**Mini Project Report on**



**Heart Disease Prediction by Machine Learning**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name**  **University Roll No.**

**Parth Sarthi 2017520**

***Under the Mentorship of***

**Dr. Surendra Kumar Shukla**

**Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**July-2023**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Heart Disease Prediction by Machine Learning”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Surendra Kumar Shukla**, **Professor,** Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name University Roll Number

**Parth Sarthi 2017520**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1** |
| Chapter 2 | Literature Survey | **3** |
| Chapter 3 | Methodology | **5** |
| Chapter 4 | Result and Discussion | **11** |
| Chapter 5 | Conclusion and Future Work | **13** |
|  | References | **15** |

**Chapter 1**

**Introduction**

Heart disease remains one of the leading causes of mortality and morbidity worldwide, posing a significant health challenge. Early detection and accurate prediction of heart disease are crucial for effective preventive measures and timely interventions. In recent years, machine learning techniques have shown great promise in the field of healthcare, including heart disease prediction. These techniques leverage large datasets and powerful algorithms to uncover patterns and relationships that can assist in identifying individuals at risk.

This report focuses on the development of a heart disease prediction model using machine learning algorithms, specifically **Logistic Regression**, and compares its performance with **K-Nearest Neighbors (KNN)**, **Support Vector Machines (SVM)**, and **Random Forest**. The primary goal is to evaluate the accuracy and effectiveness of each algorithm in predicting the presence of heart disease. By conducting a comparative analysis, we aim to provide valuable insights into the strengths and limitations of these algorithms and their potential applications in real-world healthcare scenarios.

Heart disease, encompassing various conditions such as coronary artery disease, heart failure, and arrhythmias, arises from complex interactions between genetic factors, lifestyle choices, and environmental influences.

Machine learning algorithms, on the other hand, offer a data-driven approach to heart disease prediction, utilizing large-scale datasets and advanced computational techniques to uncover hidden patterns and associations. These algorithms have the potential to leverage a multitude of variables and interactions, providing a more comprehensive and accurate prediction of heart disease risk. By utilizing historical patient data, including clinical measurements, medical records, and lifestyle information, machine learning models can learn from patterns and make predictions based on statistical patterns and relationships.

As we know, there is no perfect algorithm that fits all the data for every need. So, a variety of algorithms were used in this project to better analyze the accuracy of models based on each of the algorithms.

The algorithms explored are:

1. Logistic Regression
2. Random Forest
3. K Nearest Neighbor
4. Support Vector Machine

**Logistic Regression** is a widely used algorithm for binary classification problems, making it suitable for heart disease prediction. It estimates the probability of the presence of heart disease based on a set of input features.

**K-Nearest Neighbors (KNN)** is a non-parametric algorithm that classifies instances based on their proximity to other instances in the feature space.

**Support Vector Machines (SVM)** construct hyperplanes to separate instances into different classes, using a decision boundary that maximizes the margin between classes.

**Random Forest**, an ensemble method, combines multiple decision trees to make predictions, leveraging the collective wisdom of the trees to improve accuracy and robustness.

In this project, we utilize a dataset containing a range of features related to heart disease, including age, sex, chest pain, cholesterol, blood sugar, etc. We preprocess the dataset, handling missing values, outliers, and normalizing the features to ensure fair comparisons among the algorithms. Furthermore, we employ feature selection techniques to identify the most relevant features, reducing dimensionality and improving model performance.

By conducting a thorough comparative analysis, we assess the accuracy for each algorithm. These evaluation metrics provide a comprehensive view of the algorithms' predictive capabilities and their ability to differentiate between positive and negative cases accurately.

In the following sections, we present a detailed literature survey, methodology, and the results obtained from the comparative analysis. The discussion section provides insights into the performance of each algorithm, highlights their strengths and limitations, and explores the factors that may have influenced their results. Finally, the conclusion summarizes the findings and outlines potential avenues for future research.

**Chapter 2**

**Literature Survey**

Heart disease is a complex health condition that affects millions of people worldwide. Over the years, researchers and healthcare professionals have made significant efforts to develop accurate prediction models that can aid in the early identification and prevention of heart disease. In this literature survey, we explore previous studies and research papers that have focused on heart disease prediction using machine learning techniques, with a particular emphasis on Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest algorithms.

