION USING SIMPLE MACHINE LEARNING ALGORITHMS** “, has been carried out by us and it has not been submitted for the award of any degree, diploma or the mini project of any other college or university.

**ABHAY** -( ENG22CS0003)

**BILAL AHMED NAUSHAD KALBURGI**-**(**ENG22CS0032)

**FEROZKHAN HAJISAB MANTUR**-( ENG22CS0055)

### ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of a task would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts with success.

We are especially thankful to the **Dean, School of Engineering Dr.Udaya Kumar Reddy K R ,Dept of CSE** ,**Chairman Dr.** **Girisha G S**, for providing necessary departmental facilities, moral support and encouragement.

We are very much thankful to our guide **DR. T POONGODI** , Professor, Dept of CSE for providing help and suggestions in completion of this mini project successfully.

We have received a great deal of guidance and co-operation from our friends and we wish to thank all that have directly or indirectly helped us in the successful completion of this project work.

**ABHAY** -( ENG22CS0003)

**BILAL AHMED NAUSHAD KALABURGI**-**(**ENG22CS0032)

**FEROZKHAN HAJISAB MANTUR**-( ENG22CS0055)

### Table Of Contents

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Contents** | **Page no** |
| **1** | **ABSTRACT** | **1** |
| **2** | ***Chapter***1 *.***INTRODUCTION** | **3** |
|  | * 1. Aim of Project   2. Scope of the Project |  |
| **3** | ***Chapter* 2** .**PROBLEM DESCRIPTION** | **3** |
|  | 2.1 Problem Statement  2.2Additional Insights Related to the Project |  |
| **4** | ***Chapter* 3 . S/W & H/W REQUIREMENTS** | **5** |
|  | 3.1 Software Requirements  3.2 Hardware Requirements |  |
| **5** | ***Chapter* 4 .DESIGN** | **7** |
|  | 4.1 Algorithm Methodology |  |
| **6** | ***Chapter* 6 .CONCLUSION** | **16** |
| **7** | **REFERENCES** | **17** |

### ABSTRACT

The “Medical Report Analysis” project aims to predict the possibilities of patient is at risk of having a heart attack based on their past health information. It uses a dataset called “heart.csv” that contains details about the patients’ health condition.

To make the prediction we used the three different types of machine learning models: “Logistic Regression” it helps to predict/detect whether something will happen like getting the heart attack on the bases of input dataset from the heart CSV file. “Support Vector Machine (SVM)” it helps to looks at the relationship between several factors and get predict the chances of heart attack of a patients. “Multiple Linear Regression” it is a more advanced model that tries to separate data into different categories which make helps to predict the outcome easily.

The performance of each model is measured using evaluation metrics such as the Accuracy, Precision, Recall, F1-score and ROC-AUC acts. Accuracy which gives the how often the model is correct. Precision which provides how many times the model given correct output/predict. Recall is actual positive cases the model correctly identified. F1-score it is balanced between the precision and the recall . ROC-AUC this are the curves which determines how well the model can distinguish between the correct and incorrect predictions made by the model.

This project shows how well and correct the machine learning can be used to help predict health risk, making it a valuable tool for doctors and healthcare professionals in diagnosing and preventing heart attacks. This project is also helpful to get to know which model will give the efficient or accurate prediction.

### Chapter 1:Introduction

#### 1.1 Aim Of the Project

The main Aim of this project is to construct a machine learning model that can predict the chances of heart attack in a patients based on their medical history like the colostrum level, blood pressure and other factors. The project tests and compares different algorithms such as Logistic Regression, SVM, Multiple Linear Regression to determine which one provides the most accurate predicted values about the risk of heart attack. This project is also helps in the predict the possibility of the getting the heart attacks by the different factors of the person.

This is also helps to predict the number of chances of a person getting the heart attack. By this we can take the action/ concern to the specific doctors to pre-treatment of the disease. Regular checkup is very important because many people don’t know the symptoms of heart attack. Around 80% of people is unaware of their condition.

#### 1.2 Scope of the Project

The scope of the project is to *Improving early diagnosis* of the heart attack in a person. It also helps to detect potential heart attack risk at early stages by analysing the persons health condition by looking over the factors which can lead to earlier interventions and treatment of a patients. *Supporting Healthcare professionals* this model can make better-informed decisions about patient care like leading to improved outcomes or not getting the better health condition of the patients.

