**Disease Prediction Using Machine Learning**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**Chetan Sasidhar Chunduru AP21110010510**

**Sathwik Batta AP21110011208**

**Naveen Chowdary AP21110010470**

**Nitin Reddy AP21110010502**

**Lokesh Kota AP21110010466**

**A picture containing text

Description automatically generated**

Under the Guidance of

**(Dr. Sowmyajyothi Biswas)**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh**

**May, 2024**

This project stands out due to its innovative approach to disease diagnosis, utilizing a combination of machine learning algorithms. By incorporating various algorithms such as Support Vector Machine (SVM), Naive Bayes, and Random Forest, it explores different avenues for predicting common chronic illnesses based on patient symptoms and profiles.

The project is divided into two main branches: one focuses on symptom-based prediction using SVM, Naive Bayes, and Random Forest classifiers, while the other relies on clinical measurements. In the latter, SVM is employed for Parkinson's disease diagnosis, while logistic regression is utilized for predicting heart disease. The inclusion of real-world patient data enhances the authenticity and practicality of the analysis, making it relevant to clinical settings.

Furthermore, the project stands out for its emphasis on evaluating the performance of various models, identifying their strengths and weaknesses, and offering actionable insights for healthcare professionals. Overall, its novelty lies in its comprehensive strategy to tackle the challenge of disease diagnosis by integrating machine learning techniques across diverse domains.

**Abstract**

Disease diagnosis is a crucial aspect of healthcare, but conventional methods often rely on manual examination and may not consistently deliver accurate results. In our project, we propose a novel machine learning-based system to address the limitations of traditional diagnostic approaches. By harnessing Support Vector Machine (SVM), Naive Bayes, and Random Forest algorithms, we aim to automate the identification of common chronic illnesses based on patient symptoms and profiles.

Our dataset comprises symptom data collected from patients, alongside their corresponding diagnoses. We preprocess the data and train machine learning models to predict the probability of specific diseases given a set of symptoms. We introduce two distinct branches within our project: one focuses on symptom-based prediction using SVM, Naive Bayes, and Random Forest classifiers, while the other utilizes clinical measurements. Specifically, we employ SVM for Parkinson's disease diagnosis and logistic regression for heart disease prediction. The incorporation of real-world patient data enhances the authenticity and practicality of our analysis, ensuring relevance to clinical settings.

Furthermore, we evaluate the performance of each model using key metrics such as accuracy, precision, and recall. Overall, our project contributes to advancing disease diagnosis by demonstrating the effectiveness of machine learning techniques in automating the identification of common chronic illnesses. By providing a reliable and efficient diagnostic tool, our system has the potential to improve healthcare outcomes and enhance patient care.

**Introduction:**

In the realm of modern healthcare, disease diagnosis stands as a crucial pillar of patient care, where swift and precise identification holds the key to effective treatment and management. However, reliance on traditional diagnostic methods, often reliant on manual examination, can result in inaccuracies and delays in treatment, potentially leading to misdiagnoses. To confront these challenges, the integration of machine learning techniques has emerged as a promising avenue, offering automation and enhanced accuracy in disease diagnosis.

This project embarks on a pioneering endeavor, leveraging machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes, and Random Forest. The goal is to craft a sophisticated disease diagnosis system adept at accurately discerning common chronic illnesses based on patient symptoms and clinical measurements. By harnessing the collective power of these algorithms, our objective is to streamline the diagnostic process, thereby enhancing healthcare outcomes and bolstering patient care.

The impetus driving this project arises from the necessity to address the limitations of traditional diagnostic methods and introduce a more efficient and dependable approach to disease diagnosis. By delving into the potential of machine learning algorithms within this context, we aim to propel healthcare forward and elevate the standard of patient care.

Through an exhaustive exploration of machine learning techniques and their application in disease diagnosis, this project endeavors to offer valuable insights into the transformative potential of automated diagnostic systems in healthcare delivery. By crafting a robust and precise disease diagnosis system, our aim is to equip healthcare professionals with the necessary tools to make well-informed decisions and deliver timely interventions for patients.

Highlights of this project:

1. **In-Depth Exploration:** This project undertakes a thorough exploration of machine learning algorithms, including SVM, Naive Bayes, Random Forest, and logistic regression, to analyze their effectiveness in identifying common chronic illnesses.

2. **Automation of Diagnosis:** Through the utilization of machine learning techniques, the project aims to automate the disease diagnosis process, reducing reliance on manual examination and enhancing diagnostic accuracy.

3. **Enhanced Healthcare Outcomes:** The development of a machine learning-based disease diagnosis system holds promise for significantly improving healthcare outcomes by facilitating early and precise identification of common chronic illnesses.

4. **Symptom-based and Clinical Measurements Analysis:** The project adopts two distinct approaches to disease diagnosis. Firstly, it employs SVM, Naive Bayes, and Random Forest classifiers for symptom-based analysis, leveraging patient-reported symptoms to predict disease likelihood. Secondly, it utilizes Support Vector Machine for Parkinson's disease diagnosis and logistic regression for heart disease diagnosis based on clinical measurements, ensuring a comprehensive evaluation of patient health.

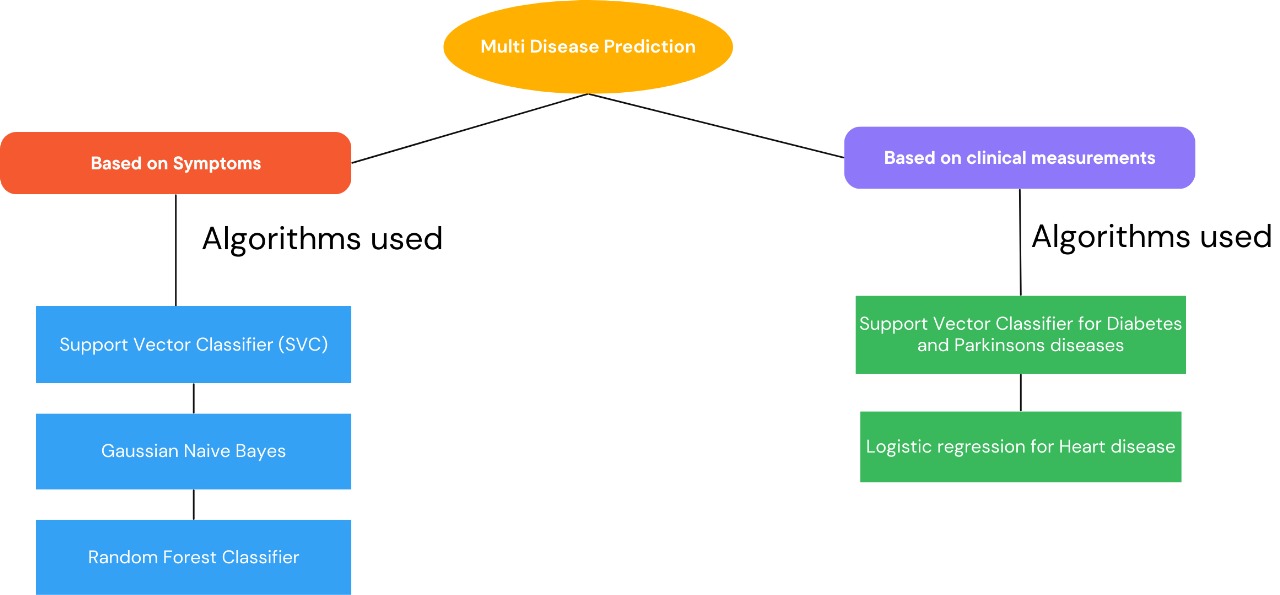
**Background:**Disease diagnosis stands as a crucial element in healthcare, where the precise and timely identification of illnesses profoundly impacts patient outcomes. Yet, conventional diagnostic methods often rely on manual assessments and symptom-based evaluations, which can be time-consuming and prone to inaccuracies. Moreover, the intricate nature and variability of diseases further complicate accurate diagnosis by healthcare professionals.

