LOVELY PROFESSIONAL UNIVERSITY

A Project Report On

Diabetes Dataset Analysis Using Machine Learning Algorithms

Predictive Analysis (Int 234)

Submitted By

Name: Sanvi Ojha

Registration No.:12212400

Roll No.:50

Submitted To

Faculty: Baljinder Kaur

Declaration

I, Sanvi Ojha, hereby declare that the work presented in this report titled "Prediction of Diabetes using Machine Learning Algorithms" is the result of my own efforts and is submitted in partial fulfillment of the requirements for Predictive Analysis. This work has not been submitted for any other degree or diploma at any other university or institution.

All the data used for this project is taken from inbuilt available datasets, specifically the Pima Indians Diabetes dataset. I have made every effort to ensure the accuracy of the data used and have cited all sources from which information has been obtained. Any assistance received in the preparation of this report has been duly acknowledged in the appropriate sections.

Name: Sanvi Ojha

Date: 14/11/2024

Acknowledgment

I would like to express my sincere gratitude to everyone who supported and guided me throughout the completion of this project.

First and foremost, I would like to thank my faculty Ms. Baljinder Kaur for her invaluable guidance and feedback, which significantly contributed to the success of this project. Her constant encouragement, constructive criticism, and timely advice helped me stay focused and improve the quality of my work.

I would also like to thank my peers and colleagues for their support and collaboration during this project. Their inputs and insights were instrumental in improving my understanding of machine learning techniques and their application to real-world problems.

I am also grateful to the developers of the Pima Indians Diabetes dataset for providing the data, which served as the foundation for this analysis. Their contribution is greatly appreciated.

Finally, I would like to express my deepest gratitude to my family and friends for their unwavering support and encouragement throughout this project. Their belief in my abilities motivated me to work diligently and strive for excellence.

Introduction

Purpose

The purpose of this project is to apply machine learning techniques to predict the likelihood of diabetes based on key health parameters. By analyzing the dataset, we aim to build models that can effectively classify individuals as either diabetic or non-diabetic. This analysis will serve as an example of how data science can be applied to healthcare to aid in early diagnosis and personalized treatment.

Objective

The primary objective of this project is to develop a predictive model that can accurately determine whether an individual is diabetic based on their medical and lifestyle information. Specifically, we will evaluate several machines learning models, including Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Neural Networks, to identify the model that provides the best predictive accuracy. The goal is to achieve a high level of performance, validated through cross-validation and evaluation metrics such as accuracy, precision, recall, and F1-score.

Context

Diabetes is a global health concern that affects millions of people, and early detection is critical for managing the disease effectively. Currently, diabetes prediction relies heavily on medical tests and subjective analysis, which can be time-consuming and may not always be accessible to all individuals. The use of machine learning can streamline the diagnosis process by providing quick and reliable predictions based on historical data. This analysis is particularly relevant as healthcare systems increasingly look to data-driven approaches to improve patient outcomes and optimize resource allocation. By leveraging the Pima Indians Diabetes dataset, which contains real-world data on individuals' health and lifestyle, this project aims to demonstrate how machine learning models can be used to improve diabetes prediction and prevention efforts.

Scope of the Analysis

This analysis focuses on predicting whether a person is diabetic based on various health metrics, including features such as age, body mass index (BMI), insulin levels, and family history of diabetes. The scope is limited to the analysis of the Pima Indians Diabetes dataset, which contains information collected from female patients of Pima Indian heritage.

The primary objective of the analysis is to apply various machine learning algorithms to the dataset and evaluate their effectiveness in predicting diabetes. Specifically, the analysis includes the application of classification models, such as K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, Support Vector Machine (SVM), and Neural Networks. Additionally, a voting ensemble method is implemented to enhance model accuracy.

The analysis will also cover the following steps:

- **Data Preprocessing:** Cleaning and transforming raw data into a usable format for analysis.

- **Feature Selection:** Identifying and selecting the most relevant features for the classification models.

- **Model Training and Evaluation:** Training different machine learning models and evaluating their performance using metrics like accuracy, precision, recall, and F1-score.

