Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220779

Customized and Automated Machine Learning-Based Models for Diabetes Type 2 Classification

Farida MOHSEN[a], Md. Rafiul BISWAS[a], Hazrat ALI[a], Tanvir ALAM[a], Mowafa HOUSEH[a] and Zubair SHAH[a,1]
[a] *College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar*

Abstract This study aims to develop models to accurately classify patients with type 2 diabetes using the Practice Fusion dataset. We use Random Forest (RF), Support Vector Classifier (SVC), AdaBoost classifier, an ensemble model, and automated machine learning (AutoML) model. We compare the performance of all models in a five-fold cross-validation scheme using four evaluation measures. Experimental results demonstrate that the AutoML model outperformed individual and ensemble models in all evaluation measures.

Keywords. Diabetes classification, Machine learning, AutoML, Ensemble model

1. Introduction

Globally, a significant proportion of people have Type 2 diabetes, and many patients may not be diagnosed until seven to ten years after onset. Undiagnosed diabetes can lead to hyperglycemia, resulting in complications such as cardiac stroke, diabetic retinopathy, and other disorders. Therefore, a timely diagnosis of diabetes will help significantly reduce its complications' severity and enhance the patient's quality of life (1).

In the literature, a variety of ML algorithms have been used by researchers for diabetes diagnosis. In one study, Sisodia et al. (2) presented a model for detecting diabetes using support vector machine (SVM), decision tree, and Naive Bayes classifiers. The Naive Bayes classifier achieved the highest accuracy of 76.3%. A recent study by Li et al. (3) on the Pima Indian Diabetes (PID) dataset used the Coefficient of Variation (CV) feature selection to eliminate features with low dispersion. After filtering out irrelevant features, two classifiers were used, decision tree and multilayer perceptron. The multilayer perceptron model obtained an accuracy of 77%. Another study (4) proposed a customized ensemble model for diabetes classification using the Practice Fusion dataset. The study selected 17 features and built a weighted average ensemble model of SVM, RF, and gradient boosting classifiers. The study did not deal with data imbalance. The model achieved an accuracy of 86%. In this work, we address the two main limitations of the existing literature, namely a lower diagnostic accuracy and a less representative dataset, as most of the previous works used the PID dataset, which is

[1] Corresponding Author. Zubair Shah, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar; Email: zshah@hbku.edu.qa.

restricted to female patients only. We developed a combination of effective AutoML and four customized classifiers, namely, SVC (5), RF (6), AdaBoost (7), and an enhanced customized ensemble soft voting model. These models were used to diagnose type 2 diabetes in the Practice Fusion dataset. Besides, we considered imbalanced data distribution between diabetic and non-diabetic patients, thus, ensuring that each class is properly represented. We developed and evaluated the customized models using the Python Scikit-learn library[2] and implemented the AutoML model using the H2O platform. We compared the evaluation results of the proposed models using accuracy, recall, precision, and F1-score metrics. Experimental results demonstrate that the AutoML model outperformed individual and ensemble models in all the evaluation metrics.

2. Methods

2.1. Study Data

Practice Fusion, an electronic medical records company, released 9,948 records from across the United States on Kaggle in 2012. The dataset has records for diabetes patients from 2009 to 2012. From this information, we selected the patient's systolic and diastolic blood pressure, body mass index (BMI), height, weight, age, and gender as features to train our classification models. We followed existing research that focused on these features and used their subset of this data[3] (4).

Data preprocessing: For each feature, the minimum, mean value, and maximum were calculated for each patient over the four years. These calculations result in 15 features for each patient in addition to their gender and age. The target column contains either 0 or 1, indicating the absence or presence of diabetes.

Data Balancing: The dataset is highly unbalanced, with around 1904 diabetic and 8043 non-diabetic samples. This imbalance could lower the classifiers' accuracy; therefore, we selected a subset of 2000 examples from the majority class and combined them with the minority class to create a balanced dataset.

2.2. Classification Models

In this study, we built different customized and AutoML models to classify diabetic patients (see Figure 1).

AutoML Model: We used H2O's AutoML platform, which involves automated tuning and training of many models. We specified a maximum of 10 models and a 5-fold cross-validation as parameters to the AutoML function. The gradient boosting machine (GBM) was the top-performing model on the AutoML leaderboard. We selected this model for evaluation.

Customized Models: We used three different models, SVC, RF, and AdaBoost with their default parameters initially. Then, we performed a grid search for parameter tuning to optimize the model's parameters given a list of possible parameters for each model.

---

[2] https://scikit-learn.org/stable/

[3] https://github.com/akula01/Supervised-Machine-Learning-Ensemble-model-for-Type-2-Diabetes-Prediction

Moreover, we constructed a weighted average soft voting ensemble model that combines our tuned RF, SVC, and Adaboost classifiers as ensemble methods have shown improved performance on classification tasks (8). This ensemble model combines all the three models by taking each classifier's predictions, multiplying them by a weight assigned to each model, and summing the result to get the final classification outcome. To select the best weight for each classifier, we performed manual optimization by trying different combinations. We evaluated the performance of all models using a 5-fold cross-validation scheme.