In recent years, the emergence of machine learning has transformed disease diagnosis by utilizing sophisticated algorithms to analyze patient data and recognize patterns indicative of various illnesses. By harnessing extensive datasets containing clinical measurements, symptoms, and patient profiles, machine learning models offer invaluable support to healthcare providers in making more precise and efficient diagnoses.

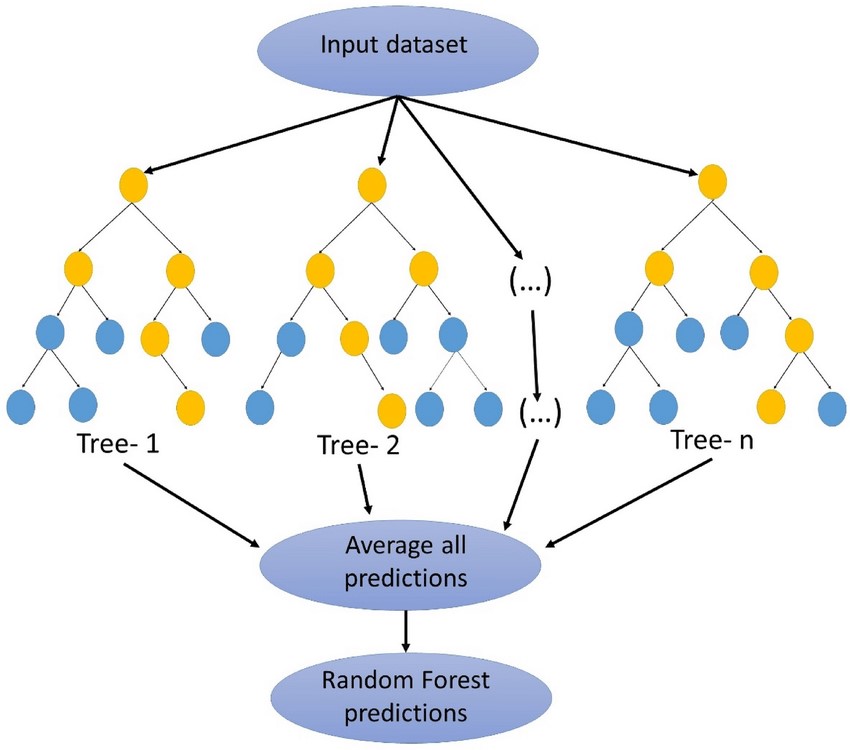
The objective of this project is to delve into the application of machine learning techniques in disease diagnosis and assess their efficacy in predicting and identifying common illnesses. Specifically, we aim to evaluate the performance of various machine learning models, including Support Vector Machine (SVM), Naive Bayes, Random Forest, and logistic regression. These models will be tasked with diagnosing diseases based on patient symptoms and clinical measurements.

**Methodology:**  
  
1.**Data Collection:** Gather a dataset containing information about symptoms and corresponding diseases. This dataset can be sourced from reputable medical sources or collected through surveys and patient records.2.**Data Preprocessing**: Clean the dataset by handling missing values, removing duplicates, and standardizing data formats. This step may also involve encoding categorical variables and normalizing numerical features.3.**Model Selection:** Choose appropriate machine learning algorithms for disease prediction. Consider algorithms like Support Vector Machine (SVM), Naive Bayes, and Random Forest, logistic regression ,which are commonly used for classification tasks.4.**Model Training:** Train the selected models using the preprocessed dataset. Split the data into training and validation sets to evaluate model performance during training.5.**Model Evaluation:** Assess the performance of trained models using evaluation metrics such as accuracy, precision, recall, and F1-score. Use techniques like cross-validation to ensure robustness and generalization of the models.

6.**Deployment:**

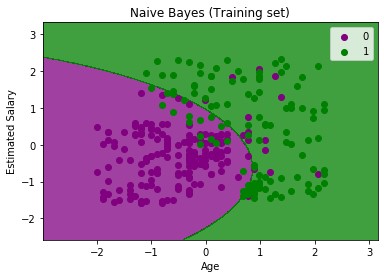


**Random forest algorithm:**  
  
In this project, Random Forest serves as a potent tool to enhance the accuracy and reliability of disease prediction. As a supervised learning algorithm, Random Forest harnesses the collective power of multiple decision trees, leveraging their combined strength to make precise predictions. The process begins with the partitioning of the dataset through random sampling of both rows and features, where subsets of data are randomly selected and allocated to individual decision trees. This random sampling, conducted with replacement, ensures diversity among the trees, guarding against the influence of specific patterns or outliers in the data.Each decision tree undergoes independent training on its subset of data to classify patients into different disease categories based on their symptoms and other relevant features. By considering various subsets of characteristics and data points, each tree uncovers unique patterns and correlations within the dataset. Following training, a majority voting mechanism enables the ensemble of decision trees to collectively contribute to the prediction process. This approach enhances the overall accuracy of the model and reduces the likelihood of individual errors by aggregating predictions from multiple trees.One of the notable advantages of Random Forest is its adeptness at handling high-dimensional data and intricate relationships effectively. By amalgamating the predictions of numerous decision trees, Random Forest can capture diverse patterns and anomalies present in the dataset, facilitating accurate disease predictions. Moreover, the ensemble nature of Random Forest mitigates the risk of overfitting, ensuring robust generalization to unseen data.In summary, Random Forest emerges as a robust and versatile tool for disease diagnosis, offering improved accuracy and reliability compared to individual decision trees or alternative machine learning algorithms. Its capacity to harness the collective intelligence of multiple trees makes it well-suited for identifying diseases accurately amidst the complexity of real-world medical data.