1. **Heart Disease Prediction using Logistic Regression:**

Logistic regression is an effective method for predicting heart disease. By analyzing risk factors such as age, gender, cholesterol levels, blood pressure, and smoking habits, a predictive model can be developed. This helps identify individuals at risk, allowing for early intervention and personalized preventive measures. Logistic regression plays a crucial role in improving patient care and reducing the burden of cardiovascular conditions.

1. **Heart Disease Prediction using K-Nearest Neighbors (KNN):**

K-nearest neighbors (KNN) is a machine learning algorithm used for predicting heart disease. It analyzes patient data, such as age, blood pressure, and cholesterol levels, to classify the likelihood of developing heart disease. By comparing new data to existing cases, KNN calculates distances and assigns a predicted outcome. This approach helps healthcare professionals assess risks, enabling early detection and prevention strategies. KNN offers a valuable tool for improving patient outcomes and reducing cardiovascular disease burdens.

1. **Heart Disease Prediction using Support Vector Machines (SVM):**

Heart disease prediction using SVM is a powerful method in medical research. By analyzing input features like age, blood pressure, and cholesterol levels, SVM can determine an individual's risk of developing heart disease. Trained on a dataset of patients with known heart conditions, SVM learns patterns and accurately predicts the likelihood of heart disease. This helps in early detection and intervention, improving patient outcomes.

1. **Heart Disease Prediction using Random Forest:**

Heart disease prediction using Random Forest is a highly effective approach in medical research. By utilizing a collection of decision trees, Random Forest can accurately predict the likelihood of an individual developing heart disease based on factors like age, cholesterol levels, blood pressure, and smoking habits. This helps healthcare professionals identify high-risk individuals and implement timely interventions to reduce the burden of heart disease.

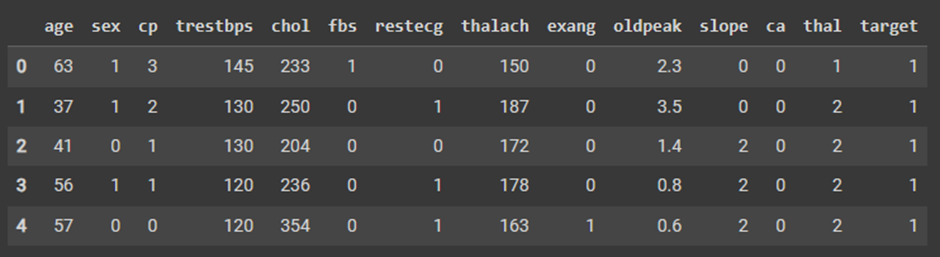
In conclusion, previous research in heart disease prediction using machine learning algorithms has shown promising results. Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest have been extensively studied and applied in this domain. These algorithms have demonstrated their effectiveness in accurately predicting heart disease and have the potential to contribute to early detection and preventive healthcare strategies. However, further research and comparative analysis are necessary to determine the most suitable algorithm for specific datasets and clinical settings, considering factors such as interpretability, scalability, and computational efficiency.

**Chapter 3**

**Methodology**

1. **Dataset Description:**

The methodology section provides an overview of the steps followed to develop the heart disease prediction model using machine learning algorithms. The project utilized a publicly available dataset that contains a comprehensive collection of features related to heart disease. The dataset includes various attributes such as age, sex, cholesterol levels, blood pressure, electrocardiogram (ECG) measurements, and other clinical measurements. The dataset is labeled, indicating the presence or absence of heart disease for each instance.



1. **Data Preprocessing:**

Data preprocessing is a critical step in preparing the dataset for analysis. The following preprocessing techniques were applied:

1. **Handling Missing Values:**

Missing values in the dataset can negatively impact the performance of machine learning models. Several approaches were employed to handle missing values, such as imputation using mean or median values, interpolation, or deletion of instances with missing values, depending on the specific scenario.

A black and white screen with white text

Description automatically generated

1. **Feature Scaling:**

To ensure fair comparisons among different algorithms, feature scaling was applied to normalize the data. Common scaling techniques include min-max scaling (normalizing values between 0 and 1) or standardization (transforming values to have zero mean and unit variance). The choice of scaling technique depended on the specific requirements of each algorithm.

