Comparative Analysis of Machine Learning Models for Diabetes Prediction

Zoran Stojanoski, Marija Kalendar and Hristijan Gjoreski
Computer Technologies and Engineering Department, Faculty of Electrical Engineering and Infomation Technologies, "SS. Cyril and Methodius University" in Skopje, Rugjer Boshkovikj Str. 18, Skopje, N. Macedonia
zoran_stojanoski@outlook.com, {marijaka, hristijang}@feit.ukim.edu.mk

Keywords: Machine Learning, Diabetes Prediction, Feature Analysis, ML Models Comparison.

Abstract: This paper focuses on analyzing the benchmark Diabetes dataset which consists of eight commonly measured characteristics. The goal of the study is to present comparative analysis of six machine learning models that predict diabetes, as well as various preprocessing techniques (under-over sampling, feature standardization). The study investigates various approaches and presents results demonstrating that machine learning algorithms can achieve high accuracy results for diabetes prediction, enabling early detection and better outcomes for patients. The paper shows that ensemble learning methods, such as Extra Trees Classifier and Random Forest Classifier, along with appropriate data pre-processing techniques, can lead to 86% accuracy in diabetes prediction classification problems. The paper highlights the potential for machine learning to play a valuable role in the prediction and management of diabetes, leading to improved quality of life and health outcomes for patients.

1 INTRODUCTION

Diabetes is a very common disease affecting millions of individuals in the world today. This disease may be detected and managed early, which could have a major positive impact on quality of life and health outcomes for patients. Thus, this research will focus on the possibilities to detect diabetes from easily measurable features and to enable prediction access to a wider audience. Machine learning, as a technique in common use today has the potential to completely transform prediction in general. Our research demonstrates that implementing machine learning in diabetes predications yields high accuracy results, proving that machine learning algorithms should be considered an important tool for medical practitioners in the early detection and management of the condition.

The main objectives of this research are to provide a simple classifier model solution for predicting diabetes from easily measurable feature variables, with the intention to be later included in a newly designed healthcare system together with a suitable mobile application for prediction of several possible diseases usable to both patients and medical personal.

The research on related work using the same dataset [1] reveals various approaches and different results. Comparing the achieved results through the accuracy metric gives us a better understanding on which methodologies to use in order to have satisfying accuracy, while maintaining efficient and effective models that can be used in a mobile application. One of the first studies showed that without data preprocessing using Naive Bayes and Decision Tree, the achieved accuracy is 79.57% [2]. In another study, a deep learning approach is used, where the authors achieve accuracy of 98% [3]. In this paper, the authors show results for diabetes prediction function using a Naive Bayes classifier of 90%. The Deep Learning approach is superior in this setting, however it has some disadvantages such as requiring significant computing resources for training and also using the model in practice.

The paper is organized as follows. Section two presents the dataset and it characteristics. Section three describes the used data pre-processing techniques, while Section four presents the used ML techniques and an overview of the results. In section five we conclude the paper.

2 DATASET ANALYSIS

The dataset was created by the National Institute of Diabetes and Digestive and Kidney Diseases, from the United States National Institutes of Health [1]. The dataset was devised with the purpose of diagnosing whether a patient has diabetes based on various diagnostic measurements. The measurements come from female Pima Indians who are at least 21 years of age. The focus on this particular population was decided due to the higher diabetes occurrences noticed in practice.

Table 1: Dataset features.

Feature Variables	Description
Pregnancies	Number of pregnancies
Glucose	Glucose level in blood
BloodPressure	Blood pressure measurement
SkinThickness	Patients' Skin Thickness
Insulin	Insulin level in blood
BMI	Body mass index
DiabetesPedigreeFunct.	Diabetes percentage
Age	Patients Age
Outcome	1 - positive, 0 - negative

The dataset includes eight independent medical predictor variables and was properly labelled with a target dependent variable. The dataset consists of entries from 768 patients. The measured features are described in Table 1. Most of the features in Table 1 are self-explanatory. Only the feature Diabetes Pedigree Function (DPF) [4] is a mathematical formula used in genetics to estimate the likelihood of an individual developing diabetes. The DPF takes into account factors, such as family health history and age, which may influence the development of the disease.