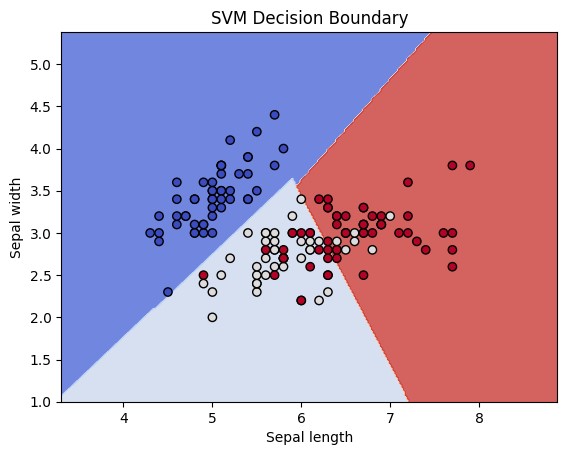
**Naive Bayes:**The Naive Bayes algorithm is a simple yet effective probabilistic classifier utilized in this project for disease diagnosis. Despite its simplistic assumptions, Naive Bayes often performs remarkably well in practice, particularly in situations with limited training data and high-dimensional feature spaces.In disease diagnosis, Naive Bayes operates by applying Bayes' theorem to calculate the probability that a patient has a particular disease given their observed symptoms. It assumes that all features (symptoms) are conditionally independent given the class label (disease), which allows for efficient computation of the posterior probability using the joint probability distribution of the features.Despite its "naive" assumption of feature independence, Naive Bayes can still produce reliable predictions, especially when applied to datasets with categorical or binary features. It is particularly well-suited for tasks involving text classification or medical diagnosis, where the presence or absence of specific symptoms contributes to the likelihood of a particular disease.Naive Bayes offers several advantages, including its simplicity, scalability, and fast training speed. It requires minimal tuning of hyperparameters and can handle missing data gracefully. Moreover, Naive Bayes performs well even with small training datasets and is robust to irrelevant features.

In summary, the Naive Bayes algorithm utilizes Bayes' theorem and the assumption of feature independence to probabilistically predict the likelihood of various classes (in this case, diseases) based on the observed features (such as symptoms).



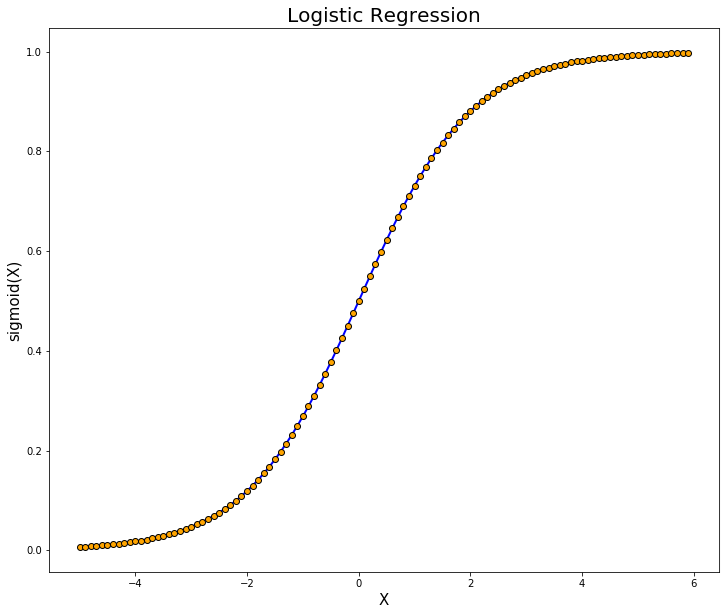
**Support Vector Machine**

The Support Vector Machine (SVM) algorithm is a powerful supervised learning technique employed in this project for disease diagnosis. SVM is particularly well-suited for binary classification tasks, where it aims to find the optimal hyperplane that separates data points belonging to different classes with the maximum margin.In the context of disease diagnosis, SVM works by mapping patient symptoms to a high-dimensional feature space and identifying the optimal boundary (hyperplane) that best separates patients with different diseases. By maximizing the margin between classes, SVM can effectively classify patients into distinct disease categories based on their symptom profiles.One of the key strengths of SVM is its ability to handle non-linear relationships between symptoms and diseases through the use of kernel functions, such as radial basis function (RBF) or polynomial kernels. These kernels enable SVM to capture complex patterns in the data and achieve high classification accuracy.Moreover, SVM exhibits robustness to overfitting and performs well in scenarios with high-dimensional feature spaces, making it suitable for disease diagnosis tasks involving a large number of symptoms and patients.



**Logistic regression**

Logistic Regression, a supervised learning algorithm, plays a vital role in our project's methodology by offering a contrasting approach to disease diagnosis. Unlike other algorithms such as Support Vector Machine (SVM) and Random Forest, Logistic Regression relies on labeled data and is proficient in modeling the probability of a binary outcome.In our research, Logistic Regression is applied to predict, based on available patient data, the probability of a particular disease occurrence. The algorithm works by fitting the input data to a logistic function, which correlates the input features to the likelihood of disease presence.Several key steps are involved in the Logistic Regression approach. Firstly, the dataset is split into training and testing sets using the train-test split technique to ensure the model's generalizability. Subsequently, the training data is utilized to train the Logistic Regression model, enabling it to predict the likelihood of disease occurrence based on patient profiles.During the training phase, the logistic loss function is minimized by iteratively optimizing the model's parameters, typically using techniques like gradient descent. Following training, the model's accuracy in classifying disease occurrences is evaluated using the testing data.Evaluation metrics such as accuracy, precision, recall, and F1-score are commonly utilized to assess the Logistic Regression model's effectiveness. These metrics provide valuable insights into the model's predictive power and its ability to accurately identify disease occurrences while minimizing false diagnoses.



**Predicted Diseases:**

1.(vertigo) Paroymsal Positional Vertigo 40. veinshepatitis A

2.AIDS 41. Parkinson’s disease

3.Acne

4.Alcoholic hepatiti

5.Allerg

6.Arthritis

7.Bronchial Asthma

8.Cervical spondylosis

9. Chicken pox

10.Chronic cholestasis

11.Common Cold

12.Dengue

13.Diabetes

14.Dimorphic hemmorhoids(piles)

15.Drug Reaction

16.Fungal infection

17.GERD

18.Gastroenteritis

19.Heart attack

20.Hepatitis B

21.Hepatitis C

22.Hepatitis D

23.Hepatitis E

24.Hypertension

25.Hyperthyroidism

26.Hypoglycemia

27.Hypothyroidism

28.Impetigo

29.Jaundice

30.Malaria

31.Migraine

32.Osteoarthristis

33.Paralysis (brain hemorrhage)

