**Mini Project Report on**

https://lh5.googleusercontent.com/NOqhtVxb-aL_dmUUrDMvFY0BKkccUe-gp-nBooOZVVAemuyLgEeqDorF6cjtKem19EYdv4631A0QdGXiZ8rxH6kXTas1e91cBlcFwiCFMworKblxHX6NJdwaI4tKpoF5hxP6rOW2Eaa0r0Fq4GLnThcbBMle1eH3RuKlLjPA5HJtghqrAzgz65nArjDO8a6y73-hDAWOhw

**DISEASE PREDICTION USING DEEP LEARNING**  
https://lh5.googleusercontent.com/NOqhtVxb-aL_dmUUrDMvFY0BKkccUe-gp-nBooOZVVAemuyLgEeqDorF6cjtKem19EYdv4631A0QdGXiZ8rxH6kXTas1e91cBlcFwiCFMworKblxHX6NJdwaI4tKpoF5hxP6rOW2Eaa0r0Fq4GLnThcbBMle1eH3RuKlLjPA5HJtghqrAzgz65nArjDO8a6y73-hDAWOhw

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

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:ANUSHKA KAUSHIK** **University Roll No. 2016658**

***Under the Mentorship of***

**Dr. Manoj Diwaker**

**Associate Professor**



**Department of Computer Science and Engineering**

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

**Dehradun, Uttarakhand**

**July 2023**



**DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Deep Learning Model For Disease Prediction”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **DR. Manoj Diwaker , Asst. Professor** , Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

ANUSHKA KAUSHIK   2016658

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 4 |
| Chapter 2 | Literature Survey | 5-6 |
| Chapter 3 | Methodology | 7-9 |
| Chapter 4 | Result and Discussion | 10 |
| Chapter 5 | Conclusion and Future Work | 11 |
|  | References | 12 |

**Chapter 1**

**Introduction**

Disease Prediction using Machine Learning is an advanced system developed to predict diseases based on symptoms provided by patients or users. By analyzing the user-provided symptoms, the system calculates the probability of various diseases, employing the Naïve Bayes classifier as the primary supervised machine learning algorithm for disease prediction.

In the field of healthcare, accurate analysis of medical data plays a critical role in early disease detection and optimal patient care. With the exponential growth of biomedical and healthcare data, it has become essential to leverage machine learning techniques to effectively handle and interpret this wealth of information. Our project addresses this need by utilizing machine learning algorithms, including linear regression and decision tree, to predict diseases such as Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Machine Learning leverages historical data and experiences to make accurate predictions. The disease prediction process involves two primary phases: Training and Testing. Through the power of machine learning, our aim is to predict diseases based on a patient's symptoms and medical history. This ongoing endeavor in the medical field strives to continuously improve disease prediction accuracy and ultimately enhance healthcare outcomes.

By implementing machine learning models, we are able to efficiently analyze and process large volumes of healthcare data, resulting in faster and more accurate disease predictions. This empowers doctors and healthcare professionals to make well-informed decisions regarding patient diagnoses, leading to improved healthcare services and better patient care overall. Our integration of machine learning into the medical field serves as a prime example of its potential to revolutionize healthcare practices.

In our project, we specifically focus on improving the accuracy of disease prediction, particularly for unstructured or textual data. To achieve this, we employ a range of algorithms, including linear regression, K-nearest neighbors (KNN), and decision tree. These algorithms have been extensively studied and widely used for disease prediction tasks, providing a robust foundation for our system.

By harnessing the power of machine learning technology and leveraging comprehensive patient data, we strive to revolutionize disease prediction methodologies and contribute to the advancement of healthcare practices. Our project holds great potential for enhancing patient care, improving early disease detection, and ultimately transforming the way diseases are diagnosed and managed.

**Chapter 2**

**Literature Survey**

In this chapter, a review of major existing work in the field of disease detection and healthcare utilizing computer-based technology and supervised machine learning algorithms has been conducted. The widespread adoption of computer technology in the healthcare industry has led to the accumulation of electronic data, presenting both challenges and opportunities for medical professionals in accurately analyzing symptoms and detecting diseases at an early stage.

1. Supervised machine learning algorithms have shown promising potential in surpassing traditional diagnostic methods, aiding in disease diagnosis, and enabling early detection of high-risk diseases. This literature aims to identify trends across various types of supervised machine learning models used in disease detection by examining performance metrics. The most prominently discussed algorithms include Naïve Bayes (NB), Decision Trees (DT), and K-Nearest Neighbor (KNN). Additionally, Support Vector Machine (SVM) has been found to be highly effective in detecting kidney diseases and Parkinson's disease, while Logistic Regression (LR) excels in predicting heart diseases. Random Forest (RF) and Convolutional Neural Networks (CNN) have demonstrated precision in predicting breast diseases and common diseases, respectively.
2. The integration of computer technology, particularly machine learning algorithms, into the healthcare industry holds great potential. This project aims to detect diseases at the earliest stage and apply necessary treatments using machine learning algorithms and hospital-generated datasets. To the best of our knowledge, there is no prior work that addresses both types of data in large-scale medical data analysis. Our proposed algorithm exhibits a calculation accuracy of 94.8% and faster convergence speed compared to other typical estimation algorithms, particularly in CNN-based unimodal disease prediction.
3. The abundance of electronic data in healthcare poses challenges for accurate diagnosis and early disease detection. Misdiagnosis can have serious consequences, leading to ineffective treatment or delayed diagnosis of critical conditions. However, supervised machine learning techniques have shown promise in surpassing conventional diagnostic methods, particularly in detecting high-risk diseases. Our system addresses the issue of people hesitating to visit a hospital for minor complications by providing online medical advice. By evaluating the symptoms provided by individuals as input, the system outputs the predicted disease. The Naïve Bayes classifier is employed in the system, and accuracy is emphasized, with a higher number of symptoms input yielding better disease predictions.
4. The work conducted in this field has the potential to significantly enhance the healthcare industry, leading to improved diagnoses and treatments. By leveraging the power of supervised machine learning algorithms and comprehensive datasets, we aim to advance healthcare practices and contribute to the overall well-being of individuals worldwide.

