➔ Regular Research Paper – NS

Diabet Prediction with KNN

Sena GORAL
Mehmet Akif Ersoy University, Turkey
scelik@mehmetakif.edu.tr

Abstract

Diabetes, a disease with an increasing incidence in the world, has a significant impact on human life. It causes damage to many parts of the body. Therefore, significant losses may also be high. Early diagnosis is very important in order to prevent diabetes or to minimize losses. Researchers have attached great importance to the classification of the disease with various classification methods. In this study, the values in the data set were studied with the k-NN algorithm. The performance of the K-NN algorithm on the data set was examined. In addition, the accuracy value was increased by expanding the data set with cross-validation. It is thought that the study will provide convenience for both the patient and the specialists.

Keywords: *Diabet, k neighbor algorithm, disease, classification, machine learning.*

1. INTRODUCTION

In cases such as underproduction or insufficient production of the insulin hormone secreted in the body, a chronic disease condition occurs [1]. This disease is called diabetes. Diabetes can cause organ loss and death. This disease may show characteristic symptoms such as thirst, weight loss and blurred vision[2,3]. The fact that various factors play a role in diabetes and human error complicates the diagnosis of this disease. A blood test does not provide sufficient information for the correct diagnosis of the disease [4]. Therefore, early diagnosis and treatment should be initiated.

With the technological developments, artificial intelligence and learning techniques have made it easier to diagnose many diseases. In this case, the diagnosis of diseases is completed in a shorter time. Many researchers use machine learning algorithms in the diagnosis of diseases. The reason why these algorithms are preferred is that they reduce the cost in diagnosing different diseases and also give more accurate and faster results.

For the classification process, which is the subject of the study, diabetes, one of the diseases that plays an important role in the formation of many fatal diseases, was chosen. The incidence of diabetes is increasing day by day. Studies show that the annual cost of diabetes is increasing every year. Diagnosis and treatment costs of diabetes and diabetes-related diseases are quite high. In addition, there is a decrease in the work capacity of the individual, a decrease in the average lifespan and various costs for the relatives of the patient. This situation shows the importance of preventive activities in health. This situation increases the importance of early and accurate diagnosis of diseases.

In this study, it is aimed to make the correct classification of the disease with the KNN algorithm in order to diagnose diabetes mellitus early. In addition, the necessary performance evaluation results were examined with KNN. In this study, it is aimed to make the correct classification of the disease with the KNN algorithm in order to diagnose diabetes mellitus early. In addition, the necessary performance evaluation results were examined with KNN.

2. RELATED WORKS

Kaur and Kumari (2018) studied disease detection with supervised machine learning algorithms [5]. Parashar et al. Comparisons were made using LDA-DVM combination and feedforward neural networks classification techniques. DVM combination accuracy reached 75% [6]. A new model was developed by Ahmed to classify diabetic treatment plans. At the end of the study, 70.8% success was achieved with the developed algorithm[7]. In 2018, Joshi and Chawan developed a system for diabetes prediction using 7 different features in their study. They used DVM, logistic regression and ANN ANN. The best results were obtained with SVM [8]. Al-Halal et al. used different classification algorithms in the diagnosis of diabetes. They achieved a success rate of 66.19% with K-NN, 72.66% with Naive Bayes, and 73.72% with Random Forest [9]. Tiwari et al. used important feature selection by Random Forest and Recursive Screening to predict diabetes. When XGBoost and ANN are compared, XGBoost provided higher accuracy than ANN with a rate of 78.91% [10].

3. MATERIAL METHOD

Based on the data set, KNN classification is used for diabetes disease diagnosis. K-nearest neighbor algorithm is one of the machine learning algorithms. The KNN method can be used to predict various types of labeled data. An object is classified by majority vote of neighboring data. Assignment to the most common class among the k nearest neighbors is performed. Checks the training table for the k nearest neighbors.