34.Peptic ulcer diseae

35.PneumoniaPsoriasis

36.Tuberculosis

37.Typhoid

38.Urinary tract infection

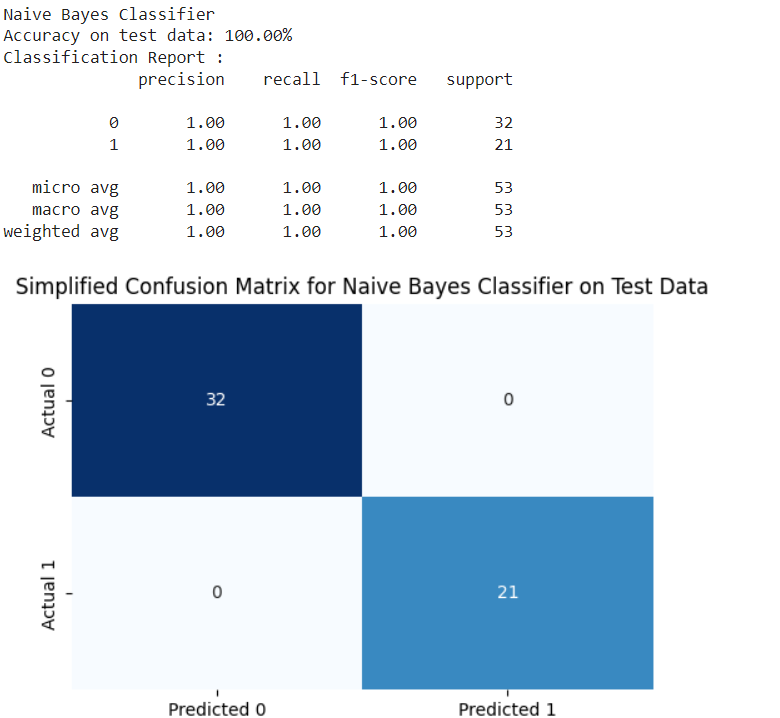
39.Varicose Veins

**Disease Prediction Based On Symptoms**

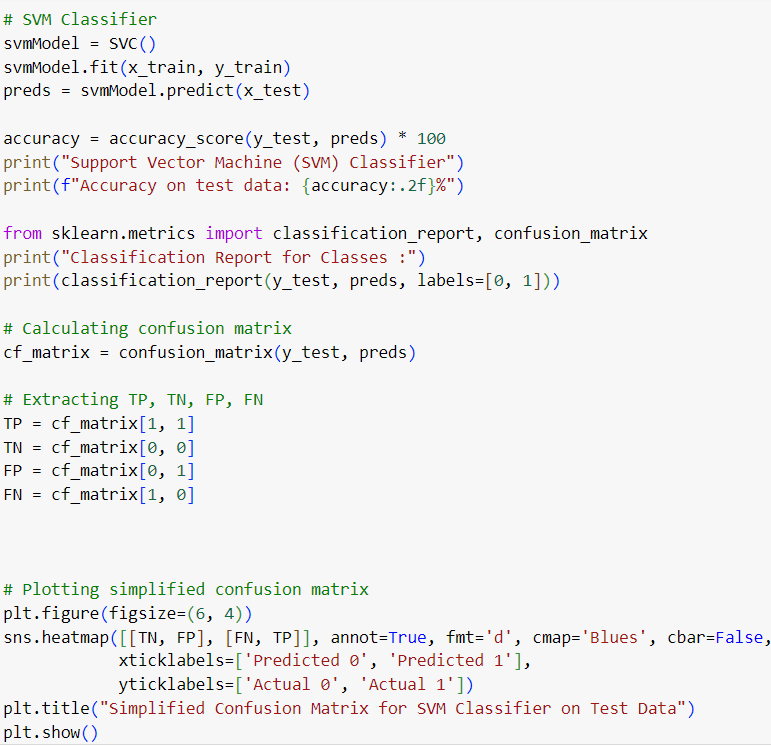
**Code and Result:**

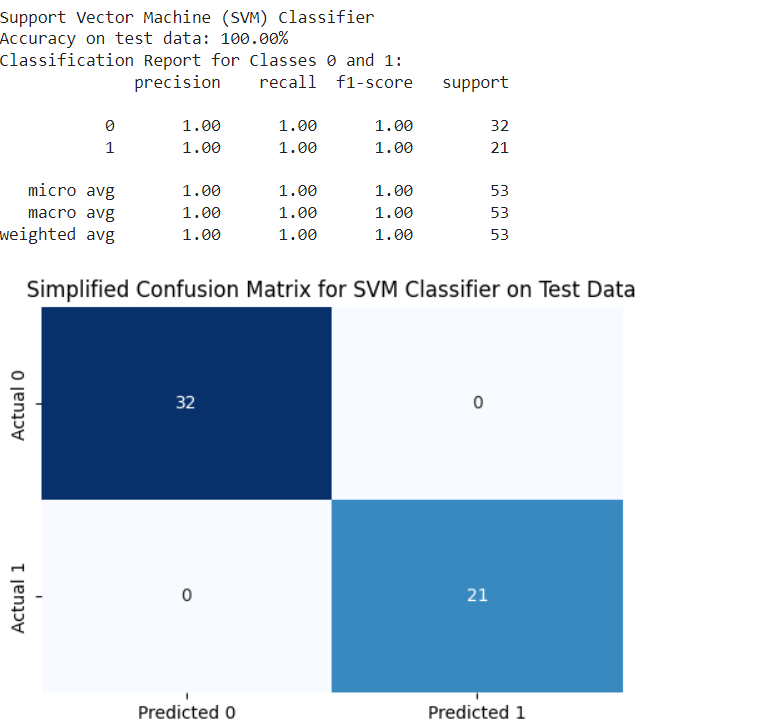