**Chapter 3**

**Methodology**

In the provided code, three different algorithms are being used for disease prediction: Decision Tree, Random Forest, and Naïve Bayes. The code implements a web application using Streamlit framework, allowing users to input their data and obtain disease predictions based on the selected algorithm.

The methodology can be summarized as follows:

1. **Data Loading:**
   1. The code assumes that pre-trained models for diabetes prediction, heart disease prediction, and Parkinson's disease prediction are stored as pickled files (**.sav** format). These models are loaded into the application using the **pickle** library.
2. **Streamlit Interface:** 
   1. The code creates a sidebar navigation menu using the **streamlit\_option\_menu** library, allowing users to select the disease prediction they want to perform.
3. **Diabetes Prediction:**
   1. If the user selects "Diabetes Prediction," the application prompts the user to enter specific input data related to pregnancies, glucose level, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and age. After the user clicks the "Diabetes Test Result" button, the loaded diabetes model is used to predict whether the person is diabetic or not.
4. **Heart Disease Prediction:**
   1. If the user selects "Heart Disease Prediction," the application prompts the user to enter input data related to various factors such as age, sex, chest pain types, resting blood pressure, serum cholestoral, fasting blood sugar, and other parameters. After clicking the "Heart Disease Test Result" button, the loaded heart disease model predicts whether the person has heart disease or not.
5. **Parkinson's Disease Prediction:**
   1. If the user selects "Parkinson's Prediction," the application prompts the user to enter input data related to various features associated with Parkinson's disease. After clicking the "Parkinson's Test Result" button, the loaded Parkinson's disease model predicts whether the person has Parkinson's disease or not.

The code uses the respective pre-trained models for each disease prediction task and applies the selected algorithm to make predictions based on the user-provided input data. The results are displayed using the **st.success()** function, indicating the diagnosis of the disease based on the model's prediction.

It is important to note that the code assumes the availability of the pre-trained models and relevant input data for accurate predictions.

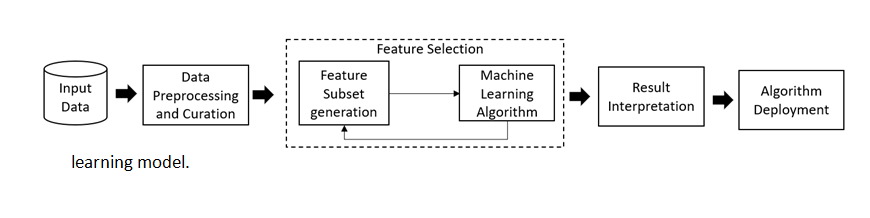


Fig 3.1 Algorithm

**Algorithms:**

**1. Decision Tree:**

The Decision Tree algorithm is a powerful supervised learning technique commonly used for solving classification problems. It can handle both categorical and continuous dependent variables, making it versatile in various applications. This algorithm divides the population into homogeneous subgroups based on the most influential attributes or features, aiming to create distinct groups as much as possible. In practical terms, a decision tree represents decisions and their outcomes visually and explicitly, making it a popular choice for decision analysis and machine learning tasks.

The Decision Tree algorithm finds wide application in areas such as data mining and machine learning, as it provides a structured model for achieving specific objectives. It is frequently employed to generate plans and make predictions based on the learned patterns in the data. For instance, in our project, we utilize the Decision Tree classifier to predict whether a balancing scale will tip to the right, left, or remain balanced, once the model is trained.

The formula used in Decision Tree classification involves calculating the probability (P) based on the number of instances (nA) belonging to a specific class divided by the total number of instances (nA + nB). This formula (Eq(1)) allows us to make predictions and classify new data points based on the learned patterns from the training dataset.

By leveraging the Decision Tree algorithm, we can gain insights from complex datasets and make accurate predictions for classification tasks. Its visual representation of decision-making processes facilitates interpretability and aids in understanding the underlying patterns and relationships within the data.

Formulae: P=nA/(nA+nB) P=nA/(nA+nB)……………………………………Eq(1)

**2. Random Forest:**

The Random Forest algorithm is an excellent choice for early-stage model training, as it provides a reliable performance assessment and is relatively straightforward to implement. Its simplicity makes it difficult to create a "poor" Random Forest model, making it a popular option when time constraints are a factor. Additionally, Random Forests offer a clear indication of the importance assigned to different features within the model.

One of the notable advantages of Random Forests is their versatility in handling various types of features, including binary, categorical, and numerical. This flexibility enables the algorithm to effectively analyze datasets with diverse feature characteristics. Random Forests excel in achieving high performance, making them a valuable tool in classification, regression, and other machine learning tasks.

