Diabetes Prediction Using Machine Learning Algorithms

Mustafa Dikici
Control and Automation Engineering
Yildiz Technical University
Istanbul, Turkey
mustafa.dikici@std.yildiz.edu.tr

Melis Sude Altun
Biomedical Engineering
Yildiz Technical University
Istanbul, Turkey
sude.altun@std.yildiz.edu.tr

Eray Mutlu
Biomedical Engineering
Yildiz Technical University
Istanbul, Turkey
eray.mutlu@std.yildiz.edu.tr

***Abstract*—Diabetes Mellitus is a dangerous chronic disease that develops when blood sugar levels remain high for an extended period of time. Diabetics are at a higher risk of ailments such as heart and kidney disease, stroke, and vision issues, among others. Diabetes prediction was chosen for this article because it affects a large number of individuals. This prediction was made using a variety of machine learning methods, including decision trees, Naive Bayes, KNN, SVM, and Ensemble. The data was preprocessed with a range of EDA approaches before being used with these algorithms. The approaches described in this paper can be utilized for pre-diagnosis, saving time and effort for doctors. The widely used PIMA Indian Dataset was used for this article's dataset. The patients have eight different qualities in this dataset.**

***Index Terms*—Diabetes Mellitus, disease, prediction, PIMA Dataset, EDA, Naive Bayes, decision trees, KNN, Ensemble**

I. Introduction

Diabetes Mellitus was chosen for this research because it is a very important disease, and many people suffer from it. People with diabetes have heart disease, kidney disease, stroke, eye problems, nerve damage, etc. high risk of contracting diseases. [1] Diabetes is a serious chronic disease that occurs when sugar levels are high in the blood for a long time. When a person is fed, the body converts the food into glucose and enters the bloodstream. Glucose needs insulin to move through the bloodstream. Insulin is a hormone produced by the pancreas that regulates blood sugar. If the body has a problem using or producing insulin, glucose cannot enter the blood cells. Hyperglycemia occurs when blood sugar levels are higher than normal, and this situation causes serious problems in the nervous system and vessels over time. [2]

The current practice in hospitals is to collect the necessary information for the diagnosis of diabetes with various tests and to provide appropriate treatment based on the diagnosis. Machine learning plays an important role in healthcare industries. And machine learning is used to find hidden patterns, discover information from data, and predict results accordingly. [3]

The aim of this study is to determine whether people have diabetes or not. This problem is a classification problem. Because there are two options to the problem: those who are diseased and those who are not. The method used in this paper is supervised learning. This method is often used for chronic diabetes. The algorithms used in the "Methods" section will be presented in more detail.

II. Dataset

A. Attributes

Pregnancies: Individuals with gastric diabetes are more likely to have type 2 diabetes later in life.

Glucose: A reading of plasma glucose concentrations was taken 2 hours after subjects were given the oral glucose test. Looking at the result, individuals with higher glucose concentrations are more likely to develop diabetes.

Blood Pressure: One of the factors that increase the likelihood of developing diabetes is blood pressure. In addition, if the person's diastolic blood pressure is 70 mmHg, diabetes may occur in these people.

Skin Thickness: The factor that determines skin thickness is collagen content. Skin thickness is higher in insulin-dependent diabetic patients. When the skinfold of the subjects was measured, it was determined that people with a skin thickness of more than 30 mm were at high risk.

Insulin: If we administer glucose for 2 hours, normal insulin levels will be 16-166 mU/L. If it is above or below this value, subjects are at high risk. [4]

Body Mass Index (BMI): Subjects are slightly more likely to have diabetes if their BMI is over 25.

Diabetes Pedigree Function: It shows the history of diabetes in people with blood relatives and provides the combination of the original person and the genetic relationship. If the DPF is high, the person is likely to be diabetic.

Age: Diabetes is a type of disease that can be seen in all age groups. However, it is most common in adults over the age of 45. Considering all these factors, it can be said that individuals in the higher age group are more likely to have diabetes.

For machine learning training, the dataset should be divided into two parts. One part should be the training part and the other part should be the testing part. The model is always adapted to the training set first. In the test part, performance measurement is required.

TABLE I
PIMA DATASET ATTRIBUTES[2]

Attribute	Description of Attributes
Pregnancies	Number of times pregnant
Glucose	Plasma concentration in oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thikness	Triceps skinfold thickness (mm)
Insulin	2hr. serum insulin (mu U/ml)
BMI	Body mass index (weight in kg / (height in m)^2)
DPF	Diabetes predigree function
Age	Age (Years)
Outcome	Class variable (0 or 1)

III. LITERATURE REVIEW

The research in [5], the publicly available dataset named as Pima Indians Diabetes Database is used for performing their experiment. Their prediction framework begins with dataset selection and then moves on to data pre-processing. After preprocessing the data, three classification algorithms: naive Bayes, SVM, and Decision tree are used. As they incorporated different evaluation metrics, they did compare the different performance measure and comparatively analyzed the accuracy. The highest accuracy achieved with their experiment was 0.76.