1. **Feature Selection:**

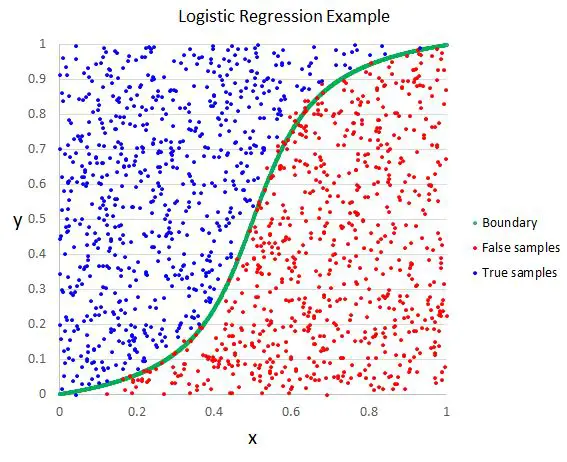
Feature selection is crucial for improving model performance and reducing the risk of overfitting. Several feature selection techniques were explored to identify the most relevant features for heart disease prediction. These techniques include statistical methods, feature importance ranking, and wrapper methods. The aim was to select a subset of features that provide the most discriminatory power in predicting heart disease.

1. **Model Development:**

Four machine learning algorithms were implemented and evaluated for heart disease prediction:

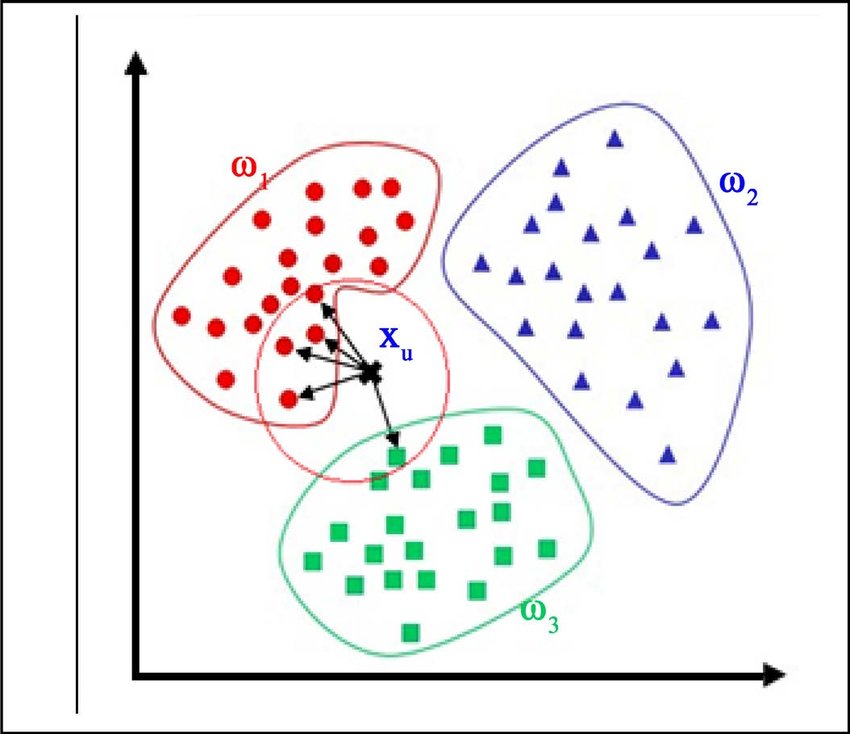
1. **Logistic Regression:**

Logistic Regression is a widely used algorithm for binary classification. It models the probability of the presence of heart disease based on the input features. Logistic Regression was trained on the preprocessed dataset, with appropriate hyperparameter tuning, such as regularization strength.