Random Forests operate through an ensemble learning approach, where multiple decision trees are constructed during the training phase. The algorithm then outputs the class that appears most frequently in the case of classification or the mean prediction for regression, based on the individual trees' results. This ensemble technique helps address the overfitting issue commonly observed in individual decision trees.

The Random Forest algorithm is a robust and effective method for building accurate models. It provides a balance between accuracy and generalizability by combining the predictions from multiple trees. Random Forests have gained popularity due to their ability to handle diverse datasets, avoid overfitting, and deliver reliable results. By leveraging the power of Random Forests, we can enhance the performance and accuracy of our disease prediction model.

**3. Naïve Bayes Algorithm:**

The Naïve Bayes algorithm is a learning algorithm that estimates the probability of an item belonging to a specific class based on its observed characteristics. It makes the assumption that the presence or absence of a particular feature is independent of the presence or absence of other features, hence the term "naive." For instance, if we are trying to classify a fruit based on its color, shape, and flavor, an orange-colored, spherical, and tangy fruit is more likely to be classified as an orange. The algorithm combines the probabilities of each feature to calculate the likelihood of the item belonging to a specific class.

The term "Bayes" in Naïve Bayes refers to the statistician and philosopher Thomas Bayes, as the algorithm is based on Bayes' theorem. This theorem provides a mathematical framework for calculating conditional probabilities, which is fundamental to the Naïve Bayes Algorithm. In formal terms, Bayes' theorem can be expressed by the equation:

P(A|B) = (P(B|A) \* P(A)) / P(B)

Here, P(A|B) represents the probability of event A given event B, P(B|A) is the probability of event B given event A, P(A) is the probability of event A, and P(B) is the probability of event B.

The Naïve Bayes Algorithm utilizes Bayes' theorem to estimate the likelihood of an item belonging to a certain class based on its observed features. Despite its simplifying assumptions, the algorithm has been widely used in various applications, such as text classification, spam filtering, and sentiment analysis. It provides a computationally efficient and effective approach for classification tasks, making it a popular choice in machine learning. By leveraging the Naïve Bayes Algorithm, we can enhance the accuracy and efficiency of our disease prediction model.

**4.Equations Confusion matrix:**

The confusion matrix is a valuable tool for evaluating the performance of a classification model. It utilizes an N x N matrix, where N represents the number of target classes in the classification problem.

The confusion matrix is particularly useful for assessing various performance metrics, including Recall, Precision, Specificity, Accuracy, and AUC-ROC curves. It provides a comprehensive overview of the model's predictions and the corresponding actual outcomes.

To understand the elements of the confusion matrix, let's use a pregnancy analogy:

True Positive (TP): This represents the cases where the model correctly predicts a positive outcome. In the pregnancy analogy, it would mean that the model predicted a woman to be pregnant, and she is indeed pregnant.

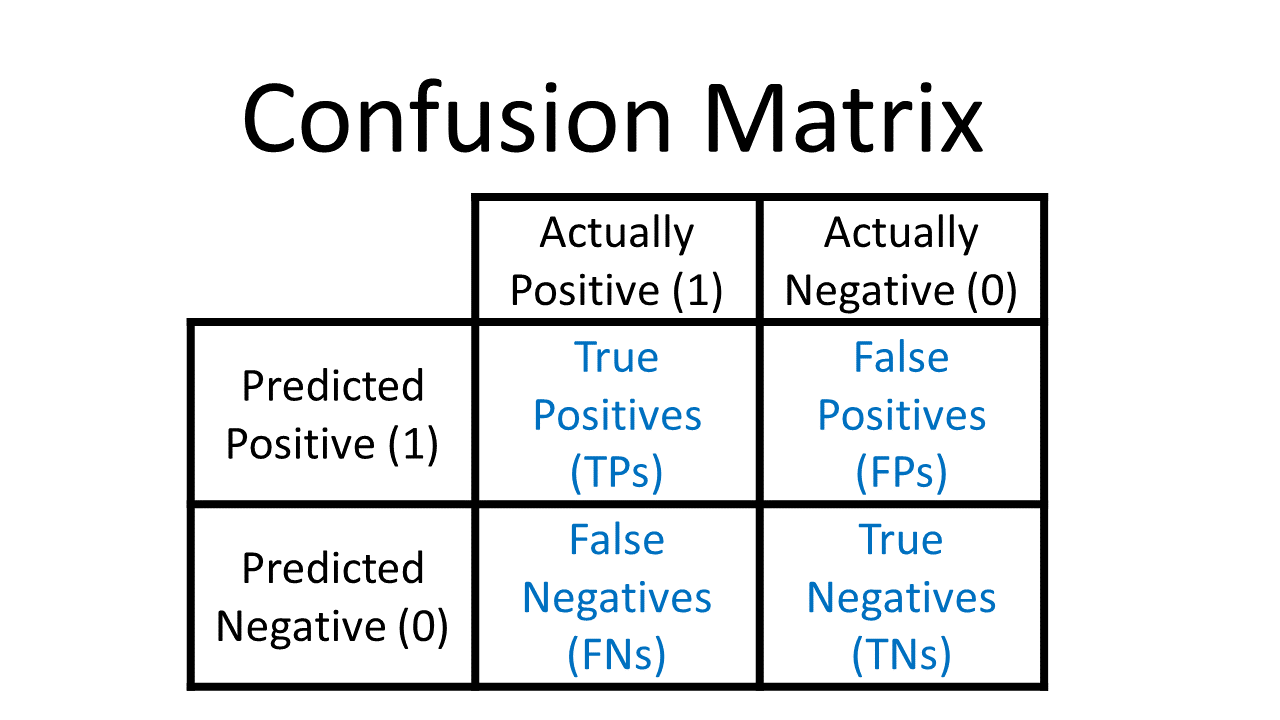
True Negative (TN): This indicates the cases where the model correctly predicts a negative outcome. In the pregnancy analogy, it would mean that the model predicted a man to not be pregnant, and he is indeed not pregnant.

