ICT 3233 Machine Learning

Continuous Assessment (ML Project)

AF/21/17702

AR/107019

K.K.S. Priyanath

Diabetes Prediction Using Random Forest Classifier

1. Title of Topic

Diabetes Prediction Using Random Forest Classifier

2. Objective / Problem Statement

The objective of this project is to develop a machine learning model that can predict whether a person is likely to have diabetes based on medical and lifestyle related attributes. Early prediction of diabetes is important for timely medical advice, lifestyle changes and prevention of health complications.

3. Chosen ML Technique

The chosen algorithm for this project is the Random Forest Classifier. Random Forest is a supervised machine learning algorithm that combines multiple decision trees to form an ensemble. Each tree gives a classification, and the forest selects the majority vote for the final prediction.

- This method was selected because:
- It provides higher accuracy compared to a single decision tree.
- It reduces the risk of overfitting by averaging results across multiple trees.
- It works well with imbalanced medical datasets like the diabetes dataset.
- It can provide probability estimates, which are useful for risk assessment.

4. Dataset Details

The dataset used is the Pima Indians Diabetes Dataset, publicly available from github.com.

Source - **https://github.com/npradaschnor/Pima-Indians-Diabetes-Dataset/blob/master/diabetes.csv**

Number of Records - 768 records and 8 input features

Features/Attributes – Pregnancies, Glucose level, Blood Pressure, Skin Thickness, Insulin level, BMI (Body Mass Index), Diabetes Pedigree Function (genetic influence), Age

Output Variable -

- 0 = No Diabetes
- 1 = Diabetes

5. Preprocessing Steps

Before training the model, the dataset underwent several preprocessing steps:

1. **Handling Invalid Values**
    - Certain features such as Glucose, Blood Pressure, Skin Thickness, Insulin and BMI contained zero values, which are not valid medical measurements. These invalid zeros were treated as missing values.
2. **Missing Value Imputation**
    - All missing values were filled using the median of each column. This approach reduces the impact of outliers and ensures that the dataset is complete for training.
3. **Separating Features and Target**
    - The dataset was divided into input features and the target variable.
4. **Train-Test Split**
    - The data was split into 80% training and 20% testing sets to evaluate model performance on unseen data.

6. Model Training & Prediction

- The Random Forest Classifier was used to train the model using the training dataset.
- The model learned patterns and relationships between the input features (medical and demographic attributes) and the target variable (Diabetes outcome).
- After training, the model was evaluated on the test dataset to measure its performance on unseen data.
- Predictions included both the predicted class (Diabetes or No Diabetes) and the probability of each class, allowing a confidence-based interpretation of results.

- Performance metrics such as accuracy, confusion matrix, and prediction probabilities were used to assess how well the model distinguishes between diabetic and non-diabetic cases.

7. Results & Plots

- ❖ The Random Forest model achieved an accuracy of approximately 78% on the test dataset, demonstrating good predictive capability for diabetes detection.
- ❖ A confusion matrix was generated to show the number of correct and incorrect predictions for both classes (Diabetes and No Diabetes). This visualization helps understand which type of cases the model predicts more accurately.
- ❖ For each user input, a probability chart was displayed, showing the likelihood of having Diabetes versus No Diabetes.
    - Probabilities were scaled to percentages (0–100%) for easy interpretation.
    - The predicted class was clearly highlighted, helping users understand the confidence level of the prediction.

❖ These visualizations make the model interactive and provide clear insight into its predictions.

Plots -

- **Confusion Matrix**

Shows how many predictions were correct or wrong for Diabetes and No Diabetes.

