Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 24, S. 53-59, Nisan 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 24, pp. 53-59, April 2021

Research Article

Diabetes Prediction Using Machine Learning Classification Algorithms

Shamriz Nahzat[1*], Mete Yağanoğlu[2]

[1] Atatürk University, Department of Computer Engineering, Faculty of Engineering, Erzurum 25240, Turkey, (ORCID: 0000-0002-0750-6392), shamriz.nahzat19@ogr.atauni.edu.tr

[2] Atatürk University, Department of Computer Engineering, Faculty of Engineering, Erzurum 25240, Turkey, (ORCID: 0000-0003-3045-169X), yaganoglu@atauni.edu.tr

(2nd International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF)-10–12 March 2021)

(**DOI:** 10.31590/ejosat.899716)

ATIF/REFERENCE: Nahzat, S. & Yağanoğlu, M. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. *European Journal of Science and Technology*, (24), 53-59.

Abstract

Artificial intelligence's use in health systems has evolved substantially in recent years. In medical diagnosis, machine learning (ML) has a wide variety of uses. Machine learning techniques are used to forecast or diagnose a variety of life-threatening illnesses, including cancer, diabetes, heart disease, thyroid disease, and so on. Chronic diabetes is one of the most common diseases worldwide and making the diagnosis process simpler and quicker would have a huge effect on the treatment process.
The fundamental goal of this work is to prepare and carry out diabetes prediction using various machine learning techniques and Conduct output analysis of those techniques to find the best classifier with the highest accuracy. This study examines diabetes prediction by taking different diabetes disease-related attributes. We use the Pima Indian Diabetes Dataset and applied the Machine Learning classification methods like K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT) for diabetes prediction. The models used in this analysis have various degrees of accuracy. This study shows a model that can correctly forecast diabetes. In comparison to other machine learning methods, the random forest has high accuracy in forecasting diabetes, according to the findings of this study.

Keywords: Machine learning (ML), Classification, Artificial Neural Network (ANN), Random Forest (RF), Decision Tree (DT).

Makine Öğrenimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini

Öz

Yapay zekanın sağlık sistemlerinde kullanımı son yıllarda önemli ölçüde gelişmiştir. Tıbbi teşhiste, makine öğreniminin (MÖ) çok çeşitli kullanımları vardır. Makine öğrenimi teknikleri, kanser, diyabet, kalp hastalığı, tiroid hastalığı v.b. dahil olmak üzere hayatı tehdit eden çeşitli hastalıkları tahmin etmek veya teşhis etmek için kullanılır. Kronik diyabet dünya çapında en yaygın hastalıklardan biridir ve teşhis sürecini daha basit ve daha hızlı hale getirmek tedavi süreci üzerinde çok büyük bir etkiye sahip olacaktır.
Bu çalışmanın temel amacı, en yüksek doğrulukla en iyi sınıflandırıcıyı bulmak için çeşitli makine öğrenimi tekniklerini kullanarak diyabet tahminini yapmak ve bu tekniklerin çıktı analizini yapmaktır. Bu çalışma, diyabet hastalığıyla ilgili farklı özellikler alarak diyabet tahminini incelemektedir. Pima Indian Diyabet Veri Kümesini kullanıyoruz ve K-En Yakın Komşu (KNN), Rastgele Orman (RO), Destek Vektör Makinesi (DVM), Yapay Sinir Ağı (YSA) ve Karar Ağacı (KA) gibi Makine Öğrenimi sınıflandırma yöntemlerini diyabet tahmin etmek için uyguladık. Bu analizde kullanılan modeller çeşitli doğruluk derecelerine sahiptir. Bu çalışma, diyabeti doğru

[*] Corresponding Author: shamriz.nahzat19@ogr.atauni.edu.tr

bir şekilde tahmin edebilen bir model göstermektedir. Bu çalışmanın bulgularına göre, diğer makine öğrenimi yöntemlerine kıyasla rastgele orman (RO), diyabet tahmininde yüksek doğruluğa sahiptir.