The distribution of positive and negative outcomes in the dataset is presented in Figure 1.

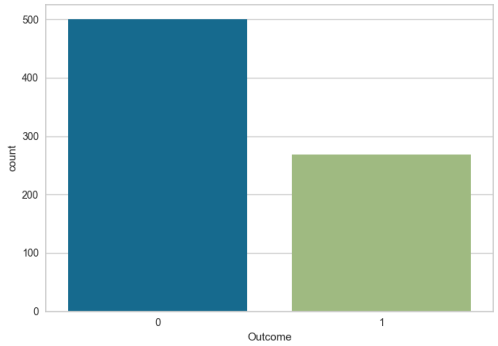

Figure 1: Distribution of positive and negative outcomes.

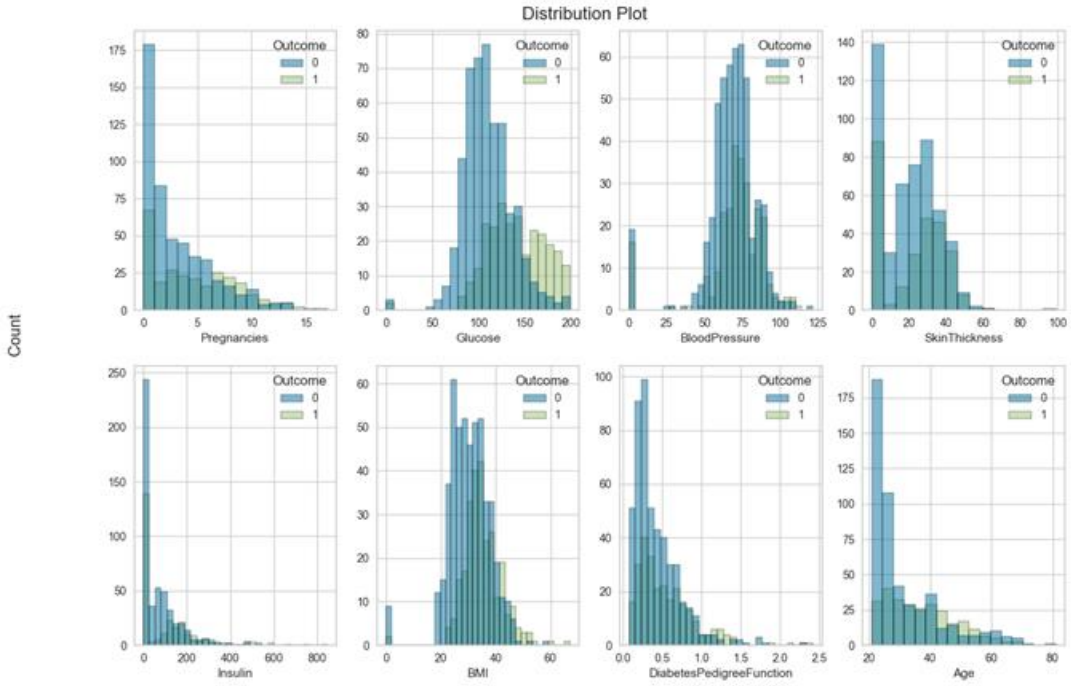

Figure 2: Ranges and distribution of dataset features.

It is easily noticeable there are more negative than positive outcomes in the dataset.

Figure 2 presents a summary of all the different feature ranges and distributions. We can also conclude that the features have varying scales.

2.1 Correlation Matrix

The correlation matrix of the dataset features provides insights into which features are most strongly correlated with the target variable, and which features are strongly correlated with each other (i.e., redundant). The correlation matrix of the used dataset (Figure 3) exhibits that our data feature variables have low correlation between each other.