The project evaluates each machine learning model performances by using the key metrics such as accuracy, precision, recall and f1-score. This helps to get to know the which model is reliable and effective for predicting the heart attacks. The project serves as the starting point for the further studies in the predictive analytics in healthcare.it lays the foundation for the future research to make the accurate prediction on the bases of data.

### Chapter 2.Problem Description

#### 2.1 Problem Statement

Heart attacks are one of the top most causes of death worldwide now a days. Detecting the heart attack in early stage can save the lot of lives, but doctors are misses the important signs or symptoms of heart attack when they diagnosing manually. This project are train some model which can help the doctors to detect the symptoms of heart attack automatically. By doing this it makes the diagnosis process more accurate and quicker and helps doctors make better decisions.

Here are the some statistics on heart attacks: In Global mortality about 16% of total deaths causes by the heart attack. Approximately 18 million people die from heart disease globally.

#### 2.2 Additional Insights Related to the Project

#### Challenges Addressed

In the data set fewer patients who experience a heart attack compared to those who don’t, this makes the creating *the imbalance in the data* *set*. this can make harder for models to learn effectively. The project will make the models to handles this type of imbalance in the data by creating the balanced data and improving the prediction accuracy.

This project ensures the models are evaluated from different angles to understands their real-world effectiveness by giving the importance to all the aspect such as precision, recall and other evaluation matrices.

#### Objectives

The primary goal of this project is to build a machine learning model that predicts the likelihood of a heart attack based on historical medical data of a patients. We also compare and test various machine learning algorithms, including Logistic Regression, Multiple Linear Regression and SVM to determine the most effective model for heart attack prediction.

By this project we can able to offer valuable insights into how machine learning can assist healthcare in making more informed decisions about heart attack risks and early diagnosis.

#### Opportunities Identified

#### 

There is an opportunity to enhance predictions further by making more and more advanced technique like “ensemble methods” which combines the multiple models or “deep learning” which learn from the complex patterns or train by the complex patterns by this we can make the prediction system even more powerful.

### Chapter 3:S/W And H/W Requirements

#### 3.1 Software Requirements

##### Python

The primary programming language employed in this sentiment analysis project serves as a robust and flexible platform, enabling efficient data manipulation and in-depth analysis. Its versatile capabilities allow for seamless integration of various data processing tasks, making it an ideal choice for extracting insights from textual data.

**Libraries:**

* *Pandas*: This library is essential for data handling and manipulation, allowing for efficient data structures and operations to clean and prepare datasets for analysis.
* *NumPy:* A foundational library for performing numerical operations, NumPy enables fast calculations and efficient handling of large multi-dimensional arrays and matrices.
* *Matplotlib and seaborn:* These visualization libraries are used to create informative and interactive data visualizations, helping to convey insights and patterns effectively through graphical representations.
* *Scikit-learn:* This library provides robust tools for implementing machine learning algorithms, including support for model evaluation metrics, making it easier to train, test, and validate predictive models.
* *Nltk or spacy*: These libraries are utilized for various natural language processing tasks if necessary, enabling the project to analyse and interpret text data effectively.

#### 3.2 Hardware Requirements

To effectively run the software, you will need a machine equipped with a minimum of 4GB of RAM, which will ensure smooth multitasking and efficient performance. Additionally, it should feature a modern CPU, such as an Intel i5 or a comparable processor, to handle processing tasks swiftly and efficiently.

To effectively manage and utilize the dataset and model files, it is important to allocate approximately 1GB of available storage space. This space will ensure that there is enough capacity to accommodate the necessary files without running into storage limitations.

### Chapter 4: Design

#### 4.1 Algorithm /Methodology

##### Data collection

Data collection means gathering information from different sources to evaluate or analysis. This contains such as the CSV file, Databases and APIs.

In this project we used the heart.CSV file to execute the models in this file the data is stored in the rows and columns, often used for structured data like heart disease information. The databases are storing large amount of data in tables, and we can special queries (SQL) to find the data we needed. APIs are allowed us to get data from online services or other applications.

The goal is to gather all data and convert it into a usable format and store it in one place where it can be easily analysed.