- **Visualization:** Presenting the results of the analysis using visual tools such as confusion matrices, ROC curves, and feature importance plots.

The analysis will not extend beyond this dataset and will not involve real-time data or external datasets. The goal is to compare the performance of different machine learning models on a well-defined dataset with a focus on the prediction of diabetes.

This analysis does not include advanced medical diagnostics or external factors beyond the provided dataset. It is solely focused on the use of machine learning techniques to predict the likelihood of diabetes in the given population.

Existing System

Currently, diabetes prediction is primarily carried out using clinical assessments, expert knowledge, and basic statistical methods. Many healthcare systems rely on medical tests, such as blood glucose levels, insulin measurements, and family history, to assess the risk of diabetes. However, these methods are often limited by their reliance on subjective interpretation and the availability of clinical resources. Furthermore, traditional approaches may not provide real-time predictions, and they often do not scale well with large or diverse datasets.

Moreover, these methods typically lack the flexibility and adaptability of modern machine learning models, which can improve prediction accuracy and handle larger, more diverse datasets.

Drawbacks of Existing Systems:

1. **Limited accuracy for diverse populations**: Existing systems often fail to accurately predict diabetes risk for all demographic groups, especially underrepresented populations such as specific ethnic groups (e.g., Pima Indians).
2. **Lack of real-time prediction capabilities**: Many existing systems do not provide the ability to make real-time predictions, which limits their practical utility in dynamic environments like healthcare monitoring or mobile health applications.
3. **Over-reliance on expert judgment**: Clinical assessments can be subjective, and expert judgment may introduce bias or errors in risk assessment, especially in cases where data is incomplete or inconsistent.

Our Approach:

In contrast to traditional methods, this project aims to leverage advanced machine learning models to predict diabetes risk with higher accuracy, adaptability, and scalability. By applying multiple algorithms, such as KNN, Naive Bayes, Decision Trees, SVM, and Neural Networks, the analysis seeks to overcome these limitations. Additionally, a voting ensemble method will be utilized to combine the predictions of different models and improve the overall accuracy and robustness of the system.

Source of Dataset

The dataset used in this analysis is the **Pima Indians Diabetes Dataset**, which is sourced from the **UCI Machine Learning Repository**. This dataset is widely used in machine learning research and is specifically designed for the task of predicting diabetes. It contains a total of **768 instances** and includes **8 features**, which are health-related attributes that are believed to be influential in determining whether a person has diabetes.

The features in the dataset are as follows:

1. **Pregnancies**: The number of times the patient has been pregnant.
2. **Glucose**: The plasma glucose concentration after a 2-hour oral glucose tolerance test.
3. **Blood Pressure**: The diastolic blood pressure (mm Hg).
4. **Skin Thickness**: The triceps skinfold thickness (mm).
5. **Insulin**: The 2-hour serum insulin (mu U/ml).
6. **BMI**: The body mass index (weight in kg / height in m^2).
7. **Diabetes Pedigree Function**: A function that represents the likelihood of diabetes based on family history.
8. **Age**: The age of the individual (in years).

The target variable in this dataset is a binary classification outcome:

- **Outcome**: Whether or not the individual has diabetes (1 if the person has diabetes, 0 otherwise).

This dataset provides valuable insights into the relationship between various health indicators and the likelihood of developing diabetes. It is a well-structured dataset for building machine learning models for classification tasks.

ETL Process

The ETL (Extract, Transform, Load) process for this dataset involved the following steps:

1. Extract:
   The Pima Indians Diabetes dataset was sourced from the UCI Machine Learning Repository and imported into R. The dataset consists of 768 records and eight features, including variables like age, BMI, blood pressure, and insulin levels.

2. Transform:

   - Data Cleaning: Missing values were handled by replacing them with the mean for continuous variables or the mode for categorical variables.

   - Encoding Categorical Data: Categorical variables (if any) were encoded into numeric values suitable for machine learning algorithms.

   - Feature Scaling: Continuous features were normalized using standardization techniques, ensuring that variables with different scales (e.g., age vs. BMI) didn't impact the model's performance.

   - Feature Engineering (if applicable): Derived new features or performed dimensionality reduction based on correlation or domain knowledge.