Anahtar Kelimeler: Makine öğrenimi (MÖ), Sınıflandırma, Yapay Sinir Ağı (YSA), Rastgele Orman (RO), Karar Ağacı (KA).

1. Introduction

Diabetes (DM) is quite possibly the most well-known disease where patients' body capacity to create and react to insulin is impeded, which may bring about expanded glucose levels in the blood (Lonappan et al.,2007). Different illnesses take arise alongside Diabetes, for example, Coronary Artery Disease (CAD), Coronary Kidney Disease (CKD), Chronic obstructive aspiratory sickness (COPD), Hypertension (HTN) and Hypothyroidism. These illnesses don't provide quite a bit of some insight until they become unsure. Thus, early finding out of these illnesses close by with suitable treatment can help the patient restoring to a superior condition (Kang et al.,2013).

Diabetes can be partitioned into two classes, type 1 diabetes (T1D) and type 2 diabetes (T2D). Patients with type 1 diabetes are regularly more youthful, generally under 30 years of age. The regular clinical indications are expanded thirst and continuous pee, high blood glucose levels (Iancu et al., 2008). This kind of diabetes can't be restored successfully with oral drugs alone and the patients have required insulin treatment. Type 2 diabetes happens all the more generally in moderately aged and old individuals, which is frequently connected with the event of fatness, hypertension, dyslipidemia, arteriosclerosis, and different illnesses (Robertson et al., 2011).

As indicated by World Health Organization (WHO), around 422 million individuals are experiencing diabetes especially from low or inactive pay nations. What's more, this could be expanded to 490 billion up to the time of 2030. In any case, diabetes is widespread in different countries such as China, Canada, India, and so on.

These days diabetes is a significant reason for death on the planet. Early forecast of sickness such as diabetes can be monitored, saving human life.

Today, machine learning techniques are used to predict or diagnose various life-threatening illnesses such as cancer, diabetes, heart disease, thyroid, and so on.

To achieve this, this work investigates forecasting of diabetes by taking different characteristics identified with diabetes sickness. For this reason, we utilize the Pima Indian Diabetes Dataset; we apply different Machine Learning (ML) categorization and outfit Techniques to anticipate diabetes. ML is a procedure that is utilized to instruct computers or machines expressly. Different ML methods give effective outcomes to gather Knowledge by building different grouping and outfit models from gathered dataset.

Various machine learning methods may be capable of forecasting, but finding the right approach is difficult. Therefore, in this study the, K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Decision Tree (DT) algorithms are applied to forecast diabetes and their performance has been examined.

2. Related works

This section of the paper is dedicated to some of the research works to predict or detect diabetes using machine learning techniques.

Soni et al. (2020) have utilized various Machine learning techniques like SVM, DT, KNN, Random Forest, Logistic Regression, and Gradient Boosting and achieved 77 percent accuracy by using the RF algorithm.

Sarwar et al. (2018) in here SVM and KNN classification models gives the highest accuracy of diabetes forecast. By using 768 records, it gives 77% accuracy.

Tejas et al. (2018) introduced Diabetes forecasting Using Machine Learning methods and to plans to anticipate diabetes by three classifiers including: SVM, Logistic regression and ANN. This study proposes a successful strategy for prior find out of the diabetes sickness.

Parashar et al. (2014) introduced a classification technique which was the LDA method, and then combined SVM classifier with Feed Forward Neural Networks. The SVM classifier shows 75.65% accuracy. Al Helal, et al. (2019) developed three categorization models which are the KNN, Naïve Bayes, and RF then their final accuracy was according to 66.19%, 72.66%, 73.72%. They were used in the Weka tool.

3. Material and Method

The fundamental goal of this work is to prepare and carry out Diabetes Prediction Using Various Machine Learning Techniques and Conduct Output Analysis of those techniques to find the best classifier with the highest accuracy. In the accompanying, we momentarily talk about the stages. Figure 1 shows The flowchart of the proposed model for diabetes prediction.