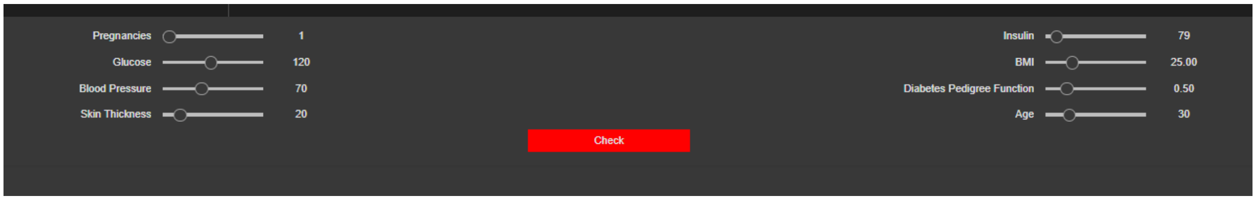
- **Prediction Probability Chart**

A bar chart shows the chance (%) of having Diabetes or No Diabetes for a given input. The higher bar shows the predicted class and how confident the model is.

- **Feature Contribution Line Chart**

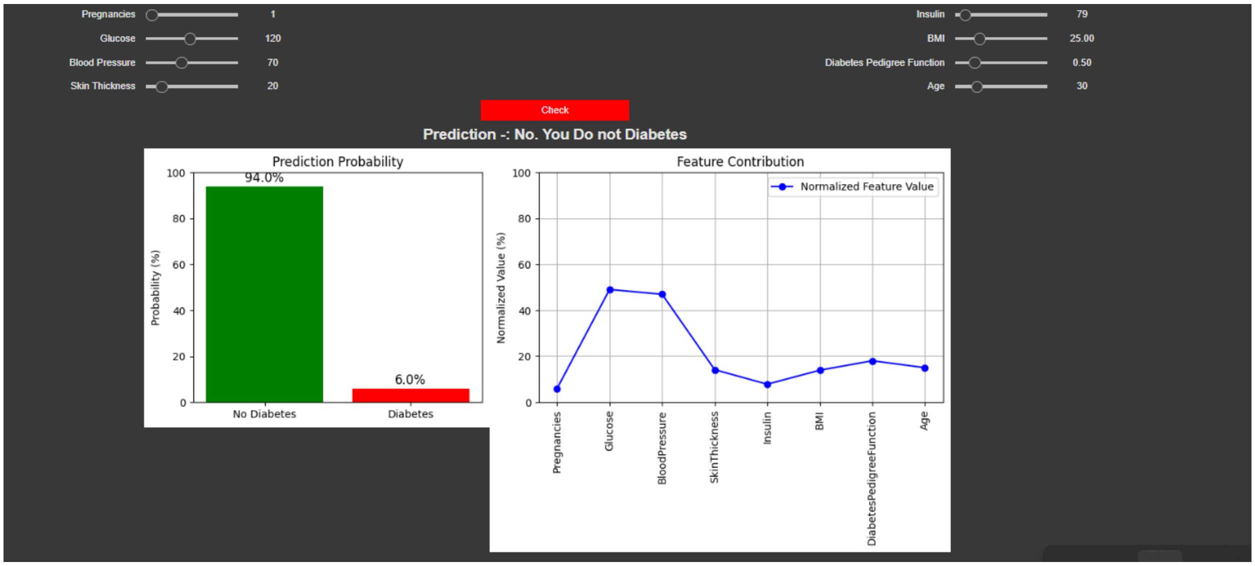
This chart visualizes the normalized values of input features (like Glucose, BMI, Age) to show which features contribute most to the prediction. It provides insight into the model's decision-making.

8. Example User Input + Prediction



Pregnancies 1
Glucose 120
Blood Pressure 70
Skin Thickness 20
Insulin 79
BMI 25.00
Diabetes Pedigree Function 0.50
Age 30

User input UI



Pregnancies 1
Glucose 120
Blood Pressure 70
Skin Thickness 20
Insulin 79
BMI 25.00
Diabetes Pedigree Function 0.50
Age 30

Prediction -: No. You Do not Diabetes

Outputs

9. Codebase Link

https://colab.research.google.com/drive/1XfwawvwvAiMSBnPX49CvVXS4treU06u8?usp=sharing

10. Conclusion / Observation

This project used a Random Forest Classifier to predict diabetes from patient data. Key features like Glucose, BMI, Age and Diabetes Pedigree Function influenced predictions. The interactive UI shows prediction probabilities and feature contributions, making results clear and intuitive. The system demonstrates how machine learning can aid in early detection and support informed healthcare decisions.