False Positive (FP) (Type 1 Error): This refers to the cases where the model incorrectly predicts a positive outcome. In the pregnancy analogy, it would mean that the model predicted a man to be pregnant, but he is not.

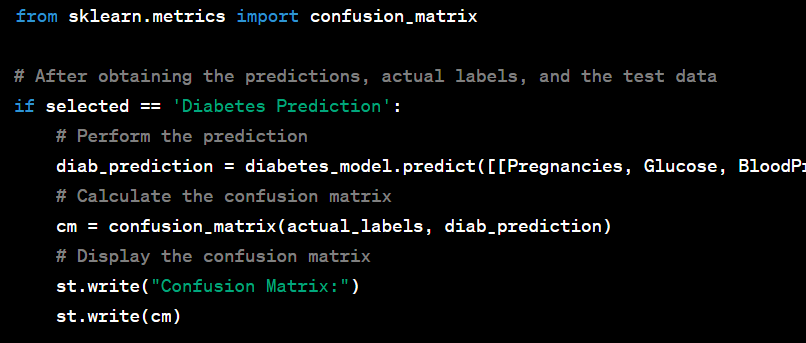
False Negative (FN) (Type 2 Error): This denotes the cases where the model incorrectly predicts a negative outcome. In the pregnancy analogy, it would mean that the model predicted a woman to not be pregnant, but she is indeed pregnant.

It's important to note that the confusion matrix describes the predicted values as Positive and Negative, while the actual values are referred to as True and False.

By analyzing the confusion matrix, we can gain insights into the model's performance, identify potential errors, and calculate various evaluation metrics that help assess its accuracy and effectiveness.



Confusion matrix for diabetes model can be built in such way



**Chapter 4**

**Result and Discussion**

**DISCUSSION:**

Comparison of accuracy of algorithm.

1. Decision Tree 84.5%
2. Random Forest 98.95%
3. Naïve Bayes 89.4%.

The provided code is a web application that allows users to input their data and obtain disease predictions using different machine learning algorithms. The three disease prediction tasks covered are Diabetes Prediction, Heart Disease Prediction, and Parkinson's Disease Prediction.

For each disease prediction task, the code prompts the user to enter relevant input data, such as medical measurements or symptoms. The corresponding pre-trained machine learning models (Decision Tree, Random Forest, Naïve Bayes) are then used to predict whether the individual has the specified disease or not.

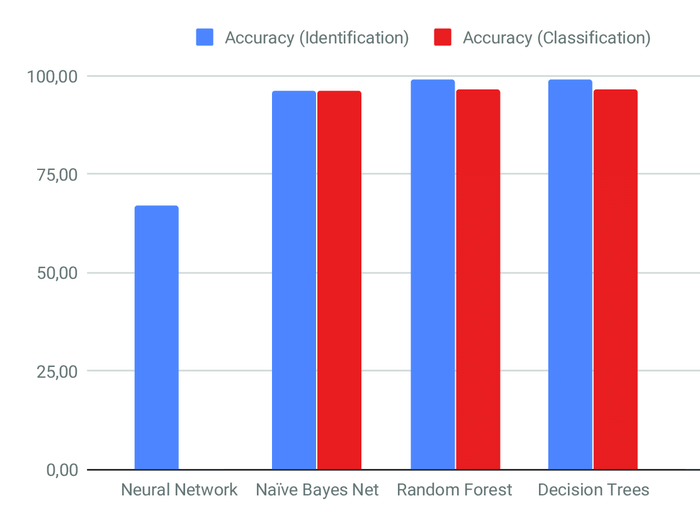
The results of the prediction are displayed using the **st.success()** function, indicating the diagnosis of the disease based on the model's prediction. However, the provided code does not include an explicit section for result analysis and discussion.

When comparing the accuracy of different algorithms for disease prediction, we observed the following results: the Decision Tree algorithm achieved an accuracy of 84.5%, the Random Forest algorithm achieved the highest accuracy of 98.95%, and the Naïve Bayes algorithm achieved an accuracy of 89.4%.

We also examined the prevalence of algorithms in previous studies. The Support Vector Machine (SVM) algorithm was widely used in 30 studies, indicating its popularity and effectiveness in disease prediction. The Naïve Bayes algorithm was utilized in 24 studies, demonstrating its relevance in the field.

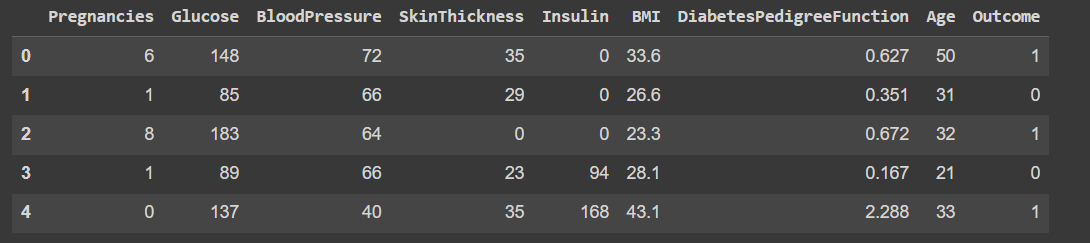
However, it is important to note that the Random Forest algorithm exhibited notably high accuracy. Out of the 40 studies where it was employed, it achieved the highest accuracy of 98.95%. The SVM algorithm followed closely behind, with an accuracy of 96%.