The KNN algorithm makes predictions on two fundamental values. These are distance and k (neighborhood number). Distance is calculated from the distance of the predicted point to other points. There are several methods for this. The Euclidean method is used in this study. K determines how many nearest neighbors will be calculated over. K value directly affects the result. If K=1, the probability of overfit is very high. Even if it is a very large value, very general results are produced. Therefore, finding the optimum k value is the most important issue.

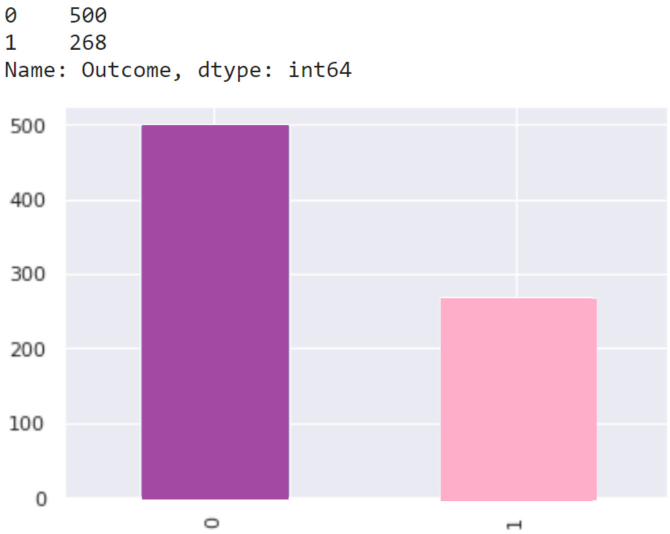
First of all, the data set used in the study was examined. In the data set, it was seen that there were 9 different features belonging to 768 people. These features are shown in Table 1.

Table 1. Features.

1	Pregnancies
2	Glocuse
3	Blood Pressure
4	Skin Thickness
5	Insulin
6	BMI
7	Diabetes Pedigree Function
8	Age

9	Outcome

First of all, necessary operations were performed with fillna() for missing values in the data set. The distribution of the features after the operations for the missing values is as in Figure 1.



```
0    500
1    268
Name: Outcome, dtype: int64
```

Figure 1. Number of diabetic and non-diabetic patient

The data set used is reserved for training and testing. It provides an accurate distribution of certain types of data points with cross validation. At the same time, it was tried to increase the accuracy by enabling the system to be tested with more situations with cross validation. The K-fold cross validation method aims to create the best model during the testing phase of the model. In classification models, k-cross validation method is used to prevent overfitting and incomplete learning.

In the k-cross validation method, the train set to be used in the train process is mixed and divided into k subsets of equal size. These operations are repeated k times, and in each iteration, the next subset is removed from the training dataset and used as the test set. When the evaluation process is completed for all parts, the cross validation model produces a performance measure and results for all data.

The model fort he classification applications is evaluated according to the performance for correct and wrong classification. The basic concepts used to evaluate the success of the model are precision, recall and f1 values. Confusion matrix is used to evaluate these criteria. In the confusion matrix seen in Figure 3, the rows represent the class numbers predicted by the model, and the columns represent the actual class numbers in the test set. The data set has been studied for the most appropriate k in the prepared system. As a result of this, as seen in Figure 2, k was determined as 11 for the most appropriateresult.The score calculated for K=11 is 0.765625.