3.1. Dataset Description

The Pima Indian Diabetes Dataset has been utilized in this paper. This dataset is open and accessible from the University of California, Irvine UCI AI respiratory (Dataset, P. I. D.). There are 768 records in this dataset with nine attributes, including the outcome attribute. In the final result, there are 768 reports, 268 cases are "tested positive," which shows the patient has diabetes, and 500 cases are "tested negative," implying that the patient has no diabetes.

3.2. Data Preprocessing

The most critical operation is data preprocessing. This operation is important for reliable outcomes and efficient prediction in order to apply ML methods efficiently on the dataset (Soni et al., 2020).

There is no missing value (NAN) value in the Indian Diabetes dataset, but there are some features with a value of zero that are meaningless here.

For this reason, we find the mean and the median of all the Columns that have zero value for the diabetic and non-diabetic patient. Then we replace the value of zero according to the diabetic and non-diabetic patient.

After normalization of Pima Indians Diabetes dataset in the proposed methodology, we used 70 percent of data for validation and training and 30 percent of data for the testing. The model is developed using the Python programming language.

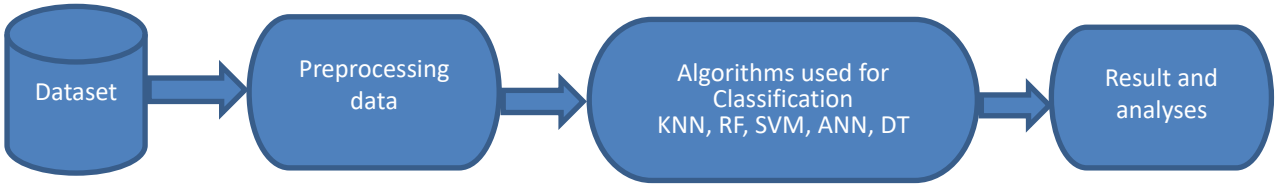

Figure 1. Proposed Model

Table 1. Dataset Description

Number	Attributes	Description
1	Pregnancies	Number of times pregnant
2	Insulin	2-Hour insulin serum (µU/ml)
3	BMI	The index of body mass
4	Age	The Age (years)
5	Glucose	Concentration of plasma glucose for 2 hours in an oral glucose tolerance check
6	Blood Pressure	Blood Pressure Diastolic (mm Hg)
7	Diabetes PedigreeFunction	Diabetes pedigree function
8	Skin Thickness	Skinfold triceps thickness (mm)
9	Outcome	Range of value: 0 and 1(0 means no 1 means yes)

