Early Prediction of Gestational Diabetes with Parameter-Tuned K-Nearest Neighbor Classifier

Tsehay Admassu Assegie 1*, Tamilarasi Suresh 2, Raguraman Purushothaman 3, Sangeetha Ganesan 4, Napa Komal Kumar 5
1 Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara, Ethiopia
2 Department of Information Technology, St. Peter’s Institute of Higher Education and Research, Chennai, India
3 Department of Artificial Intelligence and Data Science, Madanapalle Institute of Technology & Science, Angallu, Madanapalle, Andhra Pradesh, India
4 Department of Artificial Intelligence and Data Science, RMK College of Engineering Technology, Tiruvarur District, India
5 Department of Computer Science and Engineering, St. Peter’s Institute of Higher Education and Research, Tamil Nadu, India

Email: 1 tsehayadmassu2006@gmail.com
*Corresponding Author

Abstract—Diabetes is one of the quickly spreading chronic diseases causing health complications, such as diabetes retinopathy, kidney failure, and cardiovascular disease. Recently, machine-learning techniques have been widely applied to develop a model for the early prediction of diabetes. Due to its simplicity and generalization capability, K-nearest neighbor (KNN) has been one of the widely employed machine learning techniques for diabetes prediction. Early diabetes prediction has a significant role in managing and preventing complications associated with diabetes, such as retinopathy, kidney failure, and cardiovascular disease. However, the prediction of diabetes in the early stage has remained challenging due to the accuracy and reliability of the KNN model. Thus, grid search hyperparameter optimization is employed to tune the K values of the KNN model to improve its effectiveness in predicting diabetes. The developed hyperparameter-tuned KNN model was tested on the diabetes dataset collected from the UCI machine learning data repository. The dataset contains 768 instances and 8 features. The study applied Min-max scaling to scale the data before fitting it to the KNN model. The result revealed KNN model performance improves when the hyperparameter is tuned. With hyperparameter tuning, the accuracy of KNN improves by 5.29% accuracy achieving 82.5% overall accuracy for predicting diabetes in the early stage. Therefore, the developed KNN model applied to clinical decision-making in predicting diabetes at an early stage. The early identification of diabetes could aid in early intervention, personalized treatment plans, or reducing healthcare costs reducing associated risks such as retinopathy, kidney disease, and cardiovascular disease.

Keywords—Optimization; Machine Learning; Automated Diagnosis; Parameter Tuning; Classification.

I. INTRODUCTION

Diabetes is a chronic disease characterized by high glucose levels in the human blood [2]. Recently it is among one of the most rapidly spreading diseases in the world. Diabetes can be categorized into three types. The first type is Type 1 diabetes occurs when the immune system is weakened, and the cells are unable to produce enough insulin required to regulate the blood glucose level. Type 2 diabetes occurs when the body cells are unable to generate enough insulin or the body fails to utilize the insulin properly.

Gestational diabetes occurs when pregnant women acquire high blood sugar [3]. It occurs at any stage of pregnancy and causes problems for the woman and the baby during and after birth. Gestational diabetes develops in some women when they are pregnant. Most of the time, this type of diabetes goes away after the baby is born. However, women with gestational diabetes have a greater chance of developing type 2 diabetes later in life. Sometimes diabetes diagnosed during pregnancy is type 2 diabetes.

With the advancement in computing and the availability of labeled diabetes datasets in the healthcare sector, machine learning (ML) has improved the diagnosis of diabetes [4]. A pre-processing method such as feature selection is found effective in improving the performance of ML techniques for the accurate prediction of diabetes in the early stages. With feature selection as preprocessing technique, the KNN model achieves an accuracy of 76.25% on early-stage diabetes prediction.

For diabetes prediction, A.H. Osman et al. [5] investigated the performance of KNN on the Pima Indian Diabetes dataset. The study revealed that the SVM model achieves 80.39% accuracy on the test dataset. The result demonstrated higher precision for ML techniques in predicting diabetes. However, pre-processing techniques such as missing value analysis, and cross-validation technique is not investigated to validate the accuracy achieved by the model.

Another research conducted in [6] compared and tested the effectiveness of ensemble learning techniques, namely random forest (RF), extreme boosting (XGB), decision tree (DT), and light extreme boosting (LGB) for diabetes prediction. The result revealed that the highest accuracy of scored 73.5% accuracy with the ensemble learning methods. The grid search method is employed to tune the hyperparameter of the ensemble technique for diabetes prediction for improving their accuracy.