In another research [6], different dataset is used which have 10 attributes (Job Type extra). Their prediction framework begins with dataset selection and then moves on to data pre-processing. After preprocessing the data, various machine learning algorithms are used such as Random Forest, Ada Boost , Logistic Regression, KNN, Naïve Bayes, Bagging etc. Then both PIMA Dataset and their dataset are compared on same classifiers. For PIMA Dataset, Ada Boost has reached to 0.77 accuracy, while 0.96 accuracy is reached for their dataset using Logistic Regression.

Finally, Alam et al. [7] employed artificial neural network (ANN), Random Forest (RF), and K-means clustering to predict whether or not a patient has diabetes. The dataset contains 768 records of female patients after data processing, cleaning, and transformation, with 500 negative values and 268 positive values. The results showed that the ANN outperformed the two other algorithms when using association rule mining, with a 0.75 accuracy.

IV. EXPLORATORY DATA ANALYSIS (EDA)

A. Distribution of Dataset

There are 768 patients in this dataset and 268 of the have diabetes mellitus while 500 of them have not. As seen diabetes data is less than healthy data, it may cause a problem for model accuracy as the model learns from a larger dataset of healthy people.

B. Missing Values

Dataset is checked and there is not found any NaN value. But data has abnormal zeros which can be result of measurement error. These features are Glucose, Blood Pressure, Skin Thickness, Insulin and BMI.

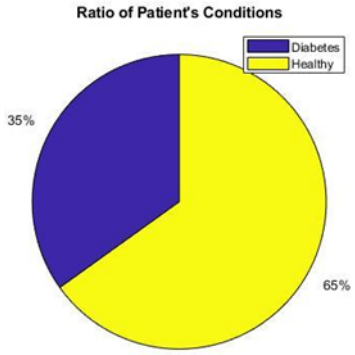

Fig. 1. Ratio of Patient's Conditions

TABLE II
MISSING VALUES

Attribute	Zero Values of Attributes
Pregnancies	111
Glucose	5
Blood Pressure	35
Skin Thikness	227
Insulin	374
BMI	11
DPF	0
Age	0

C. Replacement of Missing Values

Meaningless values are replaced by the mean of every feature itself. These values are meaningless as being zero is impossible in biological perspective for selected features. And these abnormal values decrease the accuracy of the model. In addition, mean extraction is performed after the separation of patients which have diabetes or have not.

D. Descriptive Statistics for Diabetes and Healthy

In this section, meaningless values are replaced by the mean of the every feature and then descriptive statistics are calculated. Table is constructed by the result of the MATLAB code which is send in attachment.

Standart Deviation: Standard deviation is a criterion that gives the spread of data values in probability theory and statistics.

Mean: Indicates that there is at least one point in the mean derivative of its value.

Median: It is used as a criterion indicating the central position in descriptive statistics.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
Mean	4.865	142.311	74.901	29.183	128.285	35.123	0.526	37.067
Standard Deviation	374.124	294.883	117.328	8.90553	479.841	5.99954	0.312213	10.9683
Median	4	140.5	74	27	100.3358	34.3	0.449	36
Minimum	0	78	30	7	14	22,9	0.088	21
Maximum	17	199	110	56	250	53.2	1.394	70
25th Percentile	2	119	68	22.16	100.33	30,9	0.26	28
50th Percentile	4	140.5	74	27	100.33	34.3	0.44	36
75th Percentile	8	167	82	35	167	38.7	0.72	44
# of Outliers	0	0	1	1	21	3	4	0

Fig. 2. Descriptive Statistics

E. Histograms of Features

Every features are shown in histograms after zero values are replaced by means.

