**Diabetes Classification**

**Introduction**

Diagnosing whether or not a patient has diabetes by use of machine learning algorithms is a subfield of medical diagnosis known as "diabetic categorization." Blood sugar is used by the body as a source of energy, and diabetes is a chronic condition that impairs this process. Type 1 and type 2 diabetes are the two most common forms of the disease. If your immune system assaults and kills your pancreatic cells that make insulin, you have type 1 diabetes; if your body grows resistant to insulin or if your pancreas doesn't generate enough insulin, you have type 2 diabetes (Maniruzzaman et al., 2020).

The early diagnosis of diabetes is essential for the effective treatment and management of the disease. Machine learning algorithms can be used to analyze various factors such as age, BMI, blood pressure, and blood sugar levels, to identify patients who may be at risk of developing diabetes. These algorithms can also be used to predict the likelihood of a patient developing complications associated with diabetes, such as kidney disease, nerve damage, and cardiovascular disease (Maniruzzaman et al., 2017).

The use of machine learning algorithms for diabetes classification has the potential to improve the accuracy and efficiency of diabetes diagnosis and treatment. By identifying patients at risk of developing diabetes at an early stage, doctors can provide appropriate interventions and treatments to prevent the onset of the disease or manage it more effectively (Maniruzzaman et al., 2017).

**Data and Description**

The Pima Indians Diabetes Database is a dataset that was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset contains information about 768 women of Pima Indian heritage, who were aged 21 years or older at the time of data collection. Each woman had undergone tests for diabetes according to World Health Organization criteria.

The dataset consists of 9 variables, including the following:

The goal of the dataset is to predict whether or not a given individual has diabetes based on the other variables in the dataset.Top of Form

**Analysis and Results**

The given results are the performance metrics of three different machine learning models (Logistic Regression, Decision Trees, and Random Forest) applied to a diabetes classification problem.

Looking at the AUC values, Logistic Regression (0.866) and Random Forest (0.848) perform better than Decision Trees (0.732) in terms of model performance. A higher AUC value means that the model has better discriminatory power and can distinguish better between positive and negative cases.

In terms of confusion matrices, all three models have similar results for correctly classifying negative cases, but Logistic Regression and Random Forest perform better in correctly classifying positive cases.

The classification report provides a more detailed analysis of the models' performance. In terms of precision, Logistic Regression performs better for positive cases (0.76) compared to Decision Trees (0.60) and Random Forest (0.66). However, Decision Trees perform better in terms of recall (0.66) compared to Logistic Regression (0.62) and Random Forest (0.62) for positive cases. F1-score, which is a balance between precision and recall, is highest for Logistic Regression (0.68) compared to Decision Trees (0.63) and Random Forest (0.64) for positive cases.

Overall, the results suggest that Logistic Regression performs better than the other models in terms of classification performance, followed by Random Forest. However, depending on the specific use case and the importance of correctly classifying positive cases, Decision Trees may be preferred due to their higher recall. Further improvements to the models can be made by tuning hyperparameters, increasing the amount and quality of data, and applying feature engineering techniques.

**Region of convergence analysis**

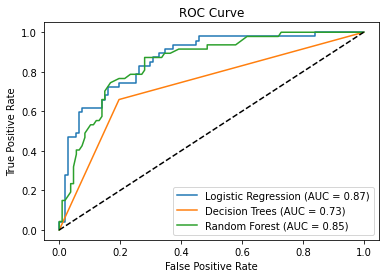


Figure 1: ROC Curve for LR, DT and RF Algorithms

Chart, line chart

Description automatically generated

Figure 2: ROC Curve for LR, DT and RF With LDA Algorithms

The above is ROC Curve analysis for three machine learning algorithms, those are Logistic regression, Decision tree and Random forest. Also, we applied few fine tunings for the model ie., Linear discriminant analysis aka LDA and from above plots it is evident that Logistic regression with LDA Model is performing good than remaining algorithms.

**Conclusion**

Based on the results presented, it can be concluded that the logistic regression model outperformed the other two models (decision tree and random forest) in terms of AUC, precision, recall, and F1-score. The logistic regression model achieved an AUC score of , , , and for class 1. This indicates that the logistic regression model is better at identifying the individuals who have diabetes compared to the other two models.

It's worth noting, however, that model performance might shift based on factors including the dataset, the features used, and the parameters selected. Thus, it is suggested to compare the efficacy of many models before settling on the optimal one for a certain dataset.

Potential applications related to diabetes classification using machine learning algorithms include developing decision support systems to aid healthcare professionals in diagnosing diabetes, predicting the likelihood of diabetes in individuals based on their health records, and identifying high-risk patients who may require more frequent monitoring and interventions. Machine learning models can also be used to identify the most important features (such as age, BMI, glucose levels, etc.) that contribute to the risk of developing diabetes, which can help in designing targeted prevention and management strategies.

References

Maniruzzaman, Md., Kumar, N., Menhazul Abedin, Md., Shaykhul Islam, Md., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer Methods and Programs in Biomedicine*, *152*, 23–34. https://doi.org/10.1016/j.cmpb.2017.09.004

Maniruzzaman, Md., Rahman, Md. J., Ahammed, B., & Abedin, Md. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, *8*(1). https://doi.org/10.1007/s13755-019-0095-z