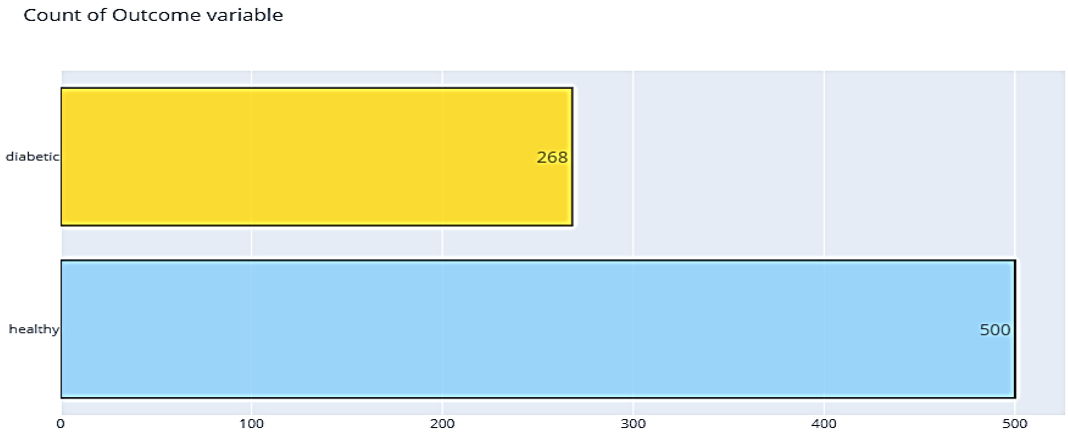

Figure 2. Diabetic and Non-Diabetic Patient

3.3. Algorithms used for Classification

At the point when our dataset has been prepared, we use Machine Learning methods to classify the dataset. KNN, RF, SVM, ANN, and DT classification algorithms have been implemented in this paper with features such as Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree and Age, which are exactly available in the dataset, also with two extra features that we have extracted from dataset by Exploratory Data Analysis technique which a diabetic is described as anyone who has a blood pressure of more than 80 and a glucose level of more than 105. Additionally, anyone with a blood pressure of over 80 is diabetic. With the mentioned features the RF classification algorithm shows the highest accuracy which is (88.31%).

3.3.1. K-Nearest Neighbors (KNN)

KNN is an algorithm for supervised machine learning and is a nonparametric and basic technique that classifies objects in the input space based on the nearest samples. The KNN Classification algorithm attempts to solve both the issues of classification and regression. The KNN algorithm belongs to the group of algorithms that have a slow learning manner. Therefore, the data generalization is delayed until to classification. To specify the class of an element that does not belong to the training set, the KNN classifier searches for k elements in the training set that are nearest to this obscure element (i.e., the shortest distance). KNN is the name given to these k elements. The classes of these k neighbors are verified, and the most common class is assigned to the obscure element's class. (Jardel das et al., 2019).

In this paper, The KNN algorithm is tested with the above-mentioned features of Pima Indians Diabetes dataset. The Confusion matrix for KNN algorithm is shown in Figure 3.

3.3.2. Random Forest (RF)

RF is an easy-to-use ML algorithm that, even without changing its meta parameters, often delivers great results. This algorithm is one of the most commonly used machine learning algorithms for both “Classification” and “Regression” due to its simplicity and usability. This algorithm would randomly create a forest. The built “forest” is actually a “Decision Trees” band. This strategy can undoubtedly deal with huge datasets. Random Forest is created by Leo Bremen. It selects samples randomly from the dataset then builds a decision tree for each sample. A prediction result is measured from each decision tree. Then vote the

prediction result after that the most votes consider the final prediction model (Fawagreh et al., 2014).

The RF algorithm is tested with the above-mentioned features of Pima Indians Diabetes dataset in this study. The Confusion matrix for the Random Forest algorithm is shown in Figure 4.

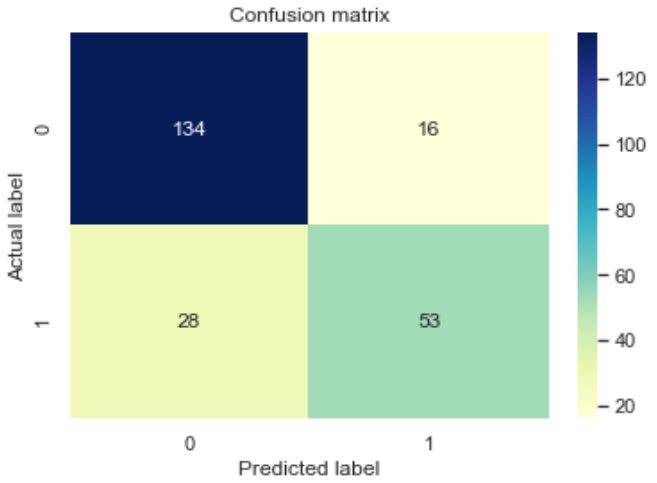

Figure 3. Confusion matrix for KNN algorithm.