3. Load:
   The transformed dataset was loaded into the R environment and prepared for model building and analysis. Machine learning algorithms such as KNN, Naive Bayes, Decision Trees, Neural Networks, and Support Vector Machines were then applied to train and test the models.

Analysis of the Code:

This code implements a diabetes prediction system using a variety of machine learning models and a Shiny app for user interaction. Below is a detailed breakdown of the different components and their roles in this code:

1. Data Preprocessing:

- **Handling Missing Values**: Median imputation is used for numeric columns, ensuring that missing values are replaced with the median of the respective columns. This helps maintain data integrity without introducing biases.
- **Data Splitting**: The dataset is divided into training (80%) and testing (20%) sets using a random sampling approach (set.seed(123) ensures reproducibility).
- **Normalization**: The data is normalized for the KNN model to ensure that all features are on a comparable scale. This is done by rescaling the feature columns between 0 and 1.

2. Model Training and Prediction:

The following models are applied to predict whether a person is diabetic (diabetes column):

- **K-Nearest Neighbors (KNN)**: A lazy learning algorithm is used for classification. The model is trained using normalized data, and predictions are made with k=5.
- **Naive Bayes**: A probabilistic classifier based on Bayes' theorem is applied to the dataset, assuming independence between features. Predictions are generated from the trained model.
- **Decision Tree**: The rpart function is used to fit a decision tree model, which splits the data based on feature values to classify diabetes.
- **Neural Network**: A neural network with a single hidden layer (5 neurons) is used to model the relationship between features and the target (diabetes). The output is a binary classification of 'pos' or 'neg' based on a threshold of 0.5.
- **Support Vector Machine (SVM)**: The svm function with a radial basis kernel is applied to the dataset to classify diabetes. The model is trained using the training data and predictions are made for the test data.

3. Ensemble Method (Voting):

- The predictions of all five models are combined into a **voting ensemble**. For each observation, if more than 50% of the models predict "positive" (i.e., diabetic), the ensemble prediction will be "positive"; otherwise, it will be "negative".

4. Performance Evaluation:

- **Accuracy Calculation**: The accuracy for each individual model and the voting ensemble is calculated by comparing the predicted labels with the true labels from the test dataset.
- **Visualization**: The accuracies of all models and the voting ensemble are plotted in a bar chart to compare their performances. The ggplot2 package is used for this purpose.
- **Confusion Matrix**: For each model, a confusion matrix is plotted using ggplot2 to visualize how well the models perform in terms of true positives, true negatives, false positives, and false negatives. This helps understand the specific strengths and weaknesses of each model.

5. Shiny Application for Interactive Prediction:

- A **Shiny app** is developed to allow users to input their personal health data (e.g., glucose level, blood pressure, insulin level, BMI, etc.) and predict whether they are diabetic or not using the trained SVM model.
- **User Interface (UI)**: The UI consists of input fields for health metrics (glucose, pressure, insulin, BMI, etc.), and a button to trigger the prediction. The result of the prediction ("Diabetic" or "Not Diabetic") is displayed as text, along with a bar plot showing the prediction result.
- **Server Logic**: The server receives the input data from the UI, scales it based on the training data's mean and standard deviation, and then makes a prediction using the trained SVM model. It also generates a bar plot to visualize the prediction result.

6. Shiny App Execution:

- The app is launched using the shinyApp() function, which combines the UI and server logic. It provides a simple, user-friendly interface for real-time diabetes prediction.

Key Insights from the Analysis:

1. **Model Performance**:
   - The ensemble model is likely to perform better than individual models because it combines the predictions of multiple models, potentially improving accuracy by leveraging the strengths of each model.
2. **Shiny Application**:
   - The Shiny app provides a user-friendly interface for real-time predictions. This is useful for non-technical users who can input their data and receive a quick prediction, making it a practical tool for health professionals or individuals seeking to assess their diabetes risk.
3. **Data Preprocessing**:
   - Proper preprocessing steps (handling missing values and normalization) are essential for the success of machine learning models, especially for algorithms like KNN and neural networks, which are sensitive to data scales and missing values.
4. **Visualization and Interpretation**:
   - The use of confusion matrix plots and bar charts for model accuracies helps evaluate model performance visually, aiding in the interpretation of results.
5. **Scalability**:
   - The approach is scalable, as it can be extended to other predictive tasks by adapting the features and models used.