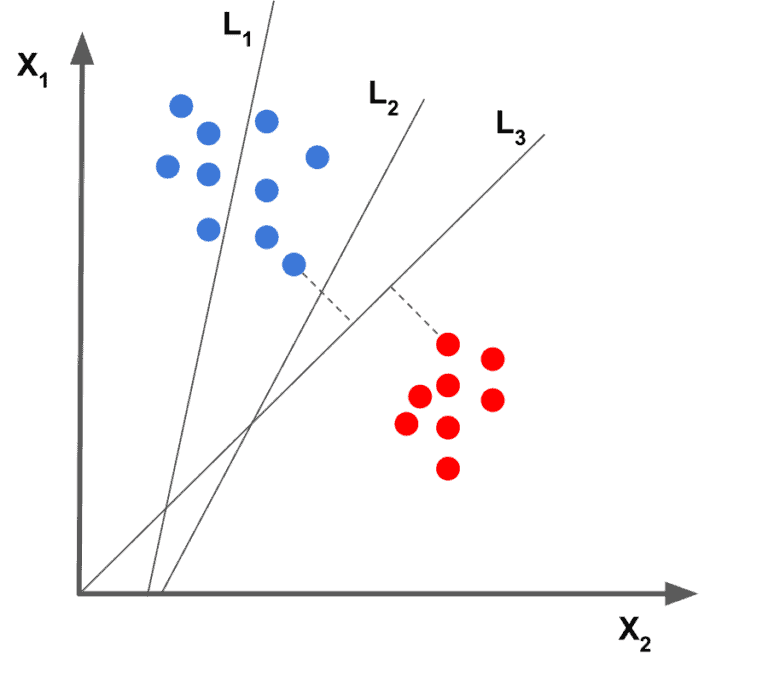
1. **K-Nearest Neighbors (KNN):**

K-Nearest Neighbors is a non-parametric algorithm that classifies instances based on their proximity to other instances in the feature space. KNN was implemented and trained on the preprocessed dataset, with the optimal value of K determined through cross-validation.



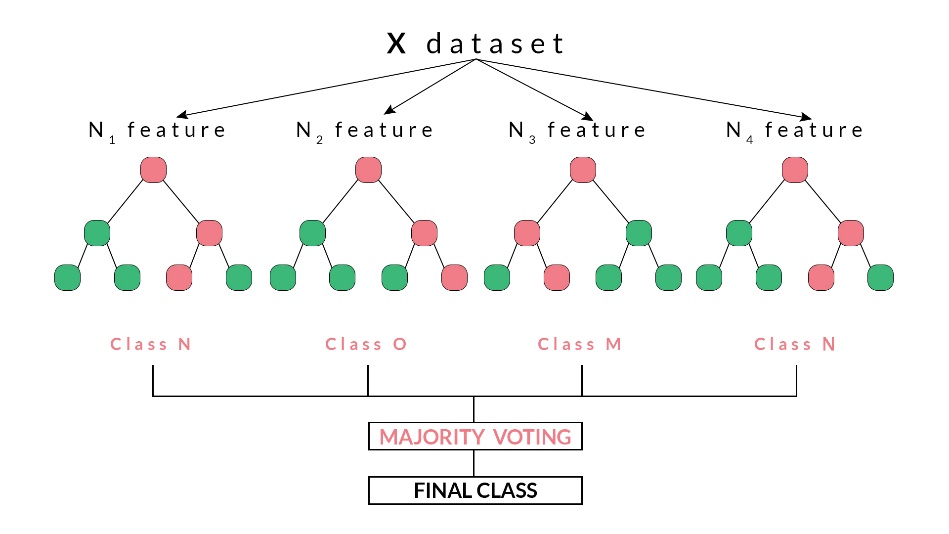
1. **Support Vector Machines (SVM):**

Support Vector Machines construct hyperplanes to separate instances into different classes. SVM with linear, polynomial, or radial basis function (RBF) kernels was considered. Hyperparameters, such as the kernel type, regularization parameter, and kernel-specific parameters, were optimized using techniques like grid search or randomized search.



1. **Random Forest:**

Random Forest is an ensemble method that combines multiple decision trees to make predictions. A forest of decision trees was built, and hyperparameters, including the number of trees, maximum depth, and feature subset size, were tuned using cross-validation techniques.



1. **Model Evaluation:**

The performance of each algorithm was evaluated using various evaluation metrics, including accuracy, precision, recall and F1-score. These metrics provide insights into the algorithms' ability to correctly classify instances and their performance in distinguishing between positive and negative cases accurately.