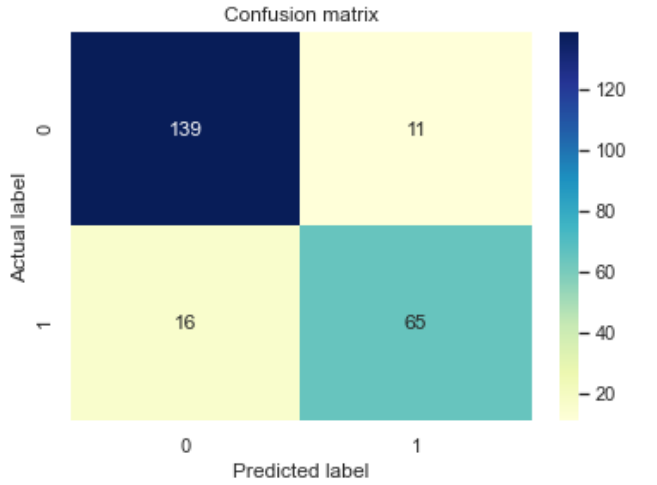

Figure 4. Confusion matrix for Random forests algorithm

3.3.3. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm. In 1963 The SVM was firstly introduced by Vapnik and Chervonenkis.

The SVM tries to locate an ideal hyperplane ready to isolate the examples of any class. This classifier specifies the hyperplane that isolates the spots to put the most noteworthy number of points of a similar class on a similar side while expands the interval of each class to such a hyperplane. The support vectors comprise of the closest points of the hyperplane. The interval from a class to a hyperplane is the littlest interval among them and the spots in that class (Jardel das et al., 2019).

The hyperplane can be utilized for grouping or regression moreover. SVM separates examples in particular groups and can likewise characterize the substances which are not upheld by data. Detachment is finished by through hyperplane plays out the partition to the nearest training spot of any group.

In this paper, the Support Vector Machine algorithm is tested with the above-mentioned features of Pima Indians Diabetes dataset. Figure 5. shows the Confusion matrix for the Support Vector Machine algorithm.

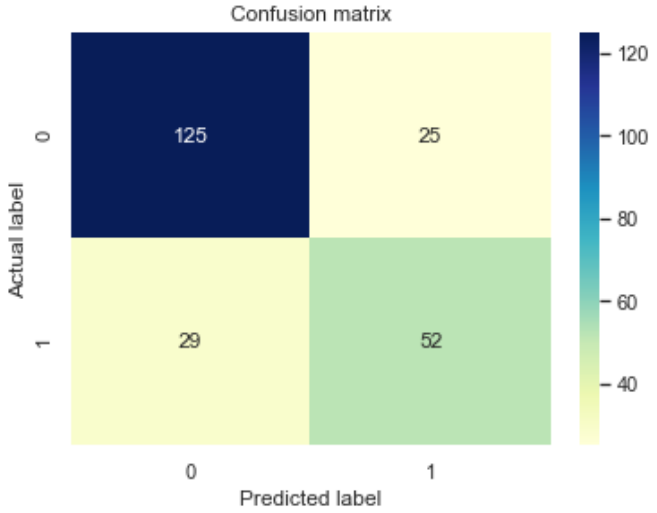

Figure 5. Confusion matrix for SVM algorithm.

3.3.4. Artificial Neural Network (ANN)

The ANN is pretty much the same as the brain's actual neural network. ANN is made up of many interconnected unit operations that cooperate to process data. They often deliver beneficial outputs as a result of it. In general, the artificial neural network (ANN) comprised of network layers and network task, which the network layers namely the input layer, hidden layer and output layer. For the data mining model, the input neurons determine all the input attribute values (Steven W et al., 2003).

Artificial Neural Networks (ANNs) are computational structures modeled on the human brain. A significant number of the new headways have been made in the field of Artificial Intelligence, utilizing Artificial Neural Networks including Voice Recognition, Image Recognition and Robotics.

In this study, the Artificial Neural Networks algorithm is tested with the above-mentioned features of Pima Indians Diabetes dataset. Figure 6 shows the Confusion matrix for the Artificial Neural Networks algorithm.