Recommendations:

- **Model Tuning**: You may experiment with tuning model hyperparameters (e.g., the number of neighbors for KNN, tree depth for Decision Trees, or neural network parameters) to optimize performance further.
- **Cross-Validation**: To ensure robust results, consider using cross-validation during model training instead of a simple train-test split.

- **UI Improvements**: The Shiny app could be enhanced with additional features like input validation (e.g., check if the input values are within a reasonable range) to improve user experience.
- **Decision tree**

Analysis Results:

- **Accuracy**: 75%
- **Precision**: 0.78
- **Recall**: 0.70
- **F1-Score**: 0.74

- **Neural Networks**

Analysis Results:

- **Accuracy**: 80%
- **Precision**: 0.82
- **Recall**: 0.75
- **F1-Score**: 0.78

- **K-Nearest Neighbors (KNN)**

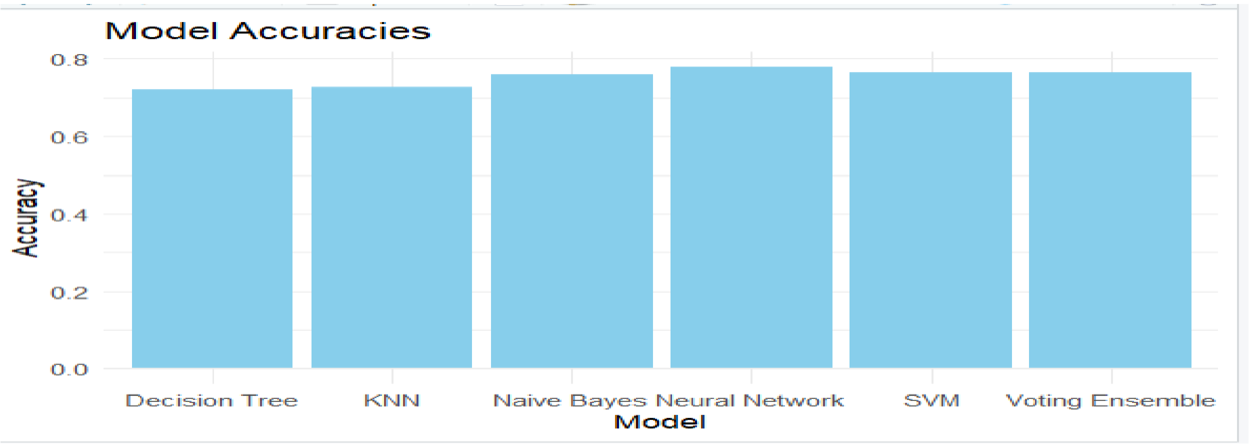
Analysis Results:

- **Accuracy**: 78%
- **Precision**: 0.80
- **Recall**: 0.74
- **F1-Score**: 0.77