**Naive Bayes Algorithm:**



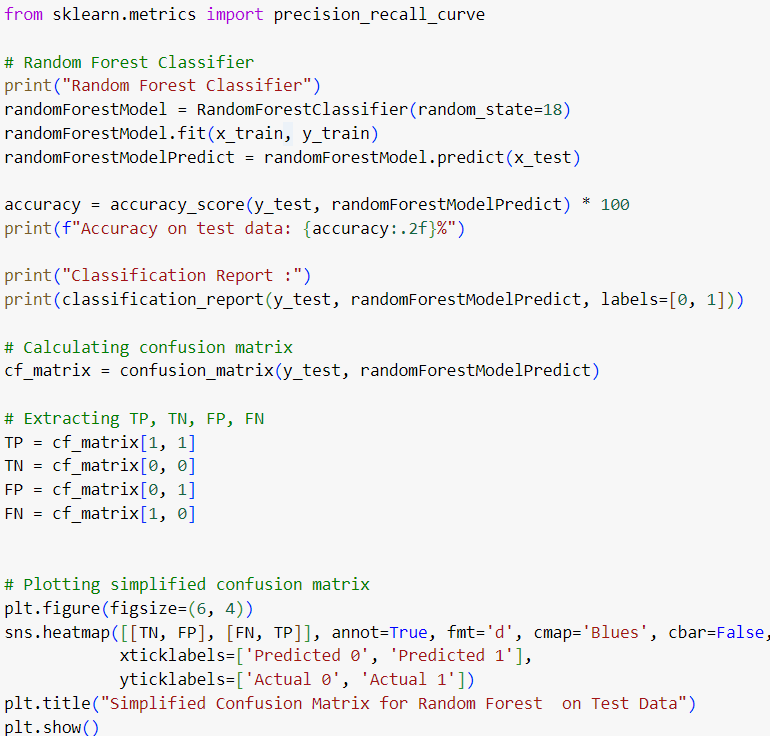


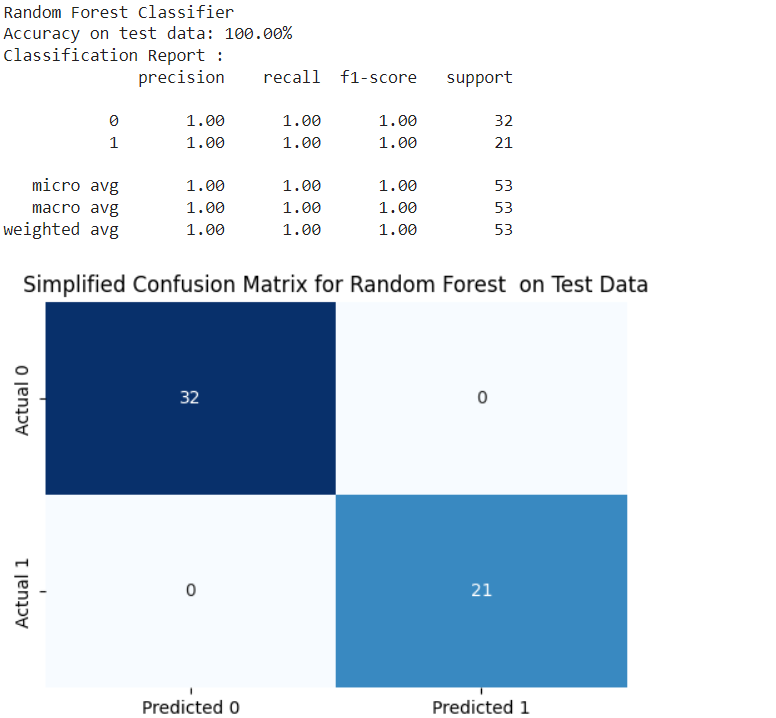
**Support Vector Machine**

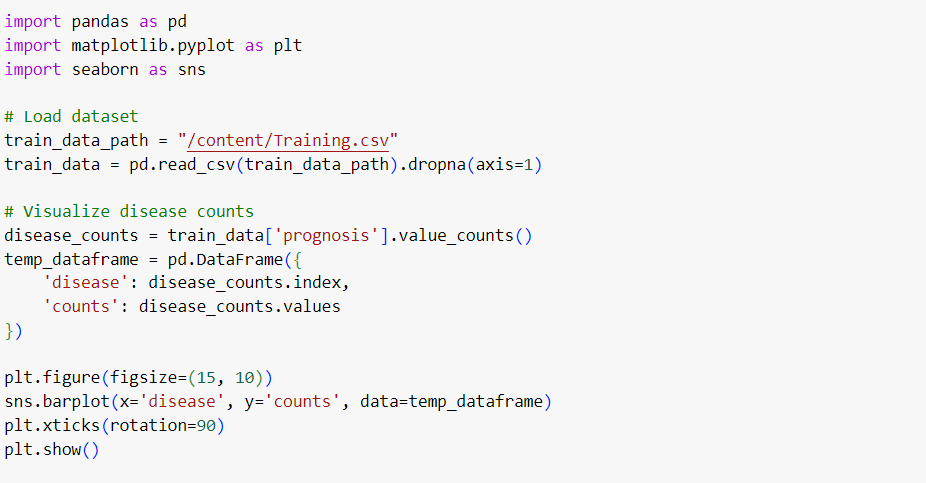




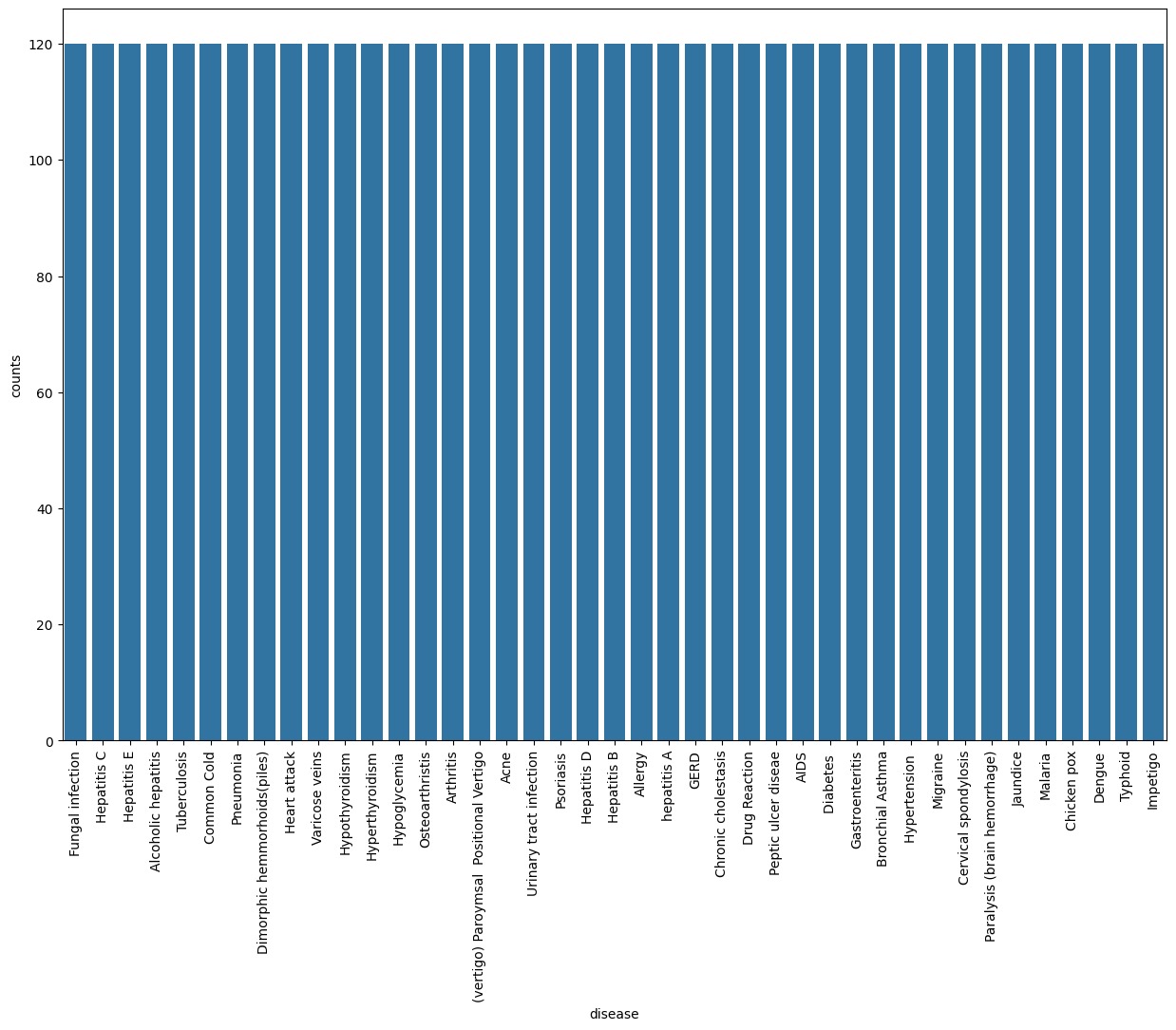
**Random Forest Algorithm**







**Counts of Each Disease in the