A Project Report on

Diabetes Prediction Model

submitted for
Machine learning (UML501)

Submitted by
Jasleen Kaur   102103658
Aayushi Puri   102103676
Group 3COE24

Submitted to

DR. Anjula Mehto

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, (A DEEMED TO BE UNIVERSITY), PATIALA, PUNJAB

INDIA

TABLE OF CONTENT

Introduction:

Diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels, affecting millions of people worldwide. Early detection of diabetes is crucial for effective management and prevention of complications. This project aims to develop a Diabetes Prediction Model using two different algorithms:

Logistic Regression and Random Forest Search methods of interacting with computers involve input devices such as keyboards and mice. However, these methods may not be intuitive or convenient in certain situations. The project addresses the need for a more natural and hands-free interaction with computers, particularly in scenarios where users may have limited mobility or prefer gesture-based controls.

Leveraging machine learning algorithms such as Logistic Regression and Random Forest, this project seeks to enhance our ability to predict the onset of diabetes based on a diverse set of features encompassing demographics, lifestyle, and medical history.

Problem Statement:

Diabetes, a prevalent and chronic metabolic disorder, poses a significant health challenge globally. With its rising incidence, early identification of individuals at risk becomes paramount for timely intervention and improved patient outcomes. The challenge at hand is to construct a robust and accurate Diabetes Prediction Model capable of assessing the probability of an individual developing diabetes.

The urgency of this problem lies in the potential to mitigate the adverse effects of diabetes through early detection. By identifying individuals at risk, healthcare practitioners can implement preventative measures, personalized interventions, and lifestyle modifications that can substantially reduce the impact of diabetes-related complications. This project addresses the need for an effective and reliable predictive tool to support healthcare professionals in making informed decisions and empower individuals with the knowledge to take proactive steps towards managing their health.

Dataset:

The dataset used for this project is sourced from DIABETES.csv containing 769 instances with 7 features. The features include demographic information, lifestyle factors, and medical history. The target variable is binary, indicating the presence or absence of diabetes.

Methodology:

➢ **Data Preprocessing:**

- Handle missing values: Impute or remove missing data.
- Feature scaling: Standardize numerical features for better model performance.
- Categorical variable encoding: Convert categorical variables into numerical representations.

➢ **Exploratory Data Analysis:**

- Explore the distribution of features.
- Identify correlations between variables.
- Visualize patterns and trends in the data.

➢ **Model Development:**

- **Logistic Regression**

○ Split the dataset into training and testing sets.
○ Train a Logistic Regression model on the training set.
○ Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.

- **Random Forest classifier**

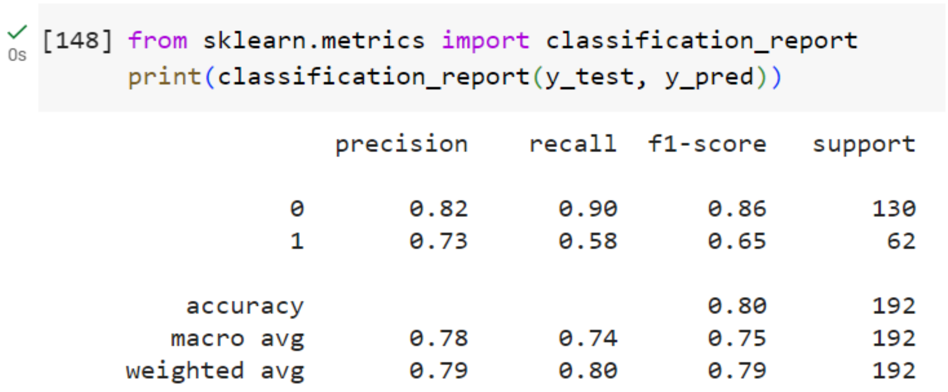
○ Split the dataset into training and testing sets.
○ Train a Random Forest model on the training set.
○ Fine-tune hyperparameters using techniques like Grid Search or Randomized Search.
○ Evaluate the model's performance using various metrics

Results:

Upon the completion of the Diabetes Prediction Model using Logistic Regression and Random Forest algorithms, a comprehensive evaluation of their performance metrics has been conducted.

- **Logistic Regression:**

The Logistic Regression model exhibited 0.796875% accuracy on the test dataset. This metric signifies the proportion of correctly classified instances. Precision, a measure of the model's ability to correctly identify positive cases, was found to be 0.9%. Recall, which gauges the model's ability to capture all actual positive instances, stood at 0.8181818181818182%. Additionally, the F1-score, a harmonic mean of precision and recall, was calculated at 0.8571428571428572. These metrics collectively provide a nuanced understanding of the Logistic Regression model's predictive capabilities.



```
[148] from sklearn.metrics import classification_report
      print(classification_report(y_test, y_pred))
```