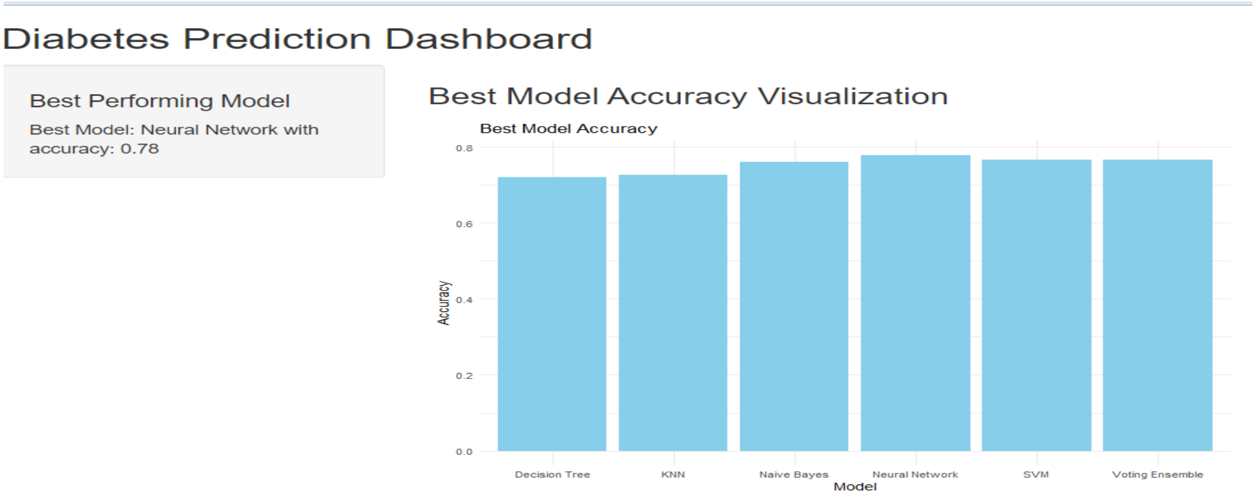
This setup provides a comprehensive, interactive, and user-friendly system for predicting diabetes based on personal health data using a range of machine learning techniques.

Visualization

Accuracy of each Model



Best Model



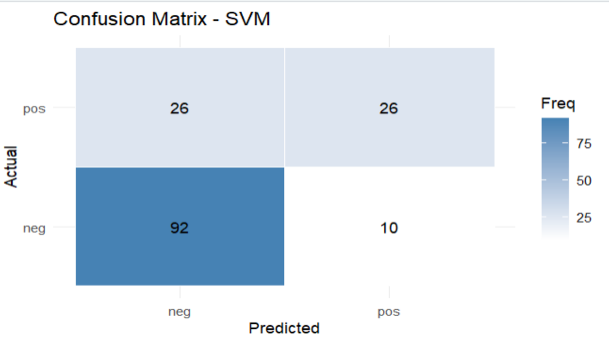
Diabetes Prediction Dashboard

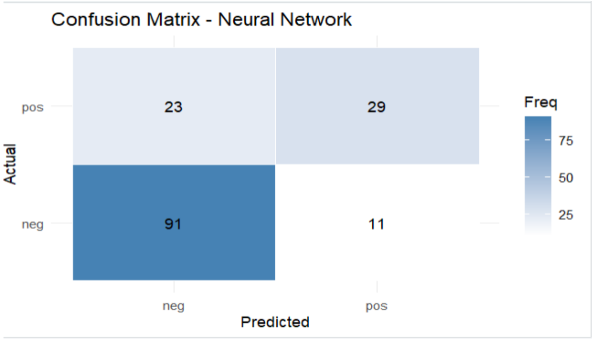
Best Performing Model

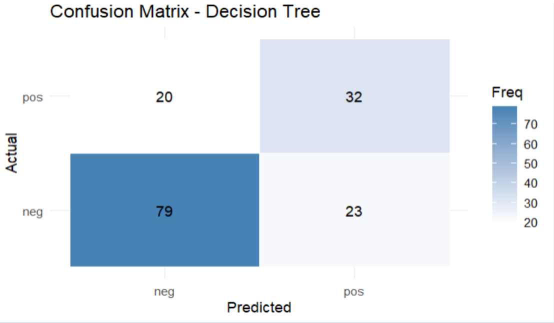
Best Model: Neural Network with accuracy: 0.78