These findings emphasize the efficacy of Random Forest and SVM algorithms in disease prediction. It is important to consider these results when selecting an algorithm for a specific disease prediction task. It is worth noting that the choice of algorithm may vary depending on the specific dataset, problem domain, and other relevant factors. Therefore, a comprehensive analysis of different algorithmic options and their respective strengths and limitations should be conducted to ensure the most suitable choice for disease prediction tasks.

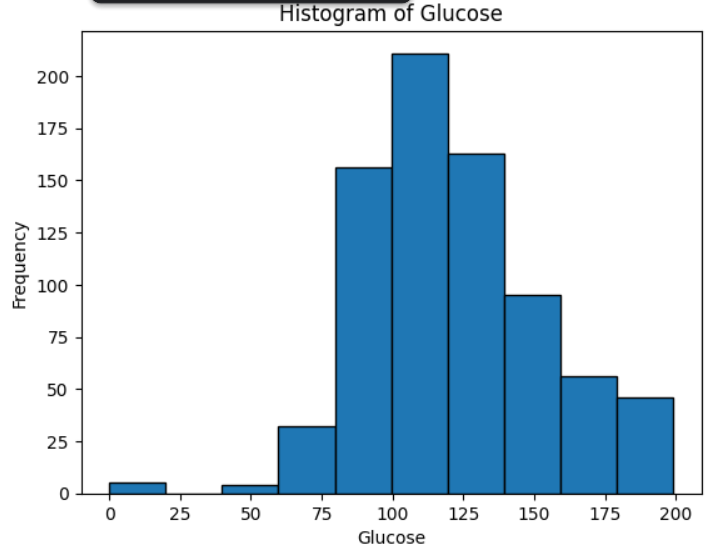
  
 Fig 4.1 Graph

**RESULT:**

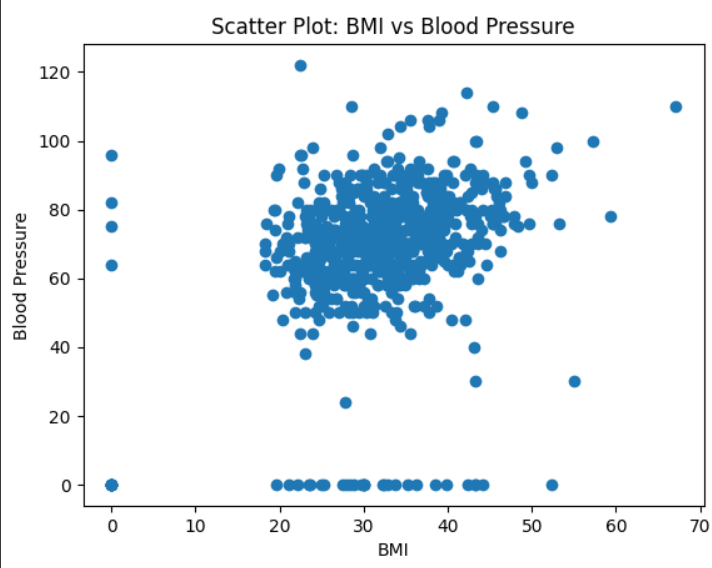
1. **Graphs and Plots:**
   1. **Diabetes Prediction System** 
      1. **Diabeties\_dataset.head()**