The diagnosis and prediction of diabetes in the early stage significantly reduces the health complications due to diabetes [7]. Studies on the application of machine learning
have proliferated in recent years for the improvement of diabetes treatment. The association between different test symptoms and test results to develop classification models that generalize and classify a given sample into the diabetic or non-diabetic class.

Research article [8] further investigated the performance of an ensemble model for gestational diabetes prediction. The investigation reveals gradient boosting model achieved a receiver operating characteristic curve of 0.71. Different health-related issues affect the performance of machine learning predictors. A study [9] investigated the effect of health-related issues on the performance machine-learning model for diabetes prediction.

Similar research [10] evaluated the performance of deep learning model for diabetes prediction. The study suggested that early prediction of diabetes significantly helps the patient to change life styles and improve their health condition. The study also revealed that deep learning model outperforms other supervised learning model.

In addition, another research article [11], [12], [13] analyzed the effectiveness of several machine learning models such as KNN, DT, RF, Naïve Bayes (NB), support vector machine (SVM), and logistic regression. The comparison of the performance of these machine-learning models reveals that the logistic regression model outperforms other models with an accuracy score of 75.32%.

A research article [14], [15], [16] predicted diabetes by developing a KNN model. The KNN model performance improves when the model is trained on pre-processed PIMA Indian diabetes dataset (PIDD). The study suggested that pre-processing with feature scaling increases the performance of KNN with an accuracy of 8.48% on PIDD.

Diabetes has several consequences such as an increased risk of retinopathy, hypertension or high blood pressure, renal damage, and cardiovascular disease [17], [18], [19], [20], [21]. Recently, diabetes is becoming one of the most prevalent diseases affecting numerous people all over the world. The advancement in machine learning has become significant in the reduction of the ever-growing risk of diabetes through early prediction of diabetes to avoid latter associated consequences such as retinopathy, and cardiovascular disease.

For the early diabetes prediction systems, the existing studies employ different machine learning models, trained on diverse amounts of data, to predict the presence of gestational diabetes. The KNN model fails to categorize the diversity of sample instances with a relatively small set of data. To address this issue, this study proposes a KNN classification model based on distance measurement to predict the presence of gestational diabetes. Furthermore, existing prediction models employ a common set of factors for constructing the model. According to the healthcare professional principle in the diagnosing process, a wide swath of health conditions results in different disease diagnoses and treatment decisions [22], [23], [24], [25], [26]. However, the existing KNN-based gestational diabetes prediction has scope for improvement for accurate prediction of gestational diabetes. For instance, a study [27] reveals 71.4% accuracy on gestational diabetes prediction with the KNN model even though the KNN model outperforms the SVM model [28]. Similarly, another research article [29] compared the performance of supervised learning algorithms. The result revealed that the KNN model achieved 74.89% accuracy in gestational diabetes prediction. The findings revealed improving the performance of the KNN model is recommended for further research. Thus, this study proposes a novel KNN model for gestational diabetes prediction. Overall, the objective of this study is summarized as follows:

- To develop a KNN model that can learn the pattern from the diabetes dataset and perform automated analysis.
- To apply the grid search technique and improve the performance of the KNN classifier.
- To apply correlation analysis and extract underlying patterns from the diabetes dataset, which will assist in determining the diabetic tendency of a new set of patients in the early stage.

The remainder of the study is structured as follows: Section 2 describes the method, explaining the dataset, preprocessing steps, and the proposed KNN model. Section 3 presents the results and discussion. The result section compares the performance of the KNN model on the original and feature-scaled dataset with the help of an accuracy metric. Finally, Section 4 concludes the study. The conclusion presents the results obtained implication of the study, limitations, and recommendations for feature work.

II. Method

The study considered four procedures for developing a predictive model for gestational diabetes. The procedures involve data acquisition, data preprocessing, and development of the KNN model and analysis of the performance of the developed model on the test set. Grid search and hyperparameter tuning have been widely employed for improving the performance of the KNN algorithm. For instance, research article [30], [31], [32]. The study employed feature scaling due to its significance in improving the performance of the KNN model for diabetes diagnosis [33]. Each step is discussed in Section A, and Section B. Fig. 1 demonstrates the flow diagram for the procedure followed to develop the proposed KNN model.
A. Data Acquisition

The PIDD is employed for the development of the KNN model for gestational diabetes prediction. The PIDD dataset is one of the standard datasets previously employed in several studies [34], [35], [36], [37] for the development of gestational diabetes prediction, prognosis, and diagnosis with machine learning algorithms. The dataset is collected from the online Kaggle repository available at the following link which if previously employed by the study [38], [39], [40] can be downloaded at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. There were 768 instances each with 8 features. The dataset consists of several medical predictor variables and one target variable, outcome. The predictor variables include the number of pregnancies the patient has had, their body mass index (BMI), insulin level, age, blood pressure (BP) glucose level, skin thickness, and diabetes pedigree function (DPF). The dataset is analyzed for missing values and the features are scaled with a standard scaler. In the analysis and dataset feature, exploration panda’s data frame is employed.