**Data Pre-processing**

Data pre-processing is the step where we clean and prepare the data before using it to train the machine learning model. This is the most important step to preform before the model get trained by the data set because the data is also is in the raw format this often make messy or incomplete which effect the prediction of model.

Feature Selection, the process of identifying and selecting the most relevant features or variables for the model, ensuring that the chosen features contribute meaningfully to the predictive power while reducing noise and improving overall model performance Feature Scaling, this process involves adjusting the range and distribution of numerical features to ensure they are on a similar scale. This can be achieved through normalization, which we found the rescales values to a range between 0 and 1, or standardization, which transforms the data to have a mean of zero and a standard deviation of one.

Applying these techniques is crucial when we are working with algorithms sensitive to the scale of data, as it helps improve the performance and convergence speed of machine learning models.

We usually check for the errors or missing values in the table of dataset if some values and information is missing or there is wrong values then we fix or remove or delete the errors to make data reliable and work efficiently with models. For sometimes the data will be filled by us by making the average , common values or remove those rows if they not necessary.

##### Model Training

1. *Train-Test Split*: Firstly we start by dividing our dataset into two distinct subsets: the training set, which we found that the values are typically comprises 80% of the data, and the testing set, which makes up the remaining 20%. This separation is crucial for evaluating the model’s performance on unseen data.

1. *Choose Model*: It means deciding which machine learning algorithm is to use for predicting heart attack risks. Different models having different strengths by looking over the each models we can chose for example “SVM” the model that tries to separate different classes based on multiple factors like the features of diagnosed / health checkup report etc then we have “logistic regression “ this model helps works well for binary classification like predicting the heart attack or no heart attack, at last we have the “multiple linear regression” by this model we can predict the continuous outcomes or values based on the multiple factors.
2. *Train Model*: Then we are fitting the chosen model to our training data. This process involves using the training set to help the model learn the underlying patterns and relationships within the data, allowing it to make predictions based on input features.

1. *Hyperparameter Tuning:* Once the model is trained, refine its performance by adjusting hyperparameters. Employ techniques such as GridSearchCV or RandomizedSearchCV, which systematically explore combinations of hyperparameters to find the optimal settings that enhance the model's predictive accuracy and overall performance.

##### Model evaluation

*Accuracy*: This metric is used to assess the overall correctness of a model by calculating the ratio of the number of correct predictions (both true positives and true negatives) to the total number of predictions made. It provides us a general idea of how often the model is correct, but it can be misleading if the dataset is imbalanced.

Formula for Accuracy is= TRUE POSITIVE+TRUE NEGATIVE

TOTAL NUMBER OF PREDICTIONS

Accuracy represent the percentage of correct predictions of each models. We have took three models like SVM, Multi-linear Regression and Logistic Regression.

In this project we got the Accuracy of each model are for SVM the percentage will be 86 %, for multi linear regression the percentage will be 85% and last for the Logistic regression 86% percentage of accuracy we got.

*Precision:* In this we are focuses specifically on the model's performance regarding positive predictions. Precision is calculated as the proportion of true positive predictions (the cases correctly identified as positive) out of all instances that the model predicted as positive (true positives plus false positives). If we get the high precision indicates that the model makes very few false positive errors.

Formula for Precision is TRUE POSITIVE

TRUE POSITIVE+ FALSE POSITIVE

For the True positive predictions among all positive predictions we get the percentage for SVM is 89% ,the precision for the Multi-linear Regression the percentage will be 93% for the Logistic Regression the percentage of precision we got 92 %.

*Recall*: It is also known as sensitivity, recall measures the model's ability to correctly identify all relevant positive cases. It is defined as the proportion of true positive predictions to the actual number of positives in the data (true positives divided by true positives plus false negatives). If we got the high recall value means that the model successfully captures most of the actual positives, making it crucial in scenarios where missing a positive case is particularly costly.

Formula for RECALL IS TRUE POSITIVE

TRUE POSITIVE+ FALSE NEGATIVE

The percentage of actual positive that the model successfully identifies. If the SVM got the recall of 80% it means 80% of SVM of all positive instance. For the Multilinear regression recall we got is 76% and the logistic regression the recall we got by the code is 79%.