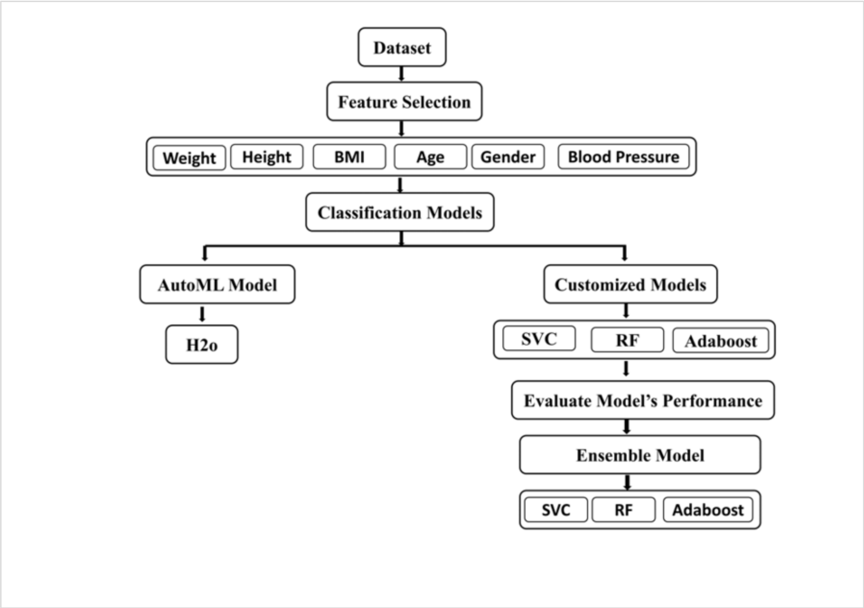

Figure 1. The proposed customized and AutoML models

3. Results and Discussion

To evaluate the performance of the proposed models, we performed a 5-fold cross-validation and calculated the average of the accuracy, recall, precision, and F1-score for each model. Table 1 shows the average scores of the 5-fold cross-validation of the top-performing model on the AutoML leaderboard, the customized ensemble classifier, and the individual classifiers.

By comparing the efficiency of the AutoML model with the customized models, as shown by the results in Table 1, the AutoML model demonstrates superior accuracy, precision, recall, and F1-score of 90%, 95.9 %, 83%, 89%, respectively. It is clear that the ensemble classifier has achieved maximum accuracy, F1 score, precision, and recall of 82.9%, 73.5%, 91.1%, and 75.85, respectively. These results show that an ensemble classifier provides better overall results compared to the individual models since the ensemble combines the predictions of various models.

By comparing the results of the individual classifiers, SVC is superior in terms of accuracy, F1-score, precision, and recall, with scores of 81.8%, 71.9%, 89.1%, and 75.7%, respectively.

Among all the models, the AutoML model achieved the best performance, exceeding the performance of other customized models. This finding indicates that AutoML tools perform reasonably well on diabetes binary classification tasks. Despite the ability of AutoML tools to limit the burden on data scientists, it cannot eliminate the need for trained data engineers to make accurate predictions.

Table 1. Performance comparison of the customized and automated classification models

Metric	SVC	RF	Adaboost	Customized Ensemble	AutoML Model
Accuracy	81.8	79.1	79.9	82.9	90.0
Precision	89.1	69.9	72.4	91.1	95.9
Recall	75.7	71.8	71.4	75.6	83.0
F1-score	71.9	68.93	69.79	73.5	89.0

4. Conclusion

We have developed different automated and customized ML models for the classification of Type 2 diabetes. The customized models included SVC, RF, Adaboost, and the customized ensemble voting of these three, while the automated AutoML model was developed using the H2O platform. All the models were evaluated in a five-fold cross-validation scheme using accuracy, precision, recall, and F1-score. Experimental results show that the AutoML model outperformed the customized ensemble in all metrics with a considerable margin. These results will help advance research on developing machine learning-based methods for the automatic classification of diabetes. However, the results in this work are based on a small downsampled version of the dataset. So, further investigations are needed on a larger dataset. Moreover, further studies are also needed to incorporate more features.

References

[1] Papatheodorou K, Papanas N, Banach M, Papazoglou D, Edmonds M. Complications of diabetes 2016. Journal of diabetes research. Hindawi; 2016.

[2] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018;132:1578-85.

[3] Li T, Fong S. A fast feature selection method based on coefficient of variation for diabetics prediction using machine learning. International Journal of Extreme Automation and Connectivity in Healthcare (IJEACH). 2019;1(1):55-65.

[4] Akula R, Nguyen N, Garibay I, editors. Supervised machine learning based ensemble model for accurate prediction of type 2 diabetes. 2019 SoutheastCon; 2019: IEEE.

[5] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273-97.

[6] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.

[7] An T-K, Kim M-H, editors. A new diverse AdaBoost classifier. 2010 International conference on artificial intelligence and computational intelligence; 2010: IEEE.

[8] Ali H, Garcez ASdA, Tran SN, Zhou X, Iqbal K. Unimodal late fusion for NIST i-vector challenge on speaker detection. Electronics Letters. 2014;50(15):1098-100.