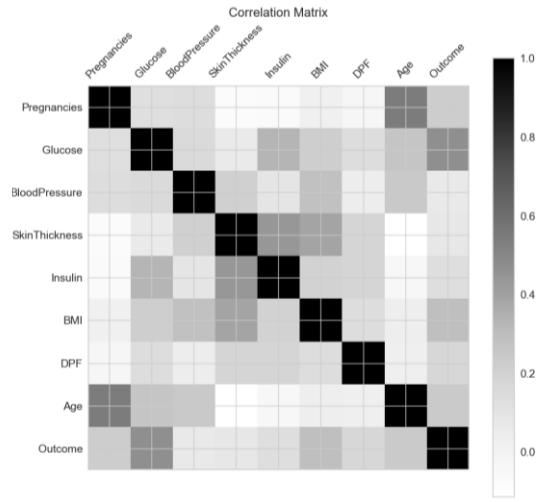

Figure 3: Correlation matrix of the dataset features.

The biggest correlation can be noticed between the age of the patients and their number of pregnancies; and the insulin level with their skin thickness. The glucose level has the biggest impact on the outcome, which is expected.

3 DATA PREPROCESSING

Handling null values in a machine learning dataset is an important pre-processing step, as many machine learning algorithms do not work well with missing values.

Furthermore, to make sure that the data is in a format that is suitable for the machine learning methods it is crucial to carry out additional pre-processing processes, such as feature scaling and normalization, even if there are no missing values. This step ensures that all features contribute equally to the analysis and prevents features with larger ranges from dominating the others.

3.1 Missing Values

The approach to handling missing values depends on the type and amount of missing data, as well as the specific problem we are trying to solve.

Common approaches to handling missing values in a machine learning datasets are dropping rows and interpolating missing values.

From the values analysis of the dataset it is clear that there are no null or missing values, which puts us in a strong position to move forward with the modelling process without having to deal with missing values.

3.2 Over-Sampling

Imbalanced classification is a common problem in machine learning where the target variable is unevenly distributed among the different classes. This can lead to a biased model that performs poorly in predicting the minority class. As presented in Figure 1, we are dealing with an imbalanced dataset.

Over-sampling is one of the techniques used to address this problem by creating synthetic samples of the minority class to balance the distribution of the target variable. Two most commonly used over-sampling techniques are Random Over-sampling [6] and Synthetic Minority Over-sampling Technique (SMOTE) [7].

Random Over-sampling involves duplicating random samples from the minority class in order to balance the distribution of the target variable. This increases the number of samples of the minority class, so that the classifier has a better chance of learning its pattern and making accurate predictions.

The SMOTE method is designed to balance the distribution of the target variable by generating synthetic samples of the minority class.

Both over-sampling methods will be applied for balancing the dataset target classes and the appropriate results are presented in the following sections.

3.3 Feature Standardization

The reason to use feature standardization is to ensure that the features have the same range and are not dominated by one feature with larger values. This allows the algorithm to give equal importance to all the features, rather than being biased towards features with larger values.

Two common techniques used in this research for feature standardization are Standard Scaler and Min Max Scaler. Standard Scaler rescales the input characteristics to give them a mean and standard deviation of 0 and 1, respectively. This is accomplished by dividing the result by the feature's standard deviation after deducting the mean of each characteristic from each data point. The standardized features that are produced have a standard deviation of one and a mean of zero. Min Max Scaler transforms the input features of a dataset to the range [0, 1]. This is done by subtracting the minimum value of each feature from each data point and dividing the result by the range (max - min) of the feature. The transformed features are then ready for input to a machine learning algorithm

In Figure 4 we present the distribution of the standardized features after applying Standard Scaler standardization.

4 CLASSIFICATION MODELS AND RESULTS

4.1 Classification Models

In this study we use six machine learning classification algorithms: Gaussian Naive Bayes [2], Random Forest Classifier [5], Extra Trees Classifier [8], Gradient Boosting Classifier [9] and XGB Classifier [10].

Gaussian Naive Bayes (GaussianNB) is a popular algorithm for classification problems which assumes that the distribution of the features is Gaussian (normal) and independent of each other. This makes it a good choice for problems where the features are continuous or real-valued, and the number of features is relatively small [2].