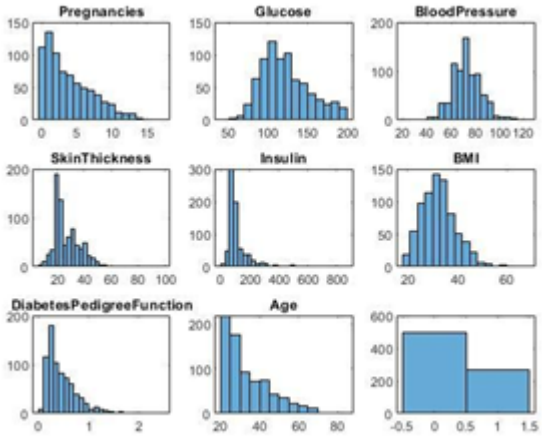

Fig. 3. Histograms of Features

F. The Correlation Matrix

The correlation matrix indicates relationships between features. Due to these relations, adding extra features is determined by multiplying or dividing each other.
These features are selected in this concept:
Feature 0: Age - Pregnancies (Since pregnancies may be zero, the feature will be illogical especially for males and the effect of age would be diminished abruptly.)
Feature 1: Skin Thickness - BMI
Feature 2: Insulin – Glucose

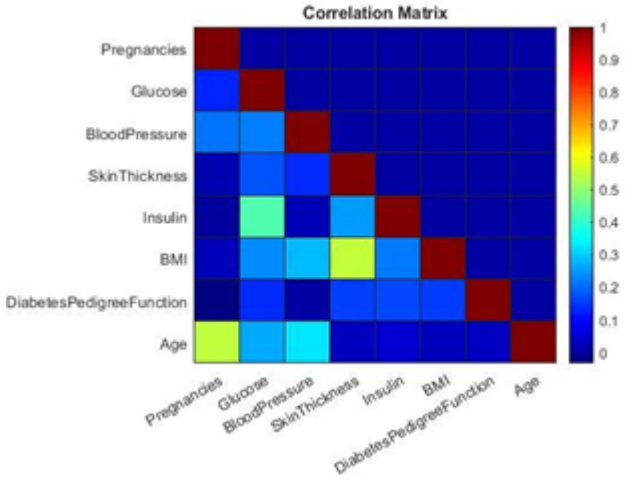

Fig. 4. Correlation Matrix

G. Normalization

Normalization by dividing data to its maximum values is done. It will improve the accuracy since dealing in the range of [0,1] is easier than multiple ranges.

H. Make Extra Features

As explained before, the relation between BMI-SkinThickness and Glucose-Insulin is found from the correlation matrix. Due to this, extra features are created, normalized and added to the data.

V. METHODS

While doing this project, the main purpose is to predict the diabetes of the subjects using different machine learning methods. In this way output reports and analysis of techniques were used. The main goal is to find the classifier that works with the highest accuracy.

A. Dataset Description

PIMA Indian Diabetes Dataset was used while preparing the project. This dataset has a total of 768 records with 8 features. In the dataset, 1s were coded as 0s for diabetics. There are diabetic patients in total (number of patients).

B. Data Preprocessing

It is necessary to process things that are applied step by step in the data processing process. The most important process here is to start from the ground up and move forward without skipping anything. Matlab was used to write this data.

C. Prediction Algorithms

The goal of this research is to compare three types of machine learning algorithms: probabilistic, probabilistic, and probabilistic.

Algorithms that are vector-based, decision-based, or statistical. As a result, Naive Bayes, Support Vector Machines, and Support Vector Machines with kernels, both linear and nonlinear, Random, Decision Tree K-nearest neighbor, Logistic Regression, and Forest For the analysis, an Adaptive Boosting Classifier was used comparison.

1) Naive Bayes Theorem: In probability and statistics, the Bayes theorem is used to create a machine learning algorithm. It is known as naive because it assumes feature independence. This approach is commonly used in supervised learning classification that is linear. It's been shown to be effective in text processing and document classification in the past. The probabilistic character of this algorithm is well-known. P(A—B)=(P(B—A)*P(A))/P(B)

2) SVMs (Support Vector Machines): Linear and nonlinear kernels, It can be used to classify data in both linear and nonlinear ways. SVM has already been shown to work well in high-dimensional spaces, where the number of dimensions is dictated by the number of features/characteristics on which the prediction is to be made. The support vector machine distinguishes the data by creating a hyperplane with respect to distinct dimensions. The forecast is usually more accurate the further the data point is from the hyperplane.

3) K-Nearest Neighbor (KNN): This popular machine learning technique uses majority polling to forecast outcomes. In comparison to other machine learning algorithms, KNN is comparatively simple. The KNN method polls the closest point to the point to be predicted and classifies it based on the majority class in the neighboring points. One of the major drawbacks of this approach is the categorization of outliers and the difference in classification when the value of K is changed.

4) Decision Tree: Decision trees are a common machine learning method that makes predictions by asking a series of questions. Depending on the answers to each question, the algorithm decides which path/branch to take next until it reaches a conclusion. One of the most significant advantages of employing Decision Trees is their low bias, or, to put it another way, their high flexibility in the correlations that they can learn. They can, however, suffer from high variance and, as a result, overfitting.

$$H(X) = -\sum_{k=1}^{K} p_k \log_2(p_k) \qquad Gini = 1 - \sum_{k=1}^{K} p_k^2$$

Fig. 5. Gini Formula

5) Adaptive Boosting Classifier (ADA Boost): A statistical classification method that adjusts its weights by combining weak rules, with the more erroneous predictions given larger weights. As a boosting classifier, it works in tandem with other algorithms that have performed poorly in their prediction tasks and is used to improve accuracy.