*F1-Score:* In this we are going to balanced metric that combines both precision and recall into a single score by calculating their harmonic mean. This metric is especially useful in situations where one needs to find a balance between precision and recall, particularly when dealing with imbalanced datasets. The F1-score takes into account both false positives and false negatives and provides a more nuanced understanding of a model’s performance than accuracy alone.

Formula for F1` is F1-Score= 2×Precision×Recall​

Precision + Recall

It is the mean of precision and recall, we got the SVM F1 score like 85%, for the multilinear regression has F1 score is 83%, and for the Logistic Regression the F1 score is 85%.

*Cross-Validation*: This statistical method is employed to evaluate a model's generalizability by partitioning the training data into 'k' subsets or folds. The model is trained on 'k-1' folds and tested on the remaining fold, and this process is repeated 'k' times, with each fold serving as the test set once. The results are then averaged to produce a more robust estimate of the model's performance. K-fold cross-validation is particularly useful in mitigating overfitting and ensuring that the model performs well on unseen data.

##### Model Visualization

*Confusion Matrix*: This tool allows us to visualize the outcomes of our classification model by comparing the true values against the predicted ones. It provides a comprehensive overview of how well the model is performing, highlighting areas of correct predictions and misclassifications for each class.

*Feature Importance*: This analysis provides and highlights the key features that significantly influence the predictions made by the model. By visualizing these important factors, we can gain insights into which variables are driving the model's decisions, helping to enhance understanding and guide further model improvements.

##### Results Interpretation

*Evaluate Model Performance:* We found values by analyse the evaluation metrics to gauge how well the model is performing. Look for key indicators such as accuracy, precision, recall, and F1 score to determine which model demonstrates the best overall effectiveness for your specific task.

*Identify Opportunities for Enhancement*: we did the explore on various avenues for improving the model’s performance. This might include collecting additional data to help the model generalize better, conducting in-depth feature engineering to create more informative attributes, or experimenting with different algorithms to find the one that yields superior results.