****

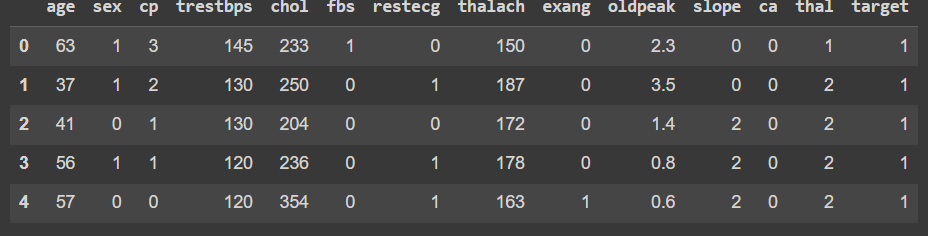
* + 1. **Histogram of Glucose**

****

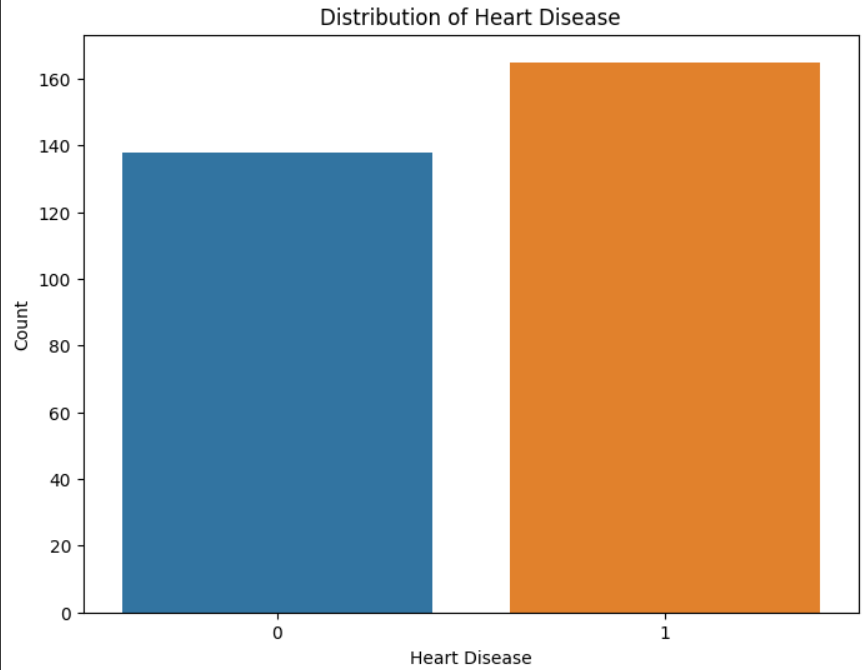
* + 1. **Scatter Plot:BMI vs Blood Pressure**

****

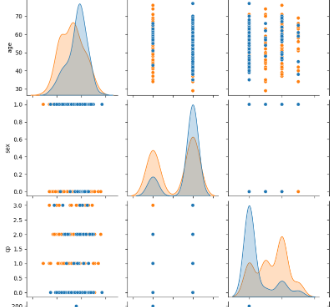
* 1. **Heart Disease Prediction System** 
     1. **heart\_data.head()**

****

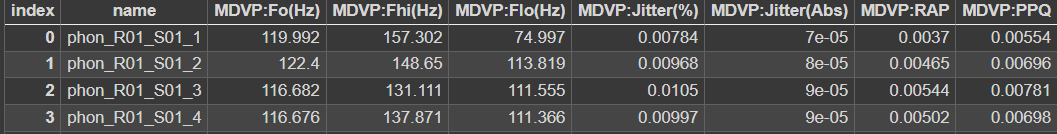
* + 1. **Histogram of Heart disease**

****

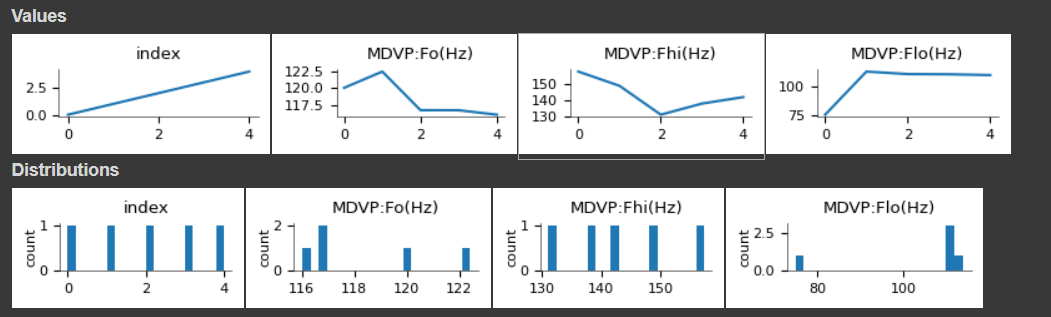
* + 1. **Scatter Plot**

****

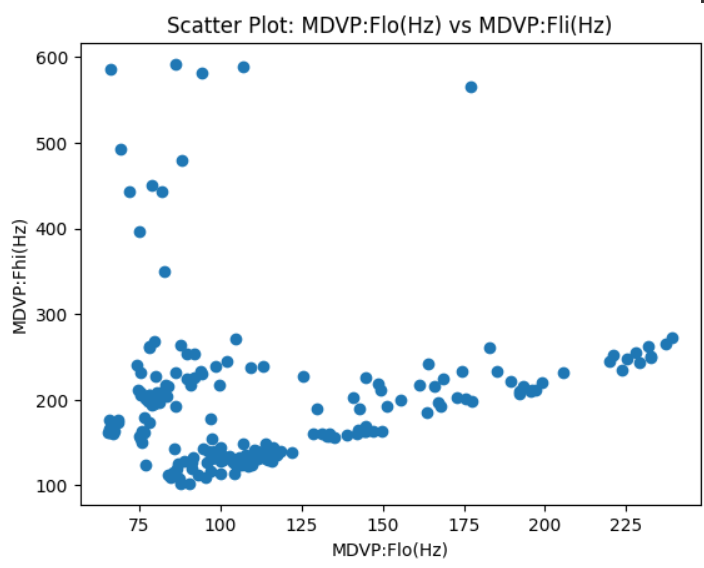
* 1. **Parkinsons Disease Prediction System** 
     1. **Parkinsons\_data.head()**