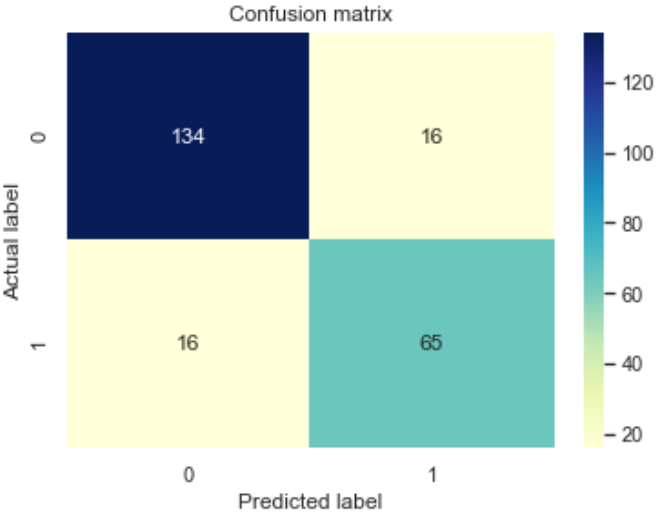

Figure 6. Confusion matrix for ANN algorithm

3.3.5. Decision Tree (DT)

DT is a fundamental classification and regression technique. DT model that has a tree structure can be used to define the mechanism of classifying instances based on characteristics (Quinlan et 1986). When the result attribute is categorical, a decision tree is used.

Both nominal and numerical features are provided by the decision tree algorithm. It has the potential to tolerate noise and unstable values. The decision tree uses a top-down approach to categorize the whole qualified dataset by partitioning the nodes from the topmost to the class node. Every node represents the instance's test attribute, with each node representing one of the several likely values for that feature attribute. From the top node to the attack class node level by level, a decision tree can easily turn the specified set of instances into meaningful patterns.

In this study, the Decision tree algorithm is tested with the above-mentioned features of Pima Indians Diabetes dataset. The Confusion matrix for the DT algorithm is shown in Figure 7.

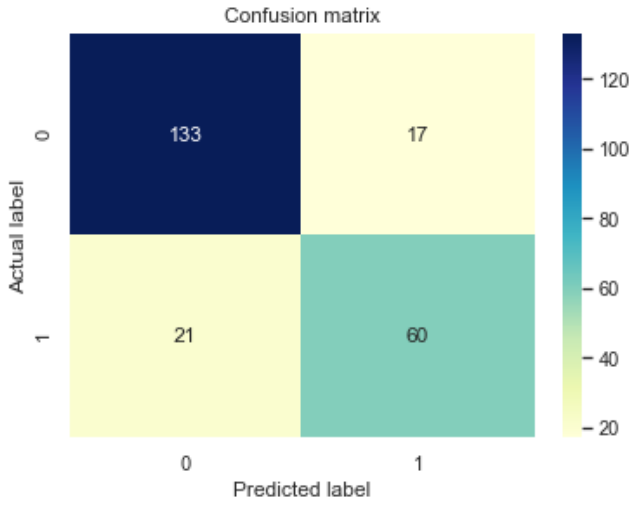

Figure 7. Confusion matrix for Decision tree algorithm.

3.4. Evaluation Measure

The quality of the outcomes produced by various machine learning algorithms is evaluated in terms of accuracy, precision, recall, and F1-score value (Sokolova et al., 2006). We measured accuracy, F1-score, recall, and precision measurements for each classification algorithm in our study by utilizing the confusion matrix.

The confusion matrix in machine learning is a table that is used to display the performance of the algorithm. The performance is determined by testing the input dataset which is given by the user. The below table shows the predicted and the actual values (Yağanoğlu and Köse, 2018).