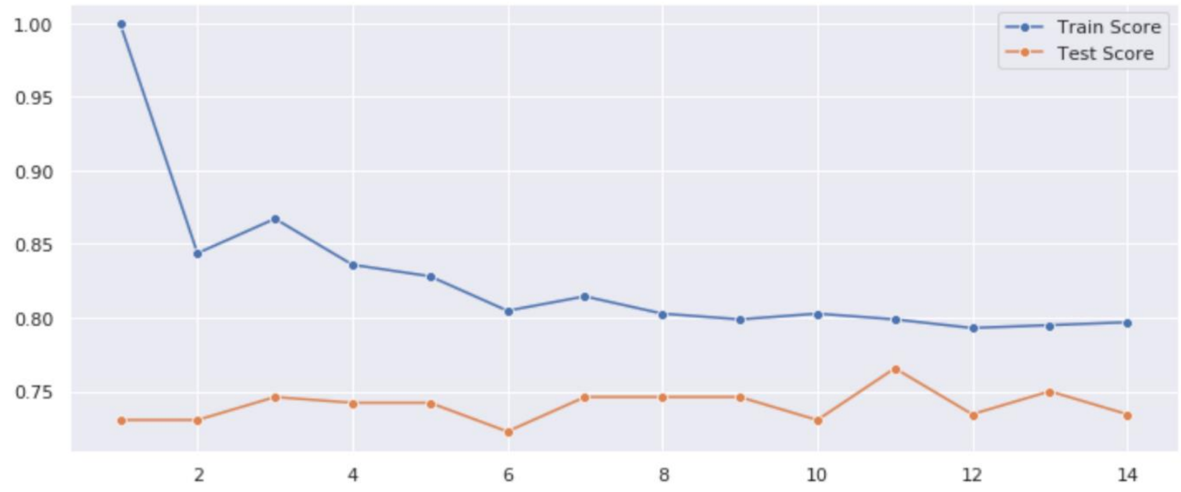


Figure 2. Test and Train Score Graphics

The confusion matrix created for the test data of the system is shown in Figure 3.

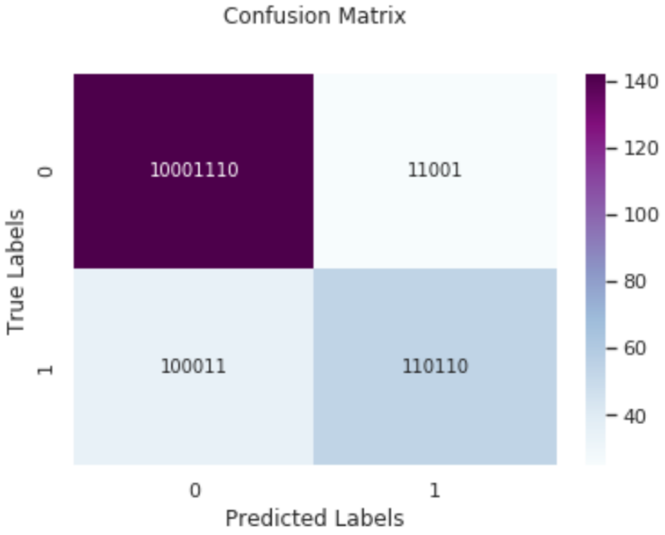


Figure 3. Confusion Matrix

3.1. Classification Report

Report which includes Precision, F1-Score, and the Recall.

Precision Score: indicates how many of the positively predicted values are positive. especially important when the cost of false positive values is high. Here it is with the value of 0.788, which is good enough for the precision metric.

TP – True Positives

FP – False Positives

Precision – The value of accuracy regarding positive detections.

Precision = TP/(TP + FP)

Recall Score: Indicates how many of the values that should be predicted positively are predicted positively. Recall score should be high as possible. A recall greater than 0.5 is good.

FN – False Negatives

Recall = TP/(TP+FN)

F1 Score: The reason why it is a harmonic mean instead of a simple mean is that we should not ignore the extreme cases. If it were a simple average calculation, a model with a Precision value of 1 and a Recall value of 0 would have an F1 Score of 0.5, which would mislead us. F1 Score value was employed rather than the Accuracy because of not to make an incorrect selection of the model formation (That is important for the data sets, which are not distributed evenly). In addition, F1 Score is very important as there is a need for a measurement metric that will include not only False Negative or False Positive but also all error costs.

F1 = 2 x (precision x recall)/(precision + recall)

The generated values are shown in Table 2.

Table 2: Classification Report

	Precision	Recall	F1- Score	Support
0	0.80	0.85	0.83	167
1	0.68	0.61	0.64	89
Micro avg	0.77	0.77	0.77	256
Macro avg	0.74	0.73	0.73	256
Weighted avg	0.76	0.77	0.76	256

3.2. ROC(Receiver Operating Characteristic)

The ROC is often used in classification problems. The ROC is known as metric evaluation that indicates whether the model is working well or not. It is an increasing function between (0,0) and (1.1). The higher the ROC value, the higher the classification success of the model. In Figure 4, the ROC Curve is shown according to the value of 11 determined for the best k. Score 0.8193500639171096 was calculated for the area under the ROC curve.

