PREDICTING DIABETES USING MACHINE LEARNING: A COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS: FINAL PHASE ABSTRACT

Abstract

The analysis underscores that diabetes shows a major healthcare challenge and a need for an accurate tool to develop early detection to eliminate the risk. Understanding the problem, the researcher has used machine learning techniques to predict diabetes, which is a chronic condition with an increased global prevalence. The secondary data collection method is used to collect the Pima Indians Diabetes Dataset, which has various clinical and demographic variables such as glucose levels, BMI, age, and family history. The major agenda is to determine the supervised learning models' effectiveness in developing early diagnosis. The reseracher has been using ***"Logistic Regression, Random Forest, and Gradient Boosting models"***, which were developed with the help of Python and scikit-learn. In this research, the researcher follows a coherent methodology, such as using "data preprocessing, exploratory data analysis (EDA), model development, evaluation, and validation". Preprocessing included median imputation of missing values, z-score normalisation of the features, and stratified splitting of data to overcome the imbalance of the classes. The EDA showed that glucose, BMI, and age had the strongest predictive value of diabetes, which sis consistent with the existing clinical knowledge. The models were developed in Python with libraries which included Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, and the unit testing was done with Pytest. Accuracy, precision, recall, F1-score, and ROC-AUC were used to measure performance and k-fold cross-validation was used to obtain generalizability. The critical analysis of results shows that the Gradient Boosting model achieved coherent performance, such as 76% accuracy and an ROC-AUC of 0.83 and outperformed the Random Forest and Logistic Regression models on the test set. It is seen from the model evaluation that the Logistic Regression shows a higher mean ROC-AUC (~0.84) in cross-validation, indicating a higher consistency across folds, even though it had lower test performance. These results indicate the trade-off between complicated ensemble models, which are more predictive, and more interpretable simple models, which can be more trustworthy in generalisation. The unit testing has helped to define whether the developed pipeline works correctly or not, and end-to-end workflow validation helps to confirm that the pipeline is accurate for developing predictions. It can be stated that this project boosted the technical skills in machine learning processes, data analysis in healthcare, as well as performance evaluation. The major issues were in handling class imbalance, optimisation of hyperparameters, and clinical interpretability. Modular ML pipeline implementation enhanced both the efficiency and reproducibility of the code, and model comparison offered cogent information about the accuracy and transparency in medical settings. In the future, the researcher will use an expanded dataset to enhance the

generalizability of the findings, and the explainable AI (XAI) methods, like SHAP or LIME, will be used to accelerate the interpretability and the implementation of the deployment through a web-based interface or API to evaluate the applicability in a real-world setting. It can be concluded that this project shows that the application of ML models, especially ensemble techniques, is viable in helping to detect diabetes early and provide preventive care. Also, it shows the significance of interpretability, fairness, and clinical integration in implementing AI-driven diagnostics. Not only did the research generate a technically sound pipeline, but it also delivered critical lessons regarding the intersection of data science and healthcare practice that will inform future academic and professional practice.