#### Flow Chart

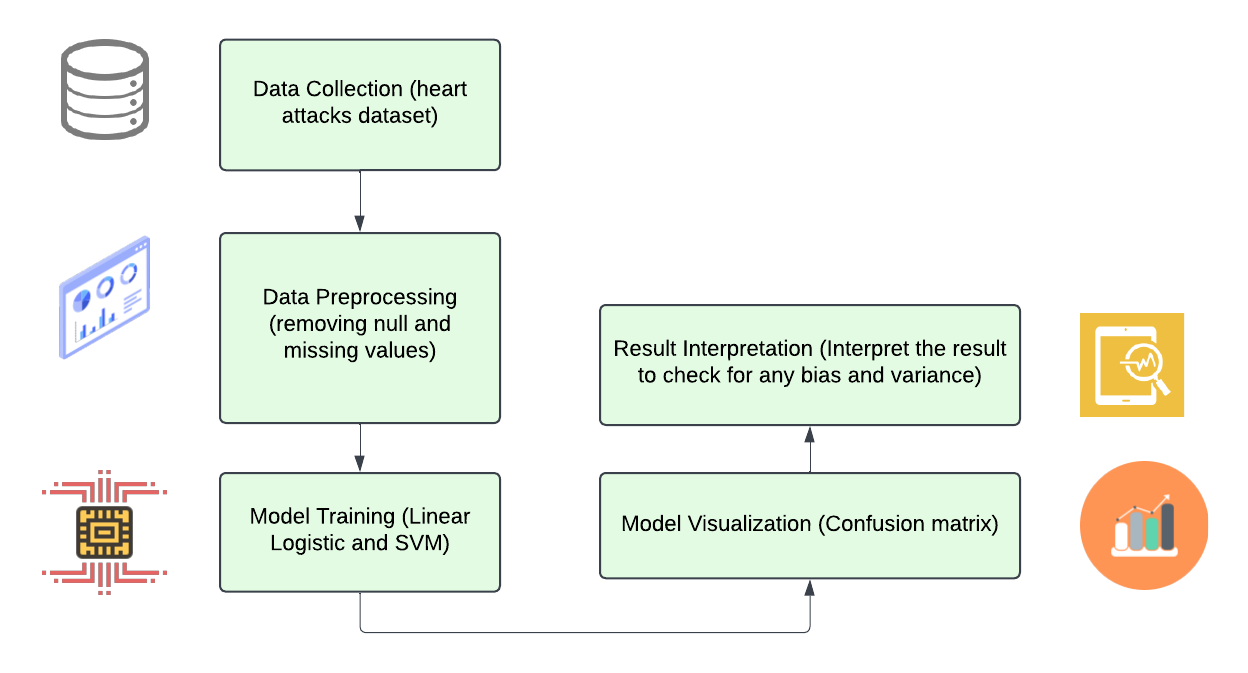


Figure Flow chart of model

##### Confusion Matrix

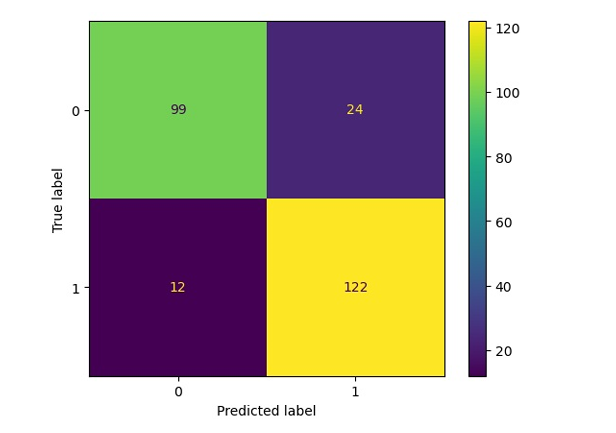


Figure 2 Confusion Matrix of SVM

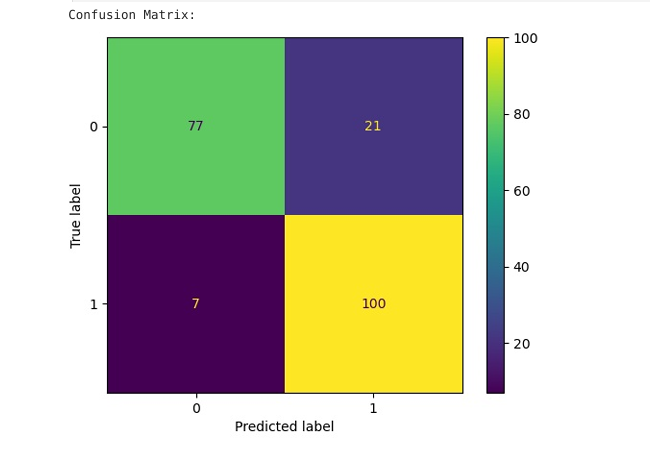