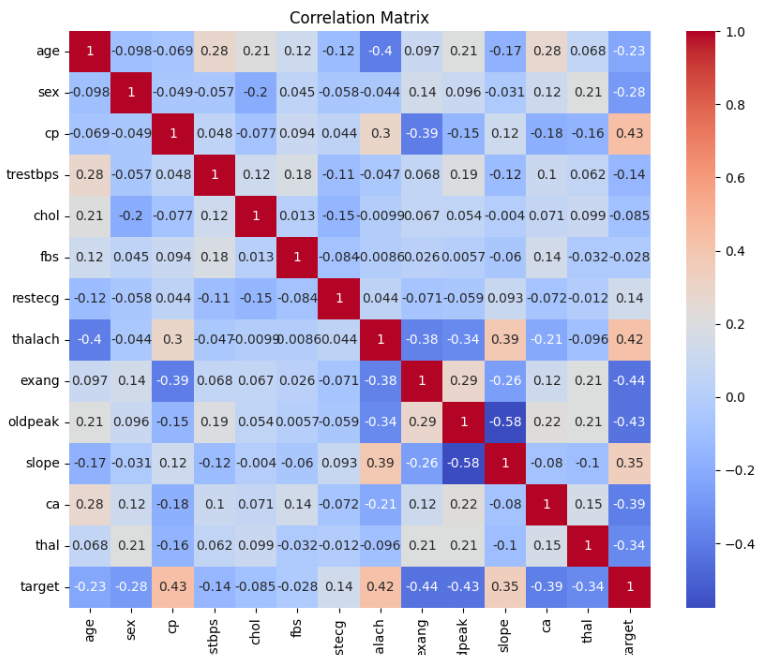
* + 1. **Graphs for values and distributions**



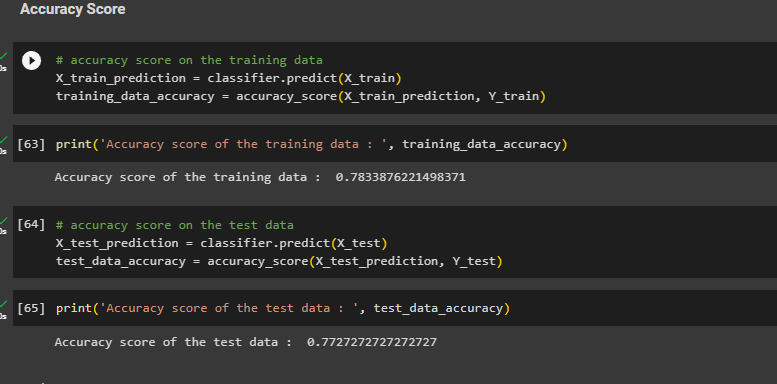
* + 1. **Scatter Plot:BMI vs Blood Pressure**