TP- The forecasted value is positive, and it is right

TN- The forecasted value is negative, and it is right

FP - The forecasted value is positive, and it is wrong

FN- The forecasted value is negative, and it is wrong

Table 2. Confusion Matrix Table

		Actual values	
		Positive (1)	Negative (0)
Forecasted values	Positive (1)	TP	FP
	Negative (0)	FN	TN

$$\textbf{Accuracy} = \frac{TP+FN}{TP+TN+FP+FN} \quad (1)$$

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\textbf{F1-score} = \frac{2*Recall*Precision}{Recall+Precision} \quad (4)$$

4. Results and Discussions

We used various classification techniques to forecast diabetes in this paper. The proposed approach uses different classification algorithms such as KNN, RF, SVM, ANN and DT with features such as Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree and Age, which are exactly available in the dataset, also with two extra features that we have extracted from dataset by Exploratory Data Analysis technique which a diabetic is described as anyone who has a blood pressure of more than 80 and a glucose level of more than 105. Additionally, anyone with a blood pressure of over 80 is diabetic.

Table 3. shows result of the different classification techniques using all existing features as well as two newly extracted features; we can see that the RF classifier works best, with 88.31% accuracy, 88% precision, 86 % recall, and 87% F1-score. With 86 percent accuracy, 85 percent precision, 85 percent recall, and 85 percent F1-score after the RF algorithm, the ANN classifier generates a considerable result.

Table 3. Comparison of the different Classification Techniques with using all available features and two new extracted features

Classification Technique	**Accuracy**	**Precision**	**Recall**	**F1 score**
KNN	81%	80%	77%	78%
RF	**88,31%**	**88%**	**86%**	**87%**
SVM	77%	74%	74%	74%
ANN	**86%**	**85%**	**85%**	**85%**
DT	84%	82%	81%	82%

Table 4 compares different Classification Strategies without using two new extracted features; we can see that the random forest (RF) classifier performs best with 87 percent accuracy, 86 percent precision, 85 percent recall, and 85 percent F1-score. The

DT classifier gives a slightly better result after the RF algorithm, with 83 percent accuracy, 82 percent precision, 81 percent recall, and 81 percent F1-score.

Table 4. Comparison of the different Classification Techniques without using two new extracted features.

Classification Technique	Accuracy	Precision	Recall	F1 score
KNN	82%	80%	79%	80%
RF	**87%**	**86%**	**85%**	**85%**
SVM	77%	75%	75%	75%
ANN	82%	81%	80%	80%
DT	83%	82%	81%	81%

In Table 5, we can see that the random forest (RF) classifier performs best with 88% accuracy, 88% precision, 86% recall, and 87% F1-score as compared to the other Classification Techniques without using Skin Thickness, Pedigree, and two new extracted features. The DT classifier performs well after the RF algorithm, with 84% accuracy, 82% precision, 81% recall, and 81% F1-score.

Table 5. Comparison of the different Classification Techniques without using Skin Thickness, Pedigree, and two new extracted features.

Classification Technique	Accuracy	Precision	Recall	F1 score
KNN	83%	81%	80%	81%
RF	**88%**	**88%**	**86%**	**87%**
SVM	78%	76%	76%	76%
ANN	83%	81%	80%	81%
DT	84%	82%	81%	81%

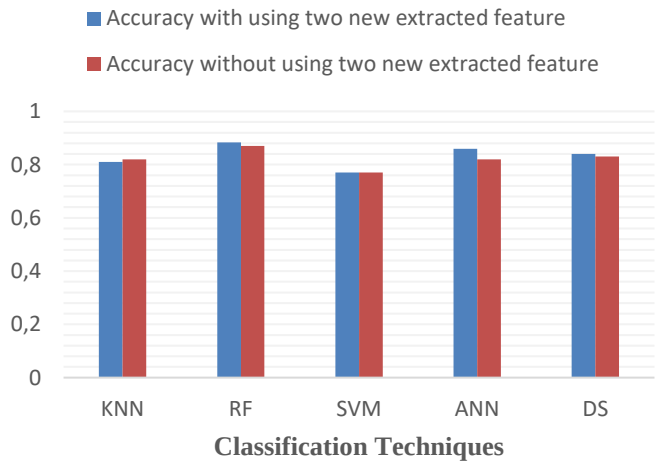

Figure 8. Accuracy Comparison of Classification Techniques using two new extracted features and without using two new extracted features.