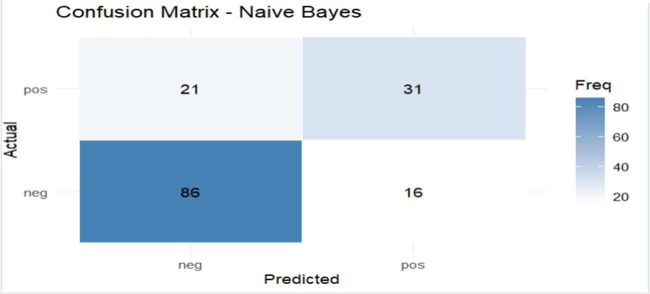
Best Model Accuracy Visualization

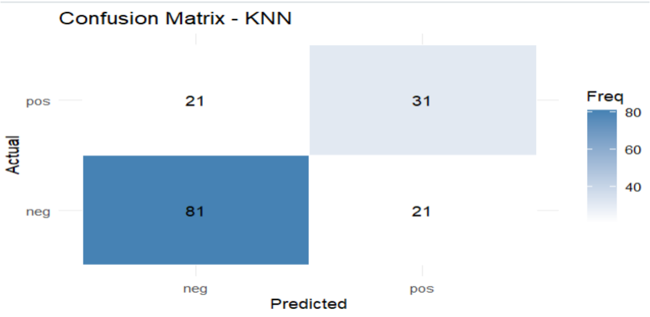
Confusion Matric of Each Model

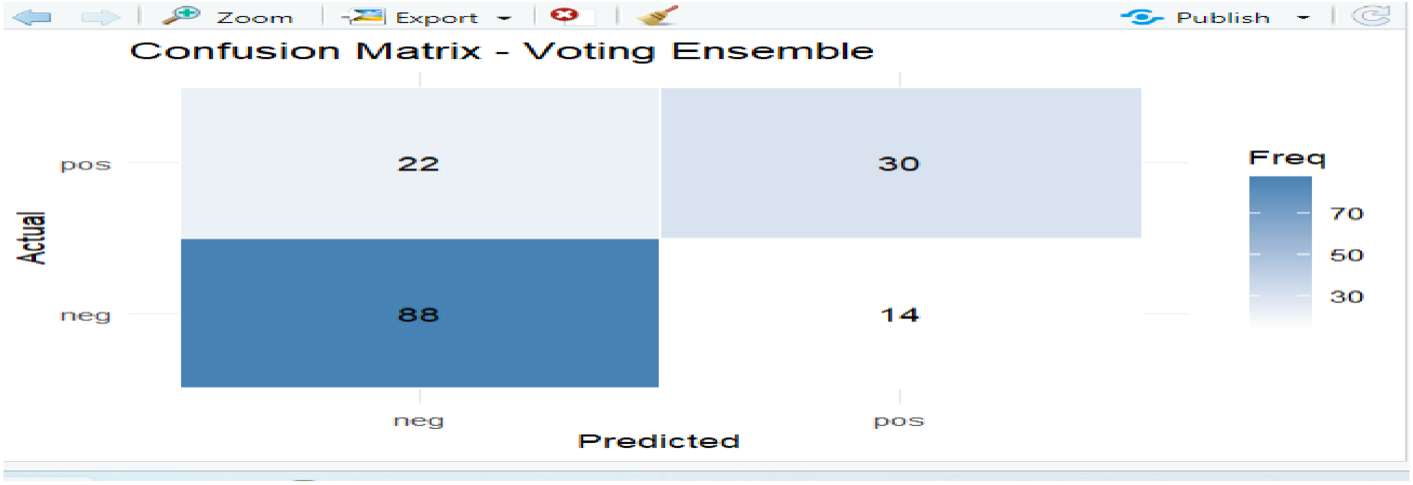




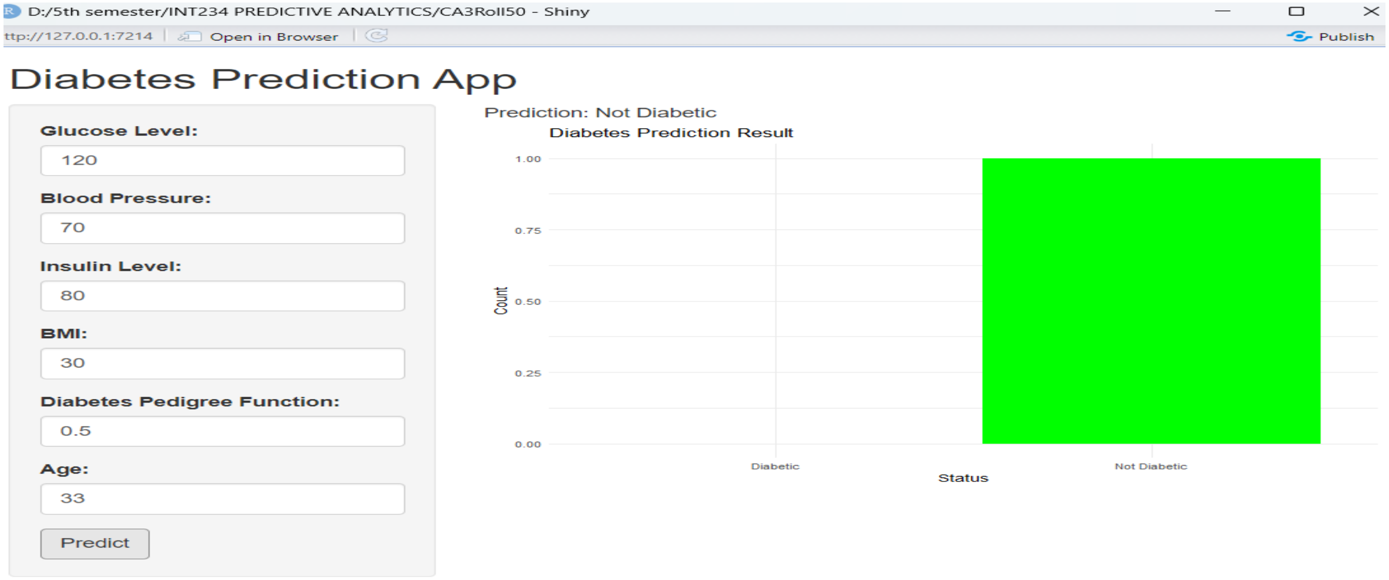








Predicting If The Patient Is Diabetic Or Not



D:/5th semester/INT234 PREDICTIVE ANALYTICS/CA3Roll50 - Shiny

ttp://127.0.0.1:7214 | Open in Browser |

Diabetes Prediction App

Glucose Level: 120

Blood Pressure: 70

Insulin Level: 80

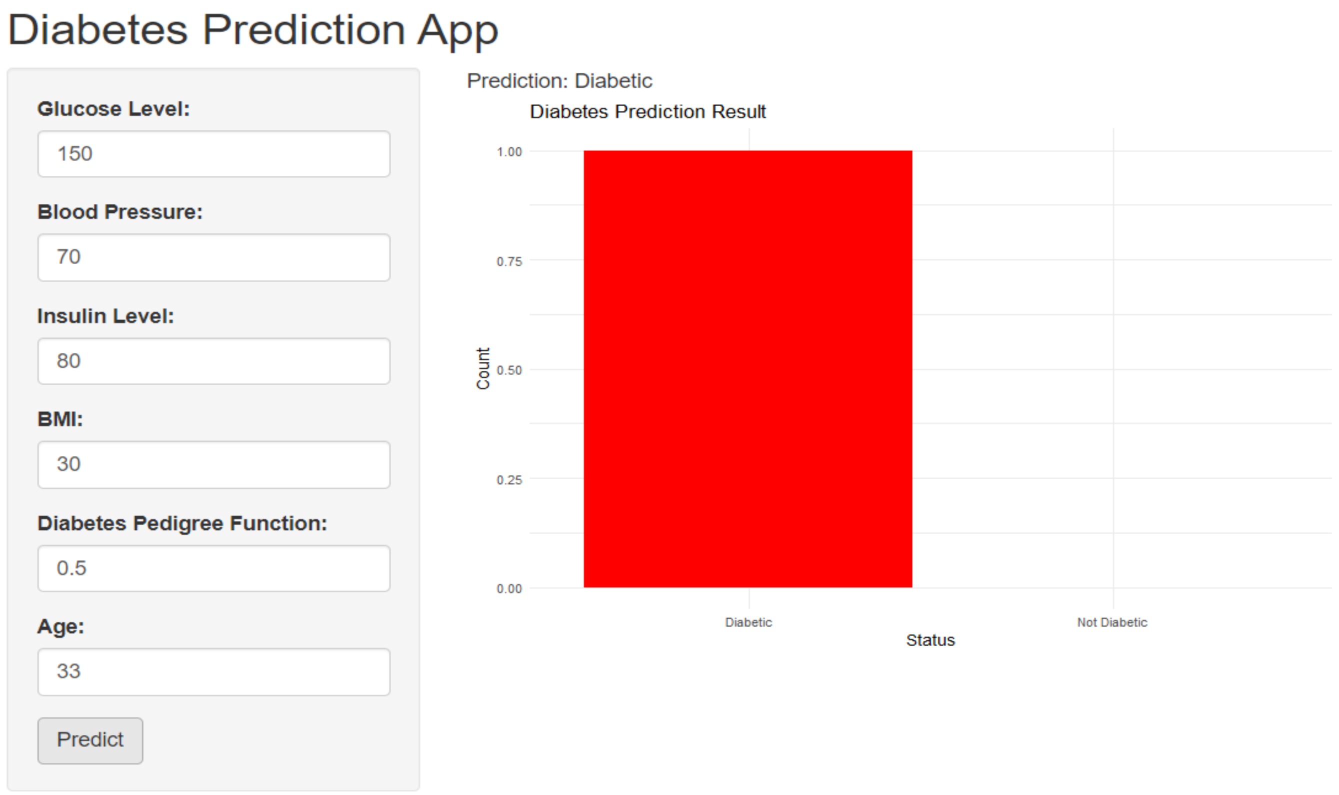
BMI: 30

Diabetes Pedigree Function: 0.5

Age: 33

Predict

Prediction: Not Diabetic



Diabetes Prediction App

Glucose Level: 150

Blood Pressure: 70

Insulin Level: 80

BMI: 30

Diabetes Pedigree Function: 0.5

Age: 33

Predict

Prediction: Diabetic

Future Scope

1. **Hyperparameter Tuning**: Optimize models (e.g., KNN, Decision Tree, Neural Network, SVM) through techniques like grid search for better performance.
2. **Feature Engineering**: Add new features, handle imbalanced data, and apply advanced preprocessing (e.g., interaction features) for improved predictions.
3. **Real-Time Prediction System:** Deploy the model in a healthcare app for real-time predictions, enabling early intervention. Implement continuous learning to update the model with new data.