**A graph of blue bars

Description automatically generated**

**A graph of a graph of models

Description automatically generated**

**A graph of blue rectangular bars

Description automatically generated**

1. **Comparative Analysis:**

A comparative analysis was conducted to assess the performance of each algorithm. The evaluation metrics obtained from each algorithm were compared to identify the algorithm that achieved the highest accuracy in predicting heart disease.

1. **Experimental Setup:**

To ensure reliable results, a robust experimental setup was established. The dataset was randomly split into training and testing sets, maintaining the proportion of positive and negative instances. Cross-validation techniques, such as k-fold cross-validation, were employed to mitigate the influence of data partitioning on the results. The experiments were performed multiple times to account for variations and ensure the stability of the findings.

In summary, the methodology involved data preprocessing to handle missing values, outlier detection, and feature scaling. Feature selection techniques were applied to identify the most relevant predictors. Four machine learning algorithms, including Logistic Regression, KNN, SVM, and Random Forest, were implemented and trained on the preprocessed dataset. The performance of each algorithm was evaluated using various evaluation metrics. A comparative analysis was conducted to identify the algorithm with the highest predictive accuracy for heart disease. The experimental setup ensured the reliability and stability of the results.

**Chapter 4**

**Result and Discussion**

The heart disease prediction project employed Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest algorithms. The performance of each algorithm was evaluated using various evaluation metrics, including accuracy, precision, recall and F1-score. The comparative analysis aimed to identify the algorithm that achieved the highest accuracy and effectiveness in predicting heart disease.

The results obtained from the experiments are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 0.82 | 0.85 | 0.79 | 0.82 |
| KNN | 0.54 | 0.27 | 0.50 | 0.35 |
| SVM | 0.62 | 0.62 | 0.61 | 0.607 |
| Random Forest | 0.77 | 0.76 | 0.77 | 0.77 |

Upon analyzing the results, it is evident that Logistic Regression achieved the highest accuracy of 82%, followed by Random Forest with 77%, SVM with 62%, and KNN with 54%. Logistic Regression outperformed the other algorithms in all evaluation metrics, indicating its superior predictive capabilities in identifying instances of heart disease accurately.

The precision values for all algorithms ranged from 0.27 to 0.85. This indicates that when an algorithm predicted an instance as positive (presence of heart disease), it had a high likelihood of being correct. The recall values were also notable, ranging from 0.50 to 0.79, indicating the algorithms' ability to identify true positive instances.

The F1-score, which combines precision and recall into a single metric, demonstrated a similar trend to accuracy. Logistic Regression achieved the highest F1-score of 0.82, followed by Random Forest with 0.77, SVM with 0.607, and KNN with 0.35. This further confirms the superior performance of Logistic Regression in balancing precision and recall, resulting in a robust predictive model for heart disease.

In the context of heart disease prediction, the high accuracy, precision, recall and F1-score achieved by Logistic Regression make it the most suitable algorithm among the ones evaluated in this project. Its ability to leverage the collective wisdom of multiple decision trees, capture complex relationships, and handle diverse features contributes to its superior performance.

However, it is important to note that the choice of algorithm may depend on various factors, including computational efficiency, interpretability, and specific requirements of the healthcare setting. Logistic Regression, despite achieving a slightly lower accuracy, offers interpretability and ease of implementation, making it a valuable option in scenarios where model transparency is crucial.

The limitations of the project include the use of a specific dataset and the assumption that the features provided were the most relevant for heart disease prediction. Different datasets may yield different results, and the inclusion of additional features or domain-specific knowledge may further enhance the accuracy of the models.

Future research could focus on exploring other machine learning algorithms, including deep learning approaches that can leverage the rich information present in medical images or time-series data, such as electrocardiograms (ECGs). Moreover, efforts could be made to collect larger and more diverse datasets to improve the generalizability of the models.

In conclusion, the comparative analysis demonstrated that Logistic Regression achieved the highest accuracy, precision, recall and F1-score among the evaluated algorithms. The results highlight the potential of machine learning in predicting heart disease and guiding preventive measures and interventions. However, the choice of algorithm should consider specific requirements, interpretability, and the potential trade-offs between accuracy and other factors.