Figure 2 Confusion Matrix for the Logistic Regression

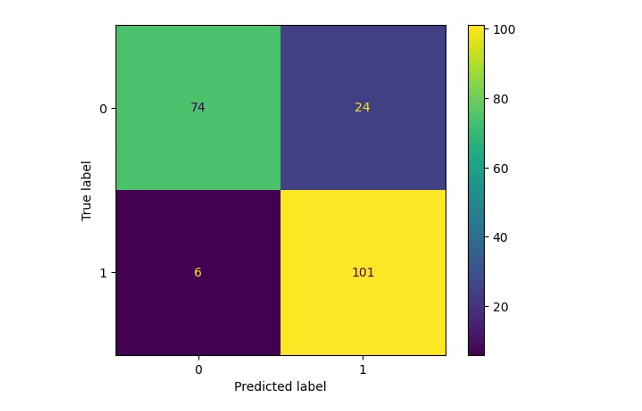


Figure 3 Confusion Matrix for the Multilinear Regression

**ROC Curve Diagram:**

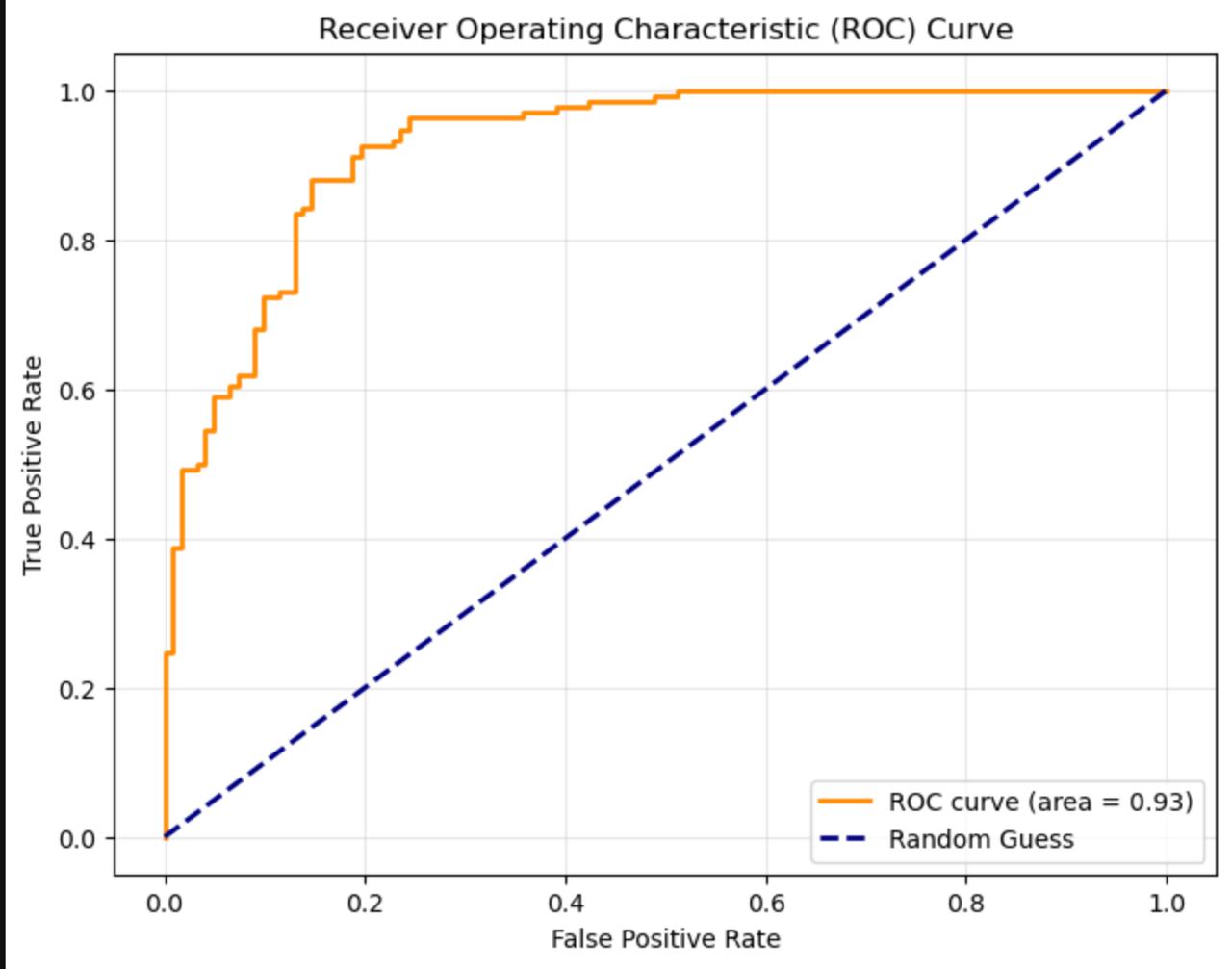


Figure SVM ROC Curve.