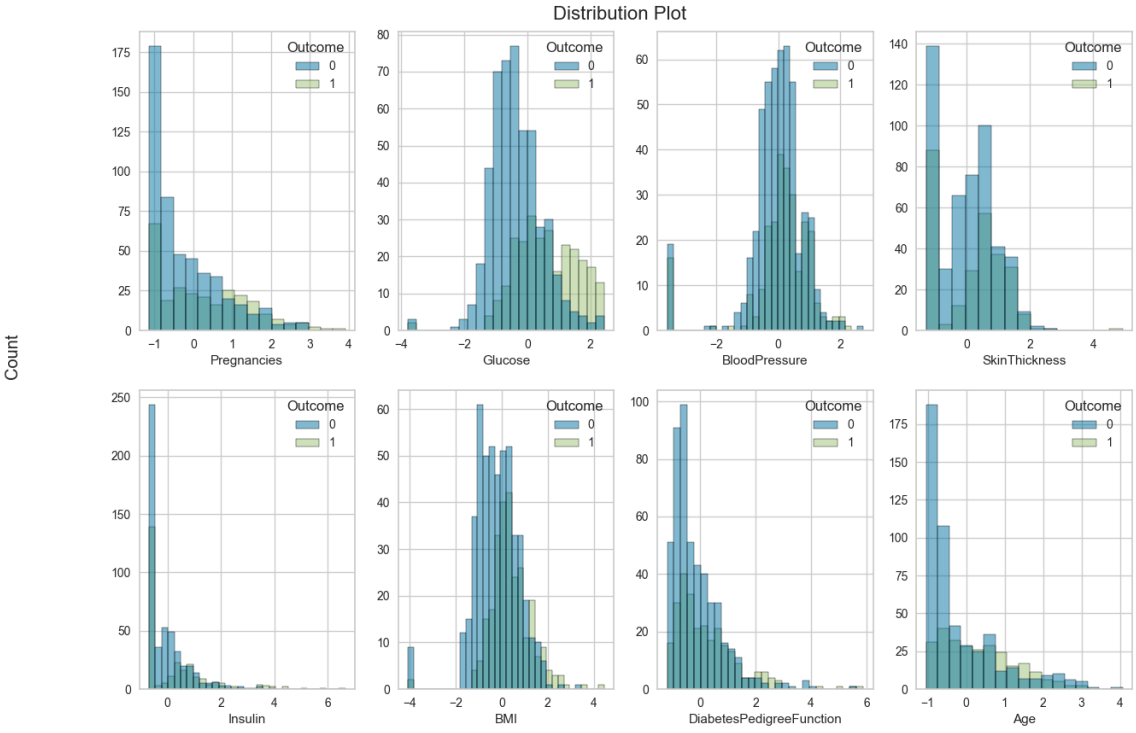

Figure 4: Ranges and distribution of dataset features after Standard Scaler standardization.

The Random Forest classifier (RandomForest) is usually a good choice for classification problems where the data has non-linear relationships, high-dimensional features, and categorical features and where overfitting is a concern [5].

Extra Trees Classifier (ExtraTrees) and Random Forest Classifier are both ensemble learning methods that use multiple decision trees to make predictions in classification problems. Extra Trees Classifier may be a better choice when dealing with many irrelevant features, imbalanced datasets, or where speed is a concern [5, 8].

Gradient Boosting Classifier (GradientBoosting) and XGB Classifier (XGB) are powerful algorithms that can handle a wide range of classification problems. If speed, scalability, or performance is a concern, XGB Classifier may be a better choice, while if ease of use and integration with scikit-learn is a concern, Gradient Boosting Classifier may be a better choice [11].

4.2 Results

Using the chosen models for training the data we can compare how different pre-processing methods perform on our dataset.

The accuracy is calculated as the ratio of the number of correct predictions to the total number of predictions made by the classifier, expressed in percentage. It is used as a simple and intuitive measure of how well the model is performing.

Table 2 presents the accuracies achieved for all the mentioned classifiers and both over-sampling techniques by applying Standard Scaler feature normalization to the data, as a more generally used method in pre-processing. From the results, we can come to the conclusion that for most of the models Random Over-sampling performs better than SMOTE, except for Gradient Boosting Classifier.

Table 2: Prediction accuracies with different over-sampling methods and standard scaler normalization.