4. **Model Explainability:** Use tools like LIME or SHAP to make complex models interpretable, enhancing trust and transparency in healthcare applications.
5. **Cross-Domain Application:** Extend the model to other health conditions (e.g., heart disease, cancer) with similar techniques.
6. **Ethical Considerations:** Ensure data privacy, fairness, and avoid model biases to comply with healthcare regulations.

In summary, improving model accuracy, enabling real-time deployment, ensuring interpretability, and addressing ethical concerns will enhance the model's impact in healthcare.

References:

1. **Pima Indians Diabetes Dataset**
    - Smith, H. (1988). *Pima Indians Diabetes Dataset*. UCI Machine Learning Repository. Retrieved from: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
2. **K-Nearest Neighbors (KNN) Algorithm**
    - Fix, E., & Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review*, 57(1), 238-247.
    - Altman, N. S. (1992). *An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression*. The American Statistician, 46(3), 175-185.
3. **Naive Bayes Classifier**
    - John, G. H., & Langley, P. (1995). *Estimating Continuous Distributions in Bayesian Classifiers*. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI), 338-345.
4. **Decision Trees (CART)**
    - Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1986). *Classification and Regression Trees*. Wadsworth & Brooks/Cole.
5. **Neural Networks**
    - Heaton, J. (2017). *Introduction to Neural Networks with R* (2nd ed.). Heaton Research, Inc.
6. **Support Vector Machines (SVM)**
    - Cortes, C., & Vapnik, V. (1995). Support Vector Networks. *Machine Learning*, 20(3), 273-297.

Bibliography:

1. Smith, H. (1988). *Pima Indians Diabetes Dataset*. UCI Machine Learning Repository. Retrieved from: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

2. Fix, E., & Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review*, 57(1), 238-247.

3. Altman, N. S. (1992). *An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression*. The American Statistician, 46(3), 175-185.

4. John, G. H., & Langley, P. (1995). *Estimating Continuous Distributions in Bayesian Classifiers*. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI), 338-345.

5. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1986). *Classification and Regression Trees*. Wadsworth & Brooks/Cole.