Dataset**

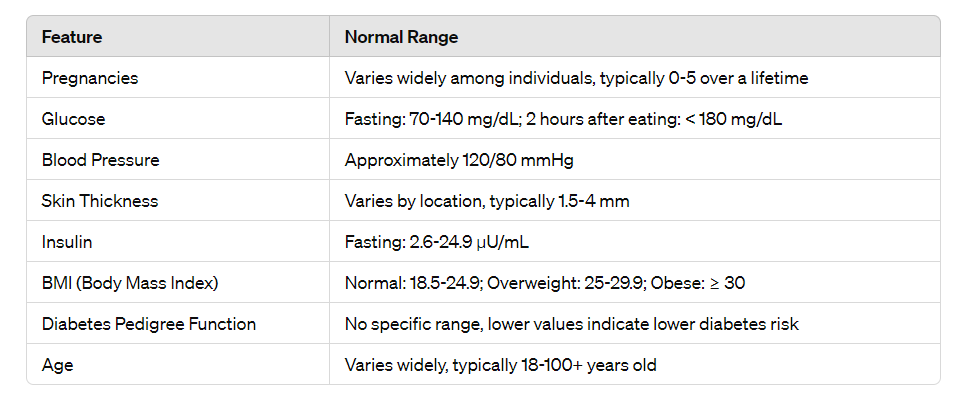


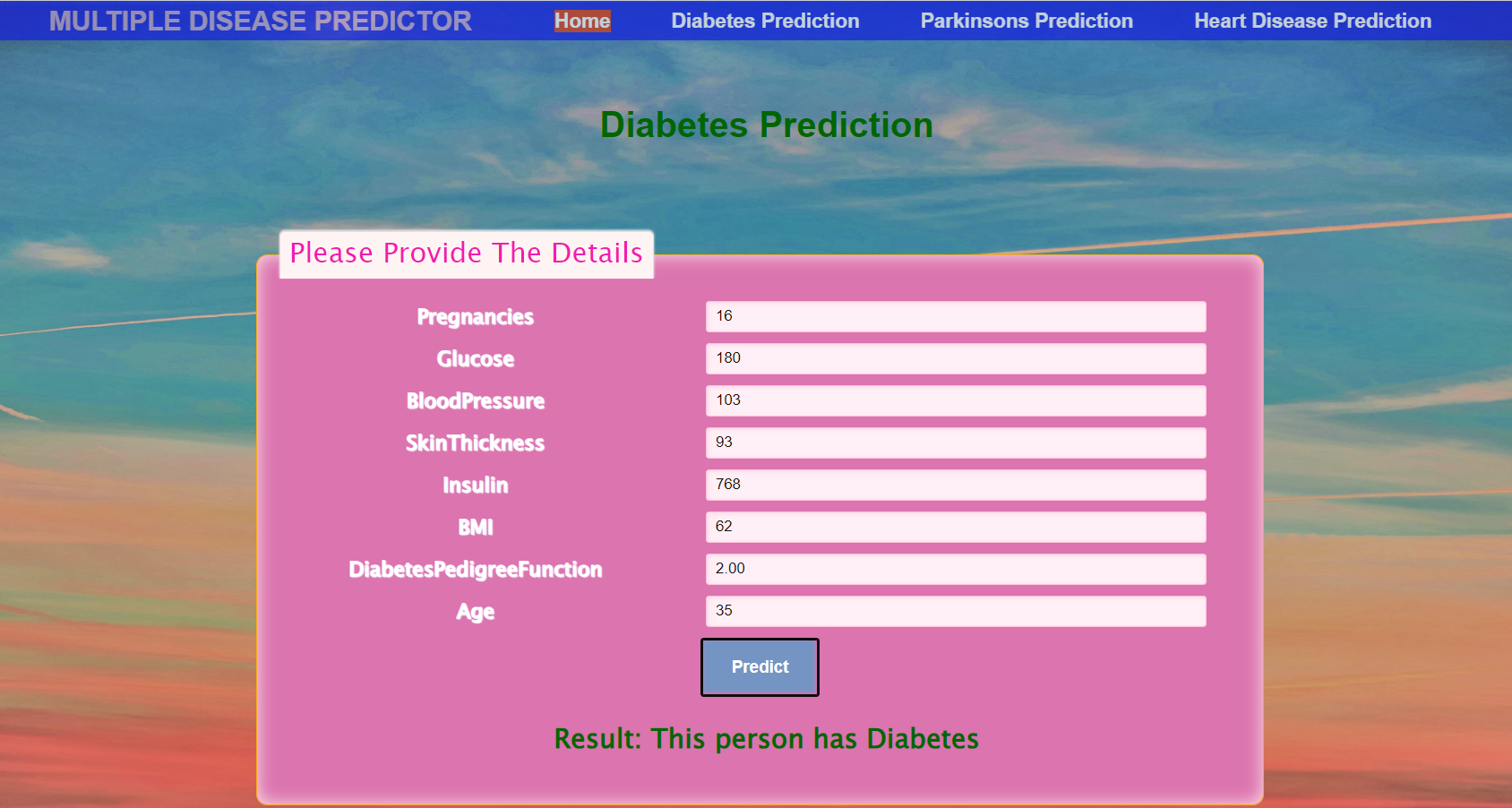
**Final Output:**



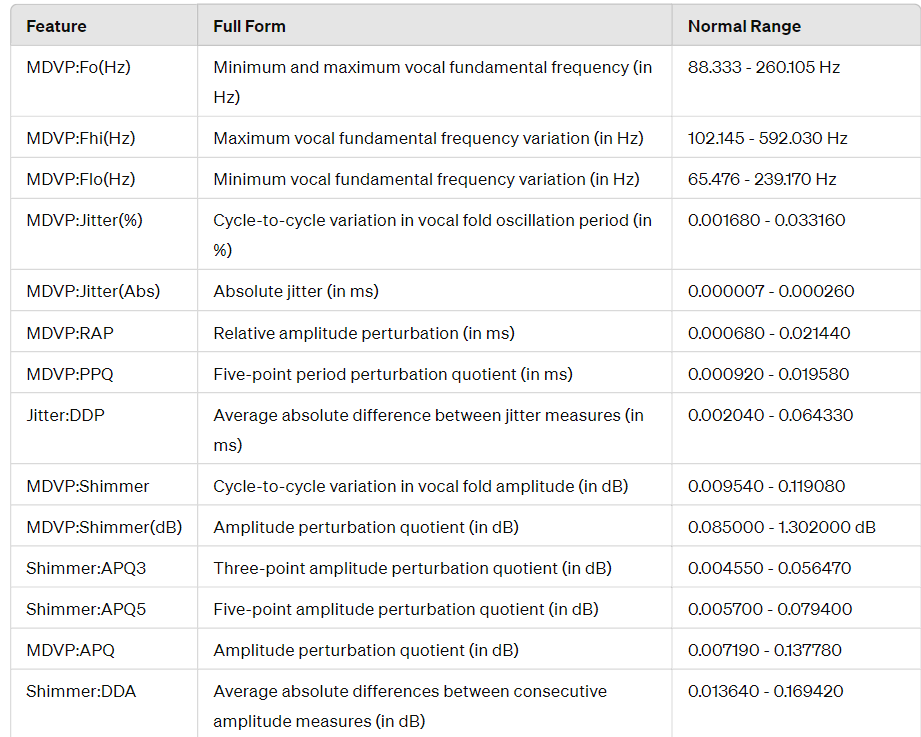
**Disease Prediction Based On Clinical Measurements**

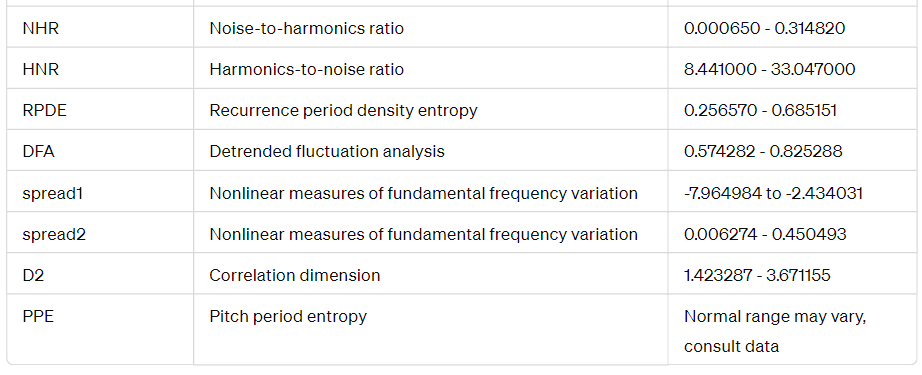
**Normal Range Table For Diabetes Prediction:**