Classifier	Rand. over-sampling	SMOTE
GaussianNB	0.76	0.74
RandomForest	0.86	0.85
ExtraTrees	0.88	0.82
GradientBoosting	0.82	0.83
XGB	0.86	0.81

Table 3, on the other hand, presents the accuracies achieved for all the classifiers and both feature normalization techniques by applying Random Over-sampling for solving the imbalanced classification problem in the dataset, again as it a more generally used method. We can conclude that both over-sampling methods work similarly, with slightly better results achieved when using Standard Scaler.

From the results in Table 1 and Table 2, we can conclude that classification models that are built on multiple decision trees give the best results of around 86% accuracy. This is because they avoid overfitting and capture complex relationships between the input features. Our further research will be focused on achieving better results with adjusting various parameters to improve the models' performance, such as the number of estimators and the depth of the trees.

Table 3: Prediction accuracies according to different feature standardization methods.

Classifier	Standard Scaler	Min Max Scaler
GaussianNB	0.76	0.76
RandomForest	0.86	0.85
ExtraTrees	0.88	0.88
GradientBoosting	0.82	0.83
XGB	0.86	0.85

Nevertheless, the results even from this preliminary research prove to have satisfying accuracy, while choosing relatively simple models that can be efficiently used in mobile or embedded applications on devices with limited resources.

5 CONCLUSION

The paper presented an analysis on the benchmark Diabetes dataset through comparison of six machine learning models that predict diabetes. It also presented comparison of different preprocessing techniques (under/over-sampling, feature standardization). The paper showed that ensemble learning methods, such as Extra Trees Classifier and Random Forest Classifier, along with appropriate data pre-processing techniques, can lead to at least 86% accuracy in diabetes prediction classification problems.

The work in this paper demonstrated that machine learning can be considered a valuable tool in the prediction of diseases, and in particular diabetes. The results of this study provide evidence that machine learning algorithms can be trained to identify patients

who are at risk of developing the condition, leading to early diagnosis and better outcomes.

Using Extra Trees Classifier, Random Forest Classifier, or other ensemble learning methods that use multiple decision trees, along with appropriate data pre-processing techniques, can often lead to high accuracy and performance in classification problems. Additionally, these ensemble models are based on Decision Trees, which can effectively be turned into simple decision rules and can run in real time even on devices with limited capacity and processing power.

ACKNOWLEDGEMENT

This work was partially supported by the WideHealth project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952279.

REFERENCES

[1] C.L. Newman, D.J. Blake, C.J. Merz, C.L. Blake, and C.J. Merz, "UCI repository of machine learning databases," 1998

[2] A. H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," 2017 Artificial Intelligence and Signal Processing Conference (AISP), Shiraz, Iran, pp. 209-212, 2017, doi: 10.1109/AISP.2017.8324083.

[3] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," Journal of diabetes and metabolic disorders vol. 19, pp. 391-403, Apr. 2020, doi: 10.1007/s40200-020-00520-5.

[4] M. Das, G. Bhattacharyya, R. Gong, and et all. "Determinants of gestational diabetes pedigree function for pima Indian females," Intern Med Open J. 2022, vol. 6(1), pp. 9-13, doi: 10.17140/IMOJ-6-121.

[5] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, no. 3, pp. 18-22, 2002.

[6] M. Hayaty, M. Siti, and S. Ghufran, "Random and synthetic over-sampling approach to resolve data imbalance in classification," International Journal of Artificial Intelligence Research vol. 4.2, pp. 86-94, 2020.

[7] J. Wang, M. Xu, H. Wang and J. Zhang, "Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding," 2006 8th international Conference on Signal Processing, Guilin, China, 2006, doi: 10.1109/ICOSP.2006.345752.

[8] A. Géron, "Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow," O'Reilly Media, Inc., 2022.

[9] P. Sven, D. Ferran, F. A. Hamprecht, and B. Nadler, "Cost efficient gradient boosting," In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), New York, USA, pp.1550-1560, 2017.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," Proceedings of the 22nd ACM SigKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[11] C. Bentéjac, A. Csörgő, and G.A. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, 2021, pp.1937-1967, doi: 10.1007/s10462-020-09896-5.