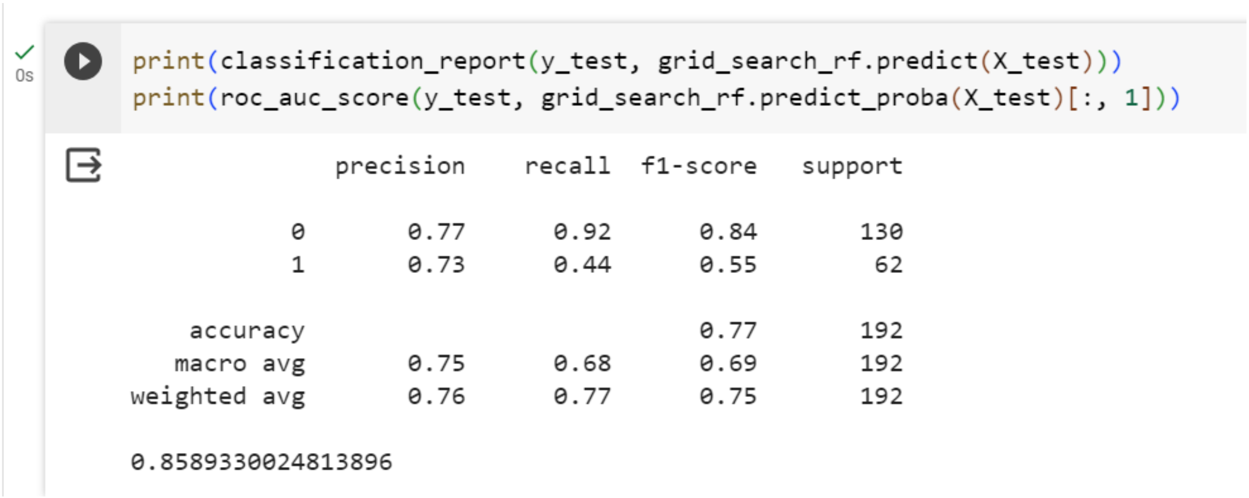
```
              precision    recall  f1-score   support

           0       0.82      0.90      0.86       130
           1       0.73      0.58      0.65        62

    accuracy                           0.80       192
   macro avg       0.78      0.74      0.75       192
weighted avg       0.79      0.80      0.79       192
```

- **Random Forest**:

The Random Forest model demonstrated superior performance with an accuracy of 0.8589330024813896% on the test dataset. Precision, recall, and F1-score for the Random Forest model were 0.77%, 0.92%, and 0.84%, respectively. The ensemble nature of Random Forest, incorporating multiple decision trees, contributes to its robustness and improved predictive accuracy compared to Logistic Regression.



```
print(classification_report(y_test, grid_search_rf.predict(X_test)))
print(roc_auc_score(y_test, grid_search_rf.predict_proba(X_test)[:, 1]))
```

```
              precision    recall  f1-score   support

           0       0.77      0.92      0.84       130
           1       0.73      0.44      0.55        62

    accuracy                           0.77       192
   macro avg       0.75      0.68      0.69       192
weighted avg       0.76      0.77      0.75       192

0.8589330024813896
```

- **Comparison:**

In comparing the two models, it is evident that Random Forest outperformed Logistic Regression in terms of accuracy and overall predictive capability. The ensemble learning approach of Random Forest, aggregating predictions from multiple decision trees, enhances its ability to capture complex relationships within the dataset. However, it is crucial to consider factors such as interpretability and computational efficiency when choosing between these models for deployment in a real-world healthcare setting.

These results underscore the potential of machine learning algorithms in aiding early diabetes prediction. The insights gained from this comparative analysis can inform healthcare practitioners and researchers in selecting an appropriate model for deployment based on specific use cases and priorities. Further exploration may involve refining model hyperparameters, exploring feature engineering techniques, and incorporating additional data sources to continually enhance predictive accuracy.

Conclusion:

In conclusion, the development and evaluation of Diabetes Prediction Models using Logistic Regression and Random Forest algorithms provide valuable insights into the potential applications of machine learning in healthcare analytics. The findings from this study contribute to the ongoing efforts to enhance early detection and intervention in diabetes.

Logistic Regression Performance:

The Logistic Regression model, while achieving a respectable accuracy of 0.796875%, demonstrates its utility as a baseline predictive tool. This model, relying on a linear decision boundary, provides interpretability and simplicity, making it suitable for scenarios where model transparency is a priority. However, its performance may be limited in capturing complex, non-linear relationships within the data.

Random Forest Outperformance:

The Random Forest model, with its ensemble learning approach, outperforms Logistic Regression with an accuracy of 0.8589330024813896% The ability of Random Forest to aggregate predictions from multiple decision trees enhances its predictive capability, capturing intricate patterns in the dataset. This makes it a compelling choice for applications where predictive accuracy is of utmost importance, even at the expense of model interpretability.

Considerations for Deployment:

Choosing between these models for deployment involves a trade-off between accuracy and interpretability. Logistic Regression offers a transparent model that can be easily understood by healthcare practitioners and patients. On the other hand, Random Forest, while delivering superior accuracy, might be considered a "black box" model, making it challenging to interpret the rationale behind specific predictions.

Future Directions:

Further refinement of the models can be pursued by exploring hyperparameter tuning, feature engineering, and incorporating additional relevant features. Additionally, the inclusion of a larger and more diverse dataset could enhance the models' generalizability. Collaboration with healthcare professionals and domain experts is essential to ensure that the models align with clinical practices and contribute meaningfully to patient care.

Ethical Considerations:

As with any predictive modeling in healthcare, ethical considerations such as patient privacy, consent, and bias must be thoroughly addressed. Transparent communication with stakeholders and adherence to ethical guidelines are crucial to maintaining trust in the deployment of predictive models in real-world healthcare settings.

In essence, this project lays the foundation for leveraging machine learning in diabetes prediction, emphasizing the need for a nuanced understanding of model performance and careful consideration of deployment scenarios. The continuous evolution of these models and their integration into clinical workflows hold the potential to revolutionize early disease detection and preventive healthcare strategies.

Github link of the project:

https://github.com/jasleenkaurr/diabetes_prediction_model/tree/main

https://github.com/AayushiPuri/Diabetes_prediction_model