According to Figure 8. we can see that using the two new extracted features in classification shows better results than the using predefined features. It's also worth mentioning that as the number of features increases, so does the algorithm's accuracy.

Overall, we implemented the most advanced Machine Learning methods to make predictions and obtain high accuracy.

After analyzing different classification algorithm, the random forest classifier with the mentioned features achieves better compared to others which are 88.31% accuracy after the random forest classifier the ANN classifier shows better result with 86% accuracy. This work was implemented using python. Several additional Python libraries are imported to solve the algorithm much efficiently. We have imported the necessary libraries like pandas, NumPy, scikit-learn and matplotlib.

5. Conclusion

ML which is a subpart of artificial intelligence has the potential to fully change diabetes risk prediction and early identification. Diabetes must be identified early on in order to be handled successfully.

The fundamental goal of this work was to prepare and carry out Diabetes Prediction Using Various Machine Learning Techniques and Conduct Output Analysis of those techniques to find the best classifier with the highest accuracy, that we have achieved successfully.

In this paper, to achieve high-performance accuracy, we extracted two new features from data set and tested various ML classification techniques. RF and ANN algorithms are more effective and produce better results than other ML classification techniques. The classification accuracy of the random forest algorithm was 88.31%.

References

Lonappan, A., Bindu, G., Thomas, V., Jacob, J., Rajasekaran, C., and Mathew, K. T. (2007). Diagnosis of diabetes mellitus using microwaves. J. Electromagnet. Wave. 21, 1393–1401. doi: 10.1163/156939307783239429

Kang, Hyun. (2013). The prevention and handling of the missing data. Korean journal of anesthesiology.

Iancu, I., Mota, M., and Iancu, E. (2008). "Method for the analysing of blood glucose dynamics in diabetes mellitus patients," in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883

Robertson, G., Lehmann, E. D., Sandham, W., and Hamilton, D. (2011). Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. J. Electr. Comput. Eng.2011:681786. doi: 10.1155/2011/681786

Soni. M and Varma. S (2020), Diabetes Prediction using Machine Learning Techniques, International Journal of Engineering Research & Technology (IJERT)

Sarwar. M, Kamal. N, Hamid. W and Shah. A (2018), International Conference on Automation and Computing (ICAC)

Tejas N. Joshi, Prof. Pramila M. Chawan, Diabetes Prediction Using Machine Learning Techniques, January 2018, Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II), pp.-09-13

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

Al Helal, M., Chowdhury, A. I., Islam, A., Ahmed, E., Mahmud, M. S., & Hossain, S. (2019, February). 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.

Dataset, P. I. D. UCI Machine Learning Repository, diambil dari http://archive. ics. uci. edu/ml/datasets. Pima+ Indians+ Diabetes. Accessed (October, 2020)

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal, 2(1), 602-609.

Quinlan, J. R. (1986). Induction on decision tree. Mach. Learn.1, 81–106. doi: 10.1007/BF00116251

Jardel das C. Rodrigues a, Pedro P. Rebouças Filho a, Eugenio Peixoto Jr b, Arun Kumar N c, Victor Hugo C. de Albuquerque b, (2019), Classification of EEG signals to detect alcoholism using machine learning techniques, Pattern Recognition Letters

Sokolova M., Japkowicz N., Szpakowicz S., (2006), Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation, American Association for Artificial Intelligence (www.aaai.org).

Steven W., Narciso C., (2003) Encyclopedia of Physical Science and Technology (Third Edition).

Yağanoğlu, M., & Köse, C., (2018), Real-time detection of important sounds with a wearable vibration based device for hearing-impaired people. Electronics, 7(4), 50.