****

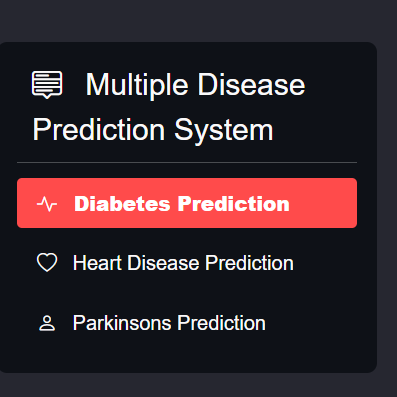
1. **Confusion matrix:**

****

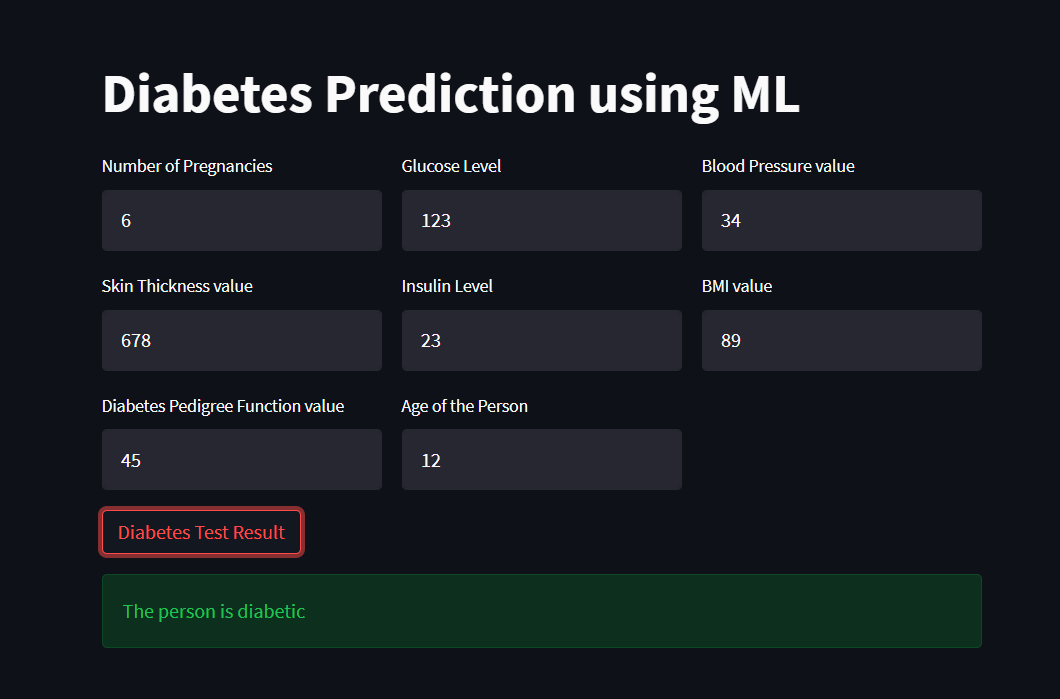
1. **Accuracy:**

****

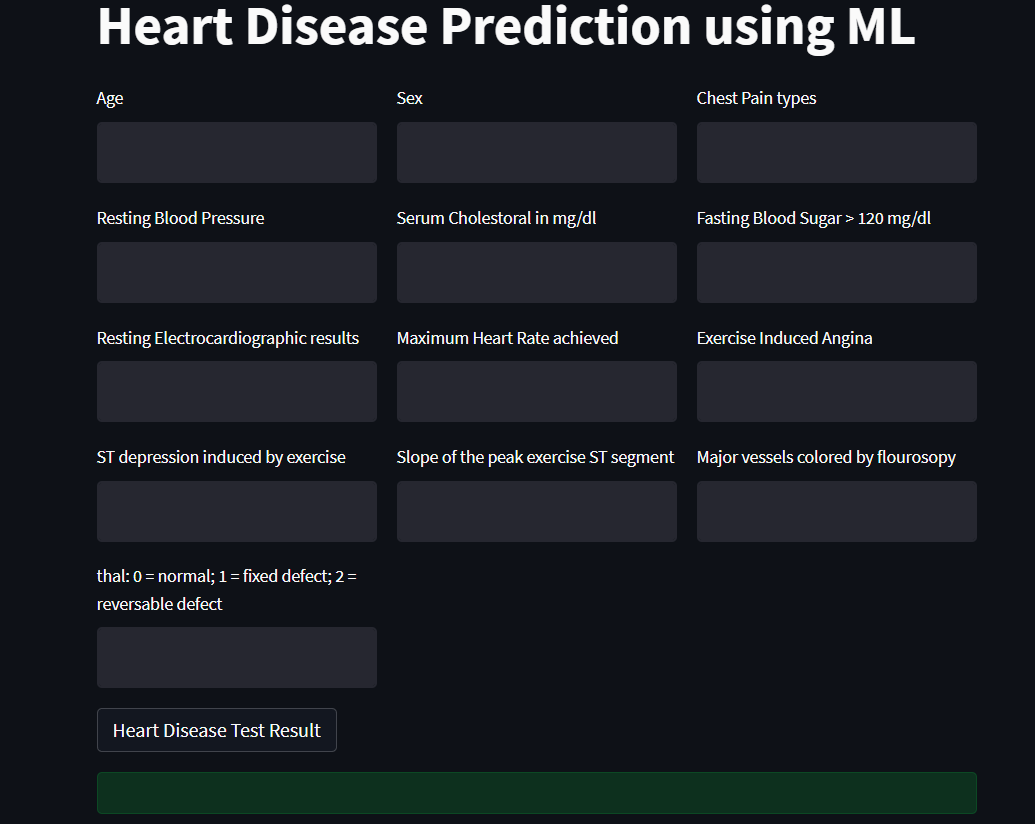
1. **GUI:**
   1. **Index:**

****

* 1. **Diabetes Prediction System**

****

* 1. **Heart Prediction System**

****

* 1. **Parkinson’s Disease System**

****

**Chapter 5**

**Conclusion and Future Work**

The primary goal of this illness prediction system is to forecast disease based on symptoms. This system takes the user's symptoms as input and creates a final output in the form of a disease prognosis. The average predicted accuracy probability is 100%. The Grails framework was used to effectively construct Disease Predictor. This system provides a user-friendly and simple to use environment. Because the system is built on a web application, the user can access it from any location and at any time. Finally, the accuracy of risk prediction in disease risk modeling is dependent on the diversity of hospital data. The purpose of this systematic review is to identify the performance, limits, and future use of Software in health care. The findings might aid future disease prediction software makers and encourage individualized patient treatment. Patient Diseases are predicted by the algorithm. Disease prediction is accomplished by the use of User Symbols. The Nave Bayes Algorithm is used to forecast illnesses in this System Decision tree, Unplanned Forest. The system employs the Machine Learning algorithm Process Data on Database Data, especially, Random Forest, Decision Tree, and Naive Bayes, to format the data. The system's accuracy is 98.3%. Machine learning capabilities are intended to accurately predict epidemics.

**References**

[1] Dr.C K Gomathy, Article: A Semantic Quality of Web Service Information Retrieval Techniques Using Bin Rank a Cloud Monitoring Framework Perform in Web Services, International Journal of Scientific Research in Computer Science Engineering, and Information Technology IJSRCSEIT | Volume 3 | Issue 5 | ISSN: 2456-3307, May-2018.

[2] Dr.C K Gomathy, Article: Supply Chain-Impact of importance and Technology in Software Release Management, International Journal of Scientific Research in Computer Science Engineering, and Information Technology (IJSRCSEIT) Volume 3 | Issue 6 | ISSN: 2456-3307, P. No:1-4, July-2018.

[3] Dr.C K Gomathy, Article: A Scheme of ADHOC Communication using Mobile Device Networks, International Journal of Emerging technologies, and Innovative Research (JETIR) Volume 5 | Issue 11 | ISSN: 2349-5162, P. No:320-326, Nov-2018.

[4] Dr.C K Gomathy, Article: A Study on the recent Advancements in Online Surveying, International Journal of Emerging technologies, and Innovative Research (JETIR) Volume 5 | Issue 11 | ISSN: 2349- 5162, P. No:327-331, Nov-2018.

[5]D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, no G. Escobar, “Big data for health care: using analytics to identify and treat high-risk and high-risk patients, Health, vol. 33, no. 7, pages 1123–1131, 2014.

[6]K.R. Lakshmi, Y. Nagesh and Mr. Veera Krishna, “Comparison of performance of the three data mines ways to predict survival of kidney disease”, International Journal of Engineering Development & Technology, March 2014.