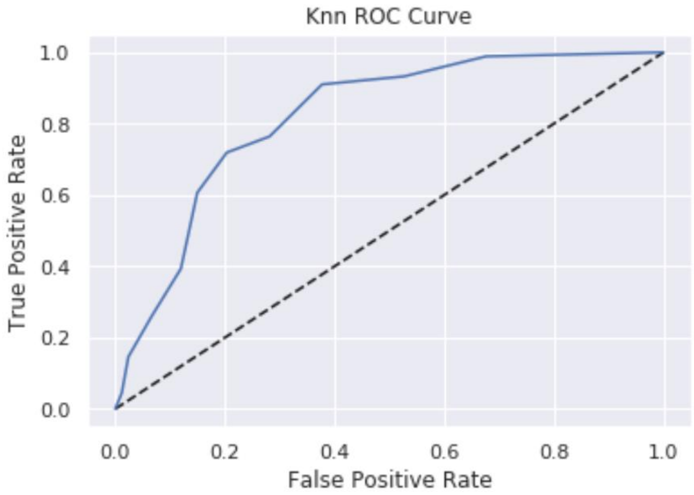


Figure 4. ROC Curve

3. CONCLUSION

The risk of developing diabetes is increasing day by day for all people. In this study, it is emphasized that early diagnosis of diabetes is very important in the treatment phase. As machine learning methods and tools that detect diabetes develop, the contribution to treatment will be greater. At the same time, as the number of data sets and people increases, the reality of the system will increase. The main purpose of this study is to achieve success in diagnosis. As a result of arranging the data set used in accordance with the system, diabetes diagnosis was made with k-NN. At the same time, the status of k-NN achievement performance was examined.

REFERENCES

[1] R. D. Howsalya Devi, A. Bai, and N. Nagarajan, “A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms,” Obesity Medicine, vol. 17, p. 100152, 2020.

[2] A. Adler et al., “Reprint of: Classification of Diabetes Mellitus,” Diabetes Research and Clinical Practice, vol. 0, no. 0, p. 108972, In Press, 2021.

[3] P. Zimmet, K. G. M. M. Alberti, and J. Shaw, “Global and societal implications of the diabetes epidemic,” Nature, vol. 414, no. 6865, pp. 782–787, 2001.

[4] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, “Diabetes Diagnostic Prediction Using Vector Support Machines,” Procedia Computer Science, vol. 170, pp. 376–381, Jan. 2020.

[5] Kaur, H., Kumari, V., 2018. Predictive modelling and analytics for diabetes using a machine learning approach.Applied Computing and Informatics.

[6] Parashar, A., Burse, K., Rawat, K., 2014. A Comparative approach for Pima Indians diabetes diagnosis using ldasupport vector machine and feed forward neural network. International Journal of Advanced Research in Computer Science and Software Engineering, 4(11), 378-383.

[7] Ahmed, T. M., 2016. Developing a predicted model for diabetes type 2 treatment plans by using data mining. Journal of Theoretical and Applied Information Technology, 90(2), 181.

[8] Joshi, T. N., Chawan, P. P. M., 2018. Diabetes Prediction Using Machine Learning Techniques. International Journal of Engineering Research and Application (Ijera), vol. 8, no.1, pp. 9-13, 2018.

[9] Al Helal, M., Chowdhury, A. I., Islam, A., Ahmed, E., Mahmud, M. S., & Hossain, S., 2019. An optimization approach to improve classification performance in cancer and diabetes prediction. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-5, IEEE.

[10] Tiwari, P., Singh, V., 2021. Diabetes disease prediction using significant attribute selection and classification approach. In Journal of Physics: Conference Series, Vol. 1714, No. 1, p. 012013.

JOMUDE
http://www.jomude.com