B. Data Preprocessing

In data preprocessing, feature scaling with a standard scaler is performed. The standard scaler is one of the most commonly employed types of feature scaling methods for distance-based learners such as KNN, and SVM [41], [42], [43], [44]. Standard scaler improves the performance of distance-based learning algorithm [45], [46], [47], [48], [49].

Feature scaling through standardization (or Z-score normalization) can be an important preprocessing step for many machine-learning algorithms [50], [51], [52], [53]. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one. The standard scaler is defined by the mathematical equation [54], [55], [56], [57], [58], [59], [60] given in equation (1).

\[ Z_{scaled} = \frac{x - \mu}{\sigma} \]  

(1)

Where \( \mu \) is the mean and \( \sigma \) is the standard deviation.

III. RESULTS AND DISCUSSION

This section presents the result of the constructed KNN model on gestational diabetes prediction. Firstly, the performance of the KNN model is evaluated on the test dataset with random K values. Secondly, the grid search technique is employed to evaluate cross-validated accuracy on K values in the range 1 to 25 and the optimal value of K is determined. Then, the performance of the KNN model is compared on random K values and the optimized K value.

A. Optimization of KNN with K value

The performance of the developed KNN model is optimized by tuning the K values, which produces the highest possible accuracy for the KNN model. After selecting the combination of hyperparameters with the highest accuracy for the KNN model, the K value that produces optimal test accuracy is determined. The variability of the training and test accuracy for different values of K is revealed in Fig. 2. As indicated in Fig. 2, the KNN model score higher test accuracy of 77.21% with K value=13.

![Fig. 2. The optimization of K value for standard KNN model](image)

In addition to the feature scaling, and grid search for improving the performance of KNN in predicting gestational diabetes, the impact of each feature on test instance is analyzed with the help of Shapley Additive Explanation (SHAP). Fig. 4 indicates that for test instances with feature values (number of times pregnant=6, age=50, glucose 148, blood pressure-72, skin thickness=35, body mass index=33.6, and diabetes pedigree function=0.627) with the outcome of diabetes, the age of the patient significantly impacts the prediction of the KNN model.

In Fig. 5 indicates the variability of the cross-validation score for varying values of K. Five-fold cross-validation score of the KNN model indicated in Fig. 5 reveals that the variability of the training score tends to decrease with an
increase in the value of K. It is evident that with K values greater than 10 the testing, and the cross-validation score tends to converge revealing better performance on diabetes prediction.

Fig. 4. The relative impact of features on test instance using KNN model

![Validation Curve for KNeighborsClassifier](image)

Fig. 5. The validation score for KNN model

### IV. CONCLUSION

In this study, hyperactive parameter tuning, and standardization of input features are employed for preprocessing methods. With these pre-processing methods, the KNN model performance improves for diabetes prediction. The result of the experiment revealed that the constructed KNN model outperforms other existing KNN models. Thus, it is suggested that feature scaling significantly improves the performance of the KNN model for the prediction of diabetes. The study improves the existing KNN model with preprocessing and grid-searching techniques. The developed model significantly improves the accuracy of the previous studies. However, the performance of the KNN model and other pre-processing methods with other datasets seek further studies to validate the findings of the current study.

The limitations of this study include the study lack of generalizability of the findings across various data sources to confirm the reliability of the KNN model performance improvement with preprocessing methods such as feature scaling and hyperparameter tuning. The researchers recommend further investigations using different datasets and validate the effectiveness of the KNN model and preprocessing methods. Additionally, a comparative analysis of the KNN model with other machine learning algorithms commonly used for diabetes prediction such as SVM, and RF is recommended for future work.

**ACKNOWLEDGMENT**

The authors would like to acknowledge Injibara University of the provision of facilities to complete this work.

**REFERENCES**


Diabetic analytics: proposed conceptual data Web
Machine learning based diabetes prediction and
19. H. Lu, J. Hirt, J. Yang, L. Mackillop, and D. Clifton, “Standardizing the assessment of caesarean birth using an oxford caesarean prediction

Tsehay Admassu Assegie, Early Prediction of Gestational Diabetes with Parameter-Tuned K-Nearest Neighbor Classifier