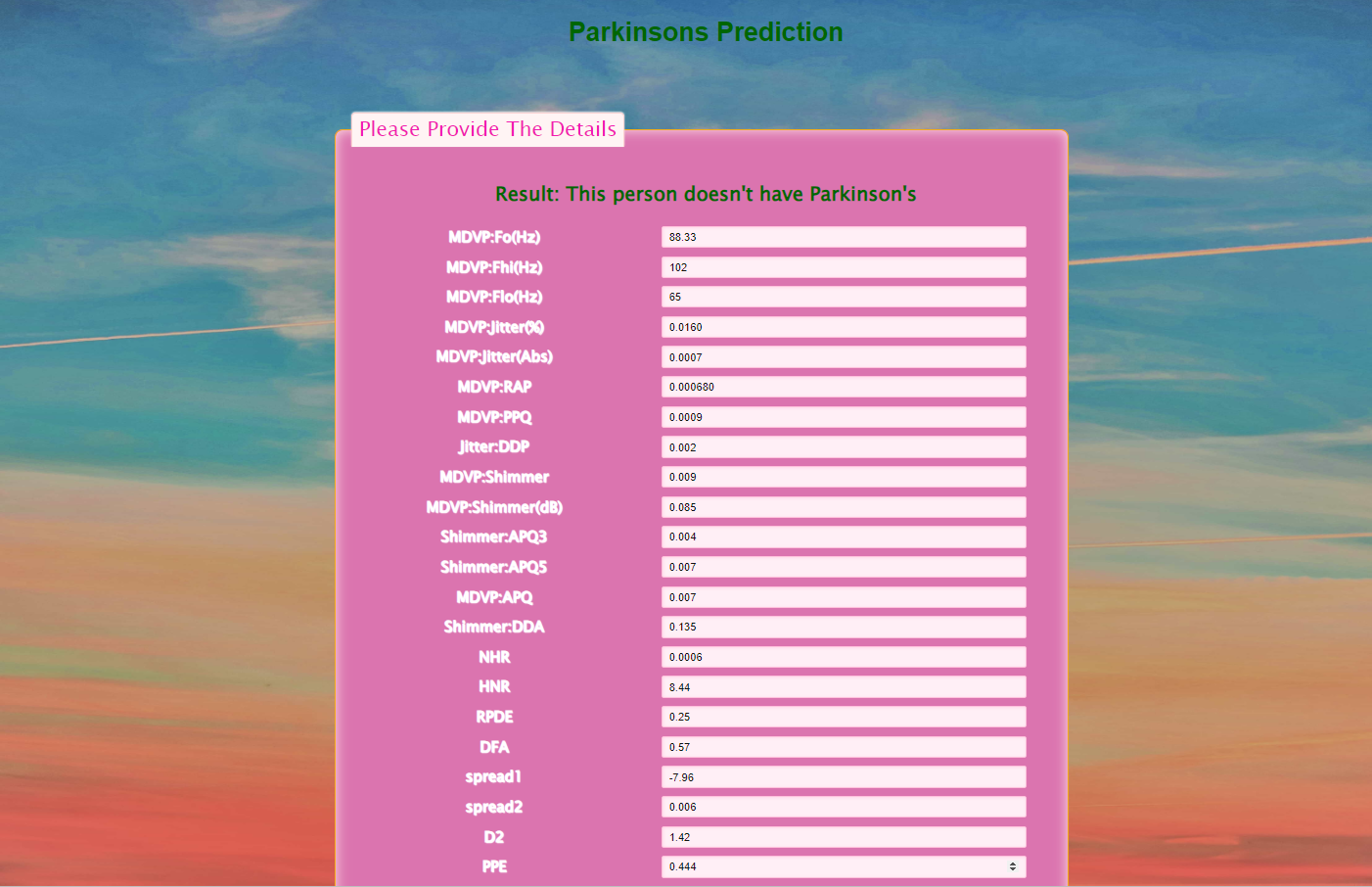




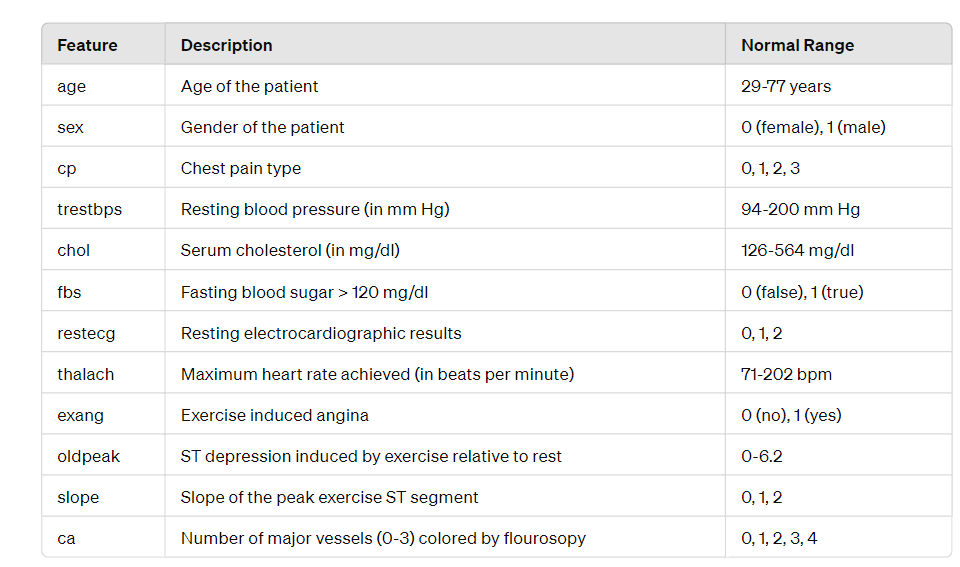
**Normal Range Table For Parkinson’s Prediction:**

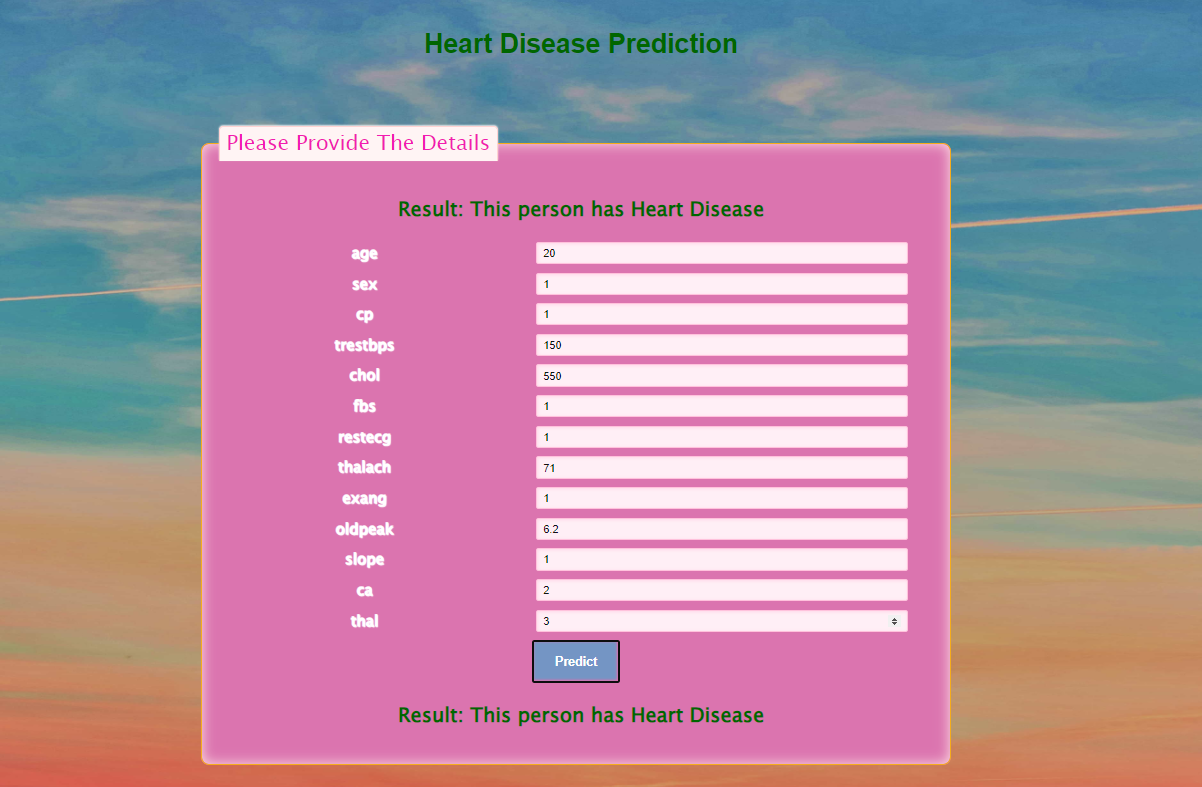






**Normal Range Table For Heart Disease Prediction:**





**Conclusion:**

In our project, we delve into disease diagnosis, leveraging machine learning algorithms to enhance the precision and effectiveness of this critical process. We focus on two primary aspects: symptom-based diagnosis and diagnosis based on clinical measurements.

For symptom-based diagnosis, we investigate the efficacy of Support Vector Machine (SVM), Naive Bayes, and Random Forest algorithms. Our goal is to develop a robust disease diagnosis system capable of accurately identifying common chronic illnesses by analyzing patient symptoms and profiles.

Simultaneously, we explore clinical measurement-based diagnosis, specifically targeting Parkinson's disease and heart disease. In Parkinson's disease diagnosis, we utilize Support Vector Machine (SVM) to predict the likelihood of disease occurrence based on clinical measurements. Conversely, for heart disease diagnosis, Logistic Regression is employed to assess the risk of cardiovascular ailments using clinical measurements.

Through meticulous analysis of these machine learning algorithms, we aim to provide valuable insights into their effectiveness in disease diagnosis. By evaluating metrics like accuracy, precision, and recall, we aim to pinpoint the most suitable algorithms for each diagnostic scenario.

Our overarching objective is to develop dependable and efficient disease diagnosis systems that empower healthcare professionals with the necessary tools to make informed decisions and offer timely interventions for patients. By integrating machine learning techniques into disease diagnosis, we aspire to revolutionize healthcare delivery and elevate patient outcomes.

**Future Scope:**

1. **Real-time Disease Monitoring Systems:** Create systems for real-time disease monitoring that continuously analyze patient data from wearable devices, electronic health records, and other sources to detect early warning signs of diseases and offer timely interventions.

2. **Personalized Medicine:** Embrace precision medicine principles, which involve tailoring treatment strategies to individual patients based on their unique genetic makeup, lifestyle factors, and environmental influences. Machine learning algorithms can aid in identifying optimal treatment plans and predicting patient responses to specific therapies.

3. **Automated Disease Diagnosis Using Cell Images:** A promising area in healthcare's future is the development of automated disease diagnosis systems from cell images. These systems, utilizing advanced image analysis techniques and machine learning algorithms, hold the potential to transform healthcare by enabling early detection, precise diagnosis, and personalized treatment strategies.