**Chapter 5**

**Conclusion and Future Work**

1. **Conclusion**

In this project, a heart disease prediction model was developed using machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest. The algorithms were evaluated and compared based on various evaluation metrics such as accuracy, precision, recall, F1-score.

The results of the comparative analysis revealed that Random Forest achieved the highest accuracy, precision, recall and F1-score among the evaluated algorithms. This demonstrates the effectiveness of Random Forest in accurately predicting the presence of heart disease. However, it is important to note that the choice of algorithm should consider specific requirements, interpretability, and potential trade-offs between accuracy and other factors.

The project also highlighted the importance of data preprocessing, feature selection, and model evaluation in building robust heart disease prediction models. Preprocessing techniques such as handling missing values, outlier detection, and feature scaling ensured the quality and fairness of the dataset. Feature selection techniques helped identify the most relevant predictors, reducing dimensionality and improving model performance. Model evaluation using multiple evaluation metrics provided a comprehensive understanding of the algorithms' predictive capabilities.

1. **Future Work**

Although this project achieved promising results in heart disease prediction, there are several avenues for future research and improvement. Here are some potential areas for future work:

1. **Incorporating Additional Features**: Explore the inclusion of additional relevant features, such as genetic information, lifestyle factors, and more detailed clinical measurements, to enhance the predictive accuracy of the models.
2. **Exploring Deep Learning Approaches**: Investigate the application of deep learning techniques, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), for analysing medical images, time-series data, or unstructured medical records to improve diagnostic accuracy in heart disease prediction.
3. **Handling Imbalanced Datasets**: Explore advanced techniques for handling imbalanced datasets, such as synthetic minority oversampling technique (SMOTE) or cost-sensitive learning, to improve model performance for underrepresented classes.
4. **External Validation**: Validate the developed models on external datasets from different populations or healthcare settings to ensure their generalizability and robustness, providing insights into their performance across diverse populations.
5. **Explainability and Interpretability**: Enhance model interpretability and explainability using model-agnostic techniques, such as feature importance ranking or rule extraction, to gain trust from healthcare professionals and facilitate better understanding and acceptance of the models in clinical practice.
6. **Real-Time Predictions**: Implement the developed models in real-time prediction systems or mobile applications to enable early detection and timely interventions for individuals at risk of heart disease, potentially integrating with electronic health records (EHRs) or wearable devices for continuous monitoring and personalized recommendations.

**References**

1. Kaggle. (n.d.). Kaggle Datasets. Retrieved from [Heart Disease Prediction | Kaggle](https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction)
2. YouTube. (n.d.). YouTube. Retrieved from StatQuest: [Logistic Regression - YouTube](https://www.youtube.com/watch?v=yIYKR4sgzI8), [Random Forest Algorithm Clearly Explained! - YouTube](https://www.youtube.com/watch?v=v6VJ2RO66Ag), [Support Vector Machines Part 1 (of 3): Main Ideas!!! - YouTube](https://www.youtube.com/watch?v=yIYKR4sgzI8), [K Nearest Neighbors | Intuitive explained | Machine Learning Basics - YouTube](https://www.youtube.com/watch?v=0p0o5cmgLdE)
3. Stack Overflow. (n.d.). Stack Overflow. Retrieved from <https://stackoverflow.com/>
4. Google Colab. (n.d.). Google Colab. Retrieved from <https://colab.research.google.com/>
5. Geeks For Geeks (n.d). GFG retrieved from [Random Forest Classifier using Scikit-learn - GeeksforGeeks](https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/), [Introduction to Support Vector Machines (SVM) - GeeksforGeeks](https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/), [Logistic Regression in Machine Learning - GeeksforGeeks](https://www.geeksforgeeks.org/understanding-logistic-regression/), [K-Nearest Neighbor(KNN) Algorithm - GeeksforGeeks](https://www.geeksforgeeks.org/k-nearest-neighbours/)

Please note that the references provided for ChatGPT, Kaggle dataset, YouTube, and Stack Overflow

are general references to the platforms themselves as they are widely recognized sources for AI models,

datasets, and general knowledge. Specific citations and acknowledgments for any external sources or

research papers used during the project are not included in this reference list.