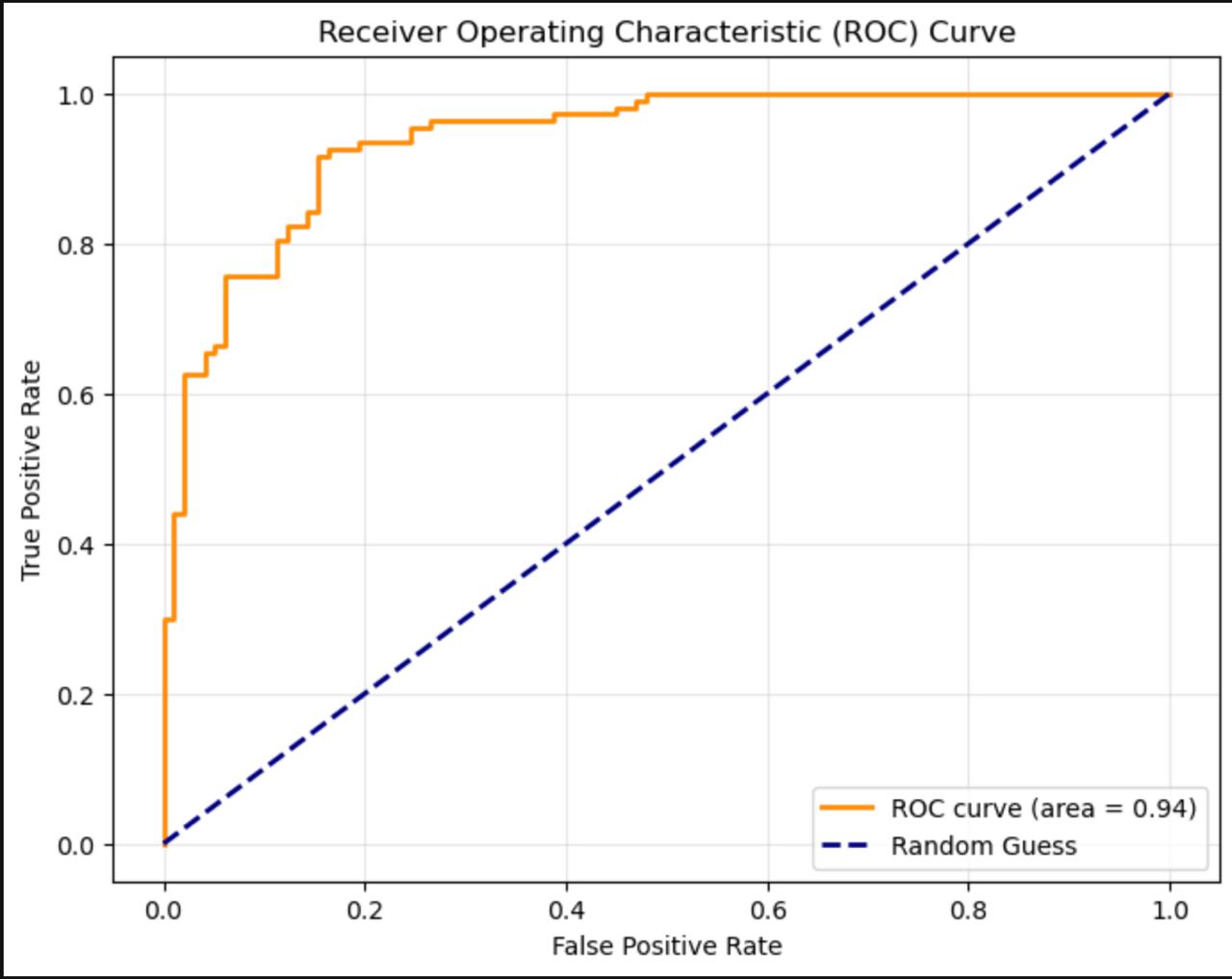


Figure Linear ROC curve

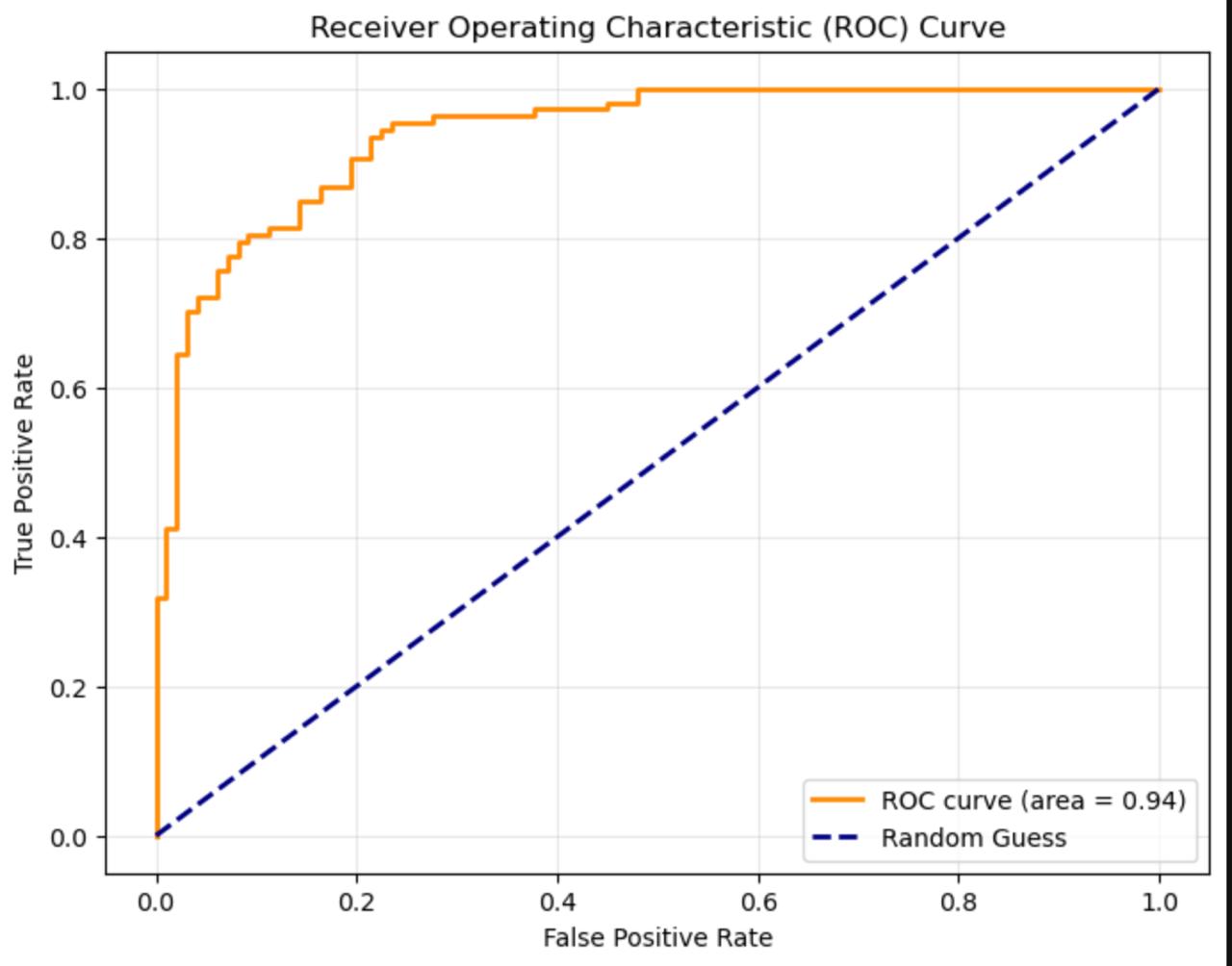


Figure Logistic Linear regression ROC Curve

### Conclusion

This project shows that machine learning models can be very effective at predicting the number of chances of getting the heart attack. After testing various models, we found that the “” performed the best in detecting the risk of the person getting the heart attack. From this project we are highlight how important it is to choose the right machine learning algorithm for the task at hand. Not all the algorithms will perform equally well for every problem, so selecting he one that works best for the specific data and problem we are working with is crucial.

Based on the evaluation metrics, SVM will become the best model for the task. It gives the highest performance across all metrics, with impressive 86% of accuracy, 89% precision and 80% recall percentages. Its F1-Score of 85% indicates a solid balance between precision and recall.

While Logistic regression takes the multiple features and performs 86% of accuracy, 92% of precision and 79% of recall, 85% of F1-score made by the logistic regression.

For multiple linear regression it takes more then two features for one variable and train the x variables to predict the value of y and in testing we got the values like 85% of accuracy , 93 % of precision, 76% of recall ,83% of F1-score .

This project proves that machine learning can be a valuable tool for predicting heart attack risk. With the correct machine learning and algorithms and improvements in the future, we could make even more accurate and reliable predictions, ultimately helping doctors and healthcare providers better assess and prevent heart attacks.

### References

#### Books

* A.Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.
* S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, 3rd ed. Birmingham,

U.K.: Packt Publishing, 2019

We have also used the given class notes (ppt, notes and the lab manual for deriving the data ).

#### Research Papers

• B. Liu and L. Zhang, "A Survey on Sentiment Analysis: Approaches, Applications, and Challenges," in *Mining Text Data*, C. C. Aggarwal and C.

Zhai, Eds. Boston, MA: Springer, 2012, pp. 415–463.

#### Online Resources

* <https://scikit-learn.org/stable/>
* [https://seaborn.pydata.org](https://seaborn.pydata.org/)