VI. RESULTS AND DISCUSSION

The results of all the implemented classifiers will be reported in this section. . In order to tune the hyperparameters of the classifiers, 'OptimizeHyperparameters' input were used in the model function. After training the previously discussed algorithms on the dataset with a ratio of training and testing of 0.8 and 0.2 respectively the following results were obtained for each algorithm. The F-scores for each algorithm employed are listed in the crosstab Table III.

TABLE III
RESULTS

Algorithms	Precision	Recall	Accuracy	F1-score
SVM	0.67	0.77	0.71	0.66
Decision Tree	0.74	0.8	0.77	0.74
Naïve Bayes	0.94	0.95	0.94	0.94
KNN	0.99	0.99	0.99	0.99
Ensemble	0.96	0.95	0.96	0.95

Table III shows that KNN method has the highest accuracy and f1-scores which are significantly closer to maximum 0.99. Also, Naïve Bayes and ensemble algorithm resulted with 0.94 and 0.95 accuracy and f1-score respectively which is very satisfying. To see true negative or false positives, confusion matrices are visualized for all algorithms in figure 5.

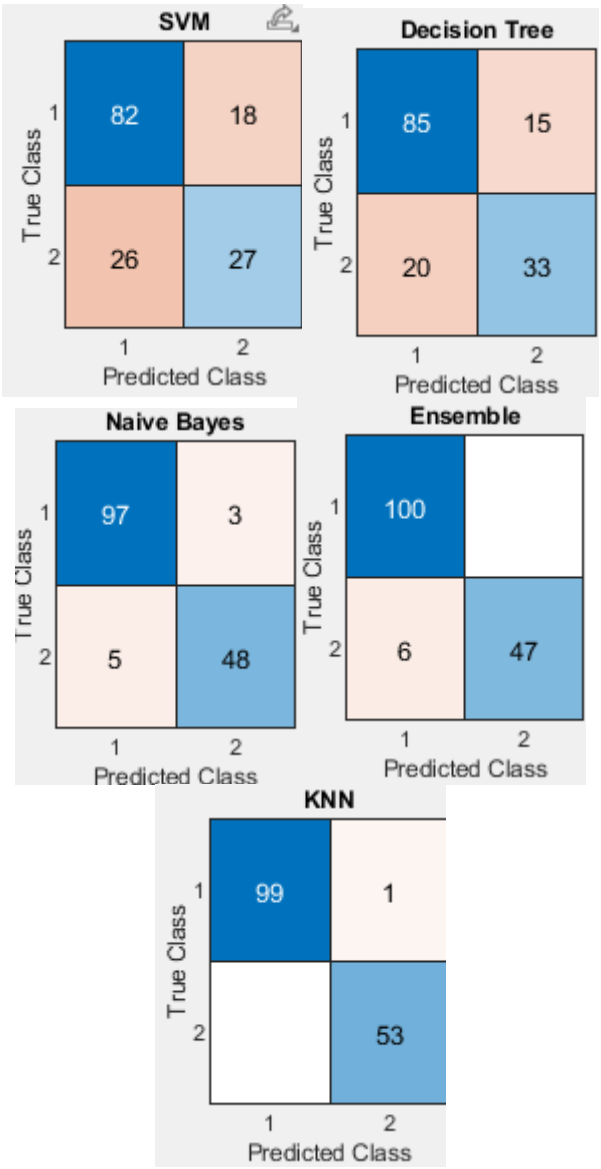

Fig. 6. Confusion Matrices

False negatives are important values to minimize for disease prediction models since patients who have diabetes are labeled healthy incorrectly. As indicated top, knn,ensemble and naïve bayes algorithms produced 0, 6 and 5 false negatives respectively. Therefore these model are desired to use for diabetes prediction.

Optimized hyperparameters are shown below in tables for each model.

These parameters are produced by the 'OptimizeHyperparameters' input in model functions, there are also model parameters which could be seen in the code.

VII. CONCLUSION

Five machine learning algorithms were used to predict diabetes. The study was conducted on K-nearest neighbor, Naive Bayes, Support Vector Machine, Decision tree and

ADA boost. A total of 768 participants were included in the investigation. Where 80 percent of the data set was partitioned for training and 20 percent was partitioned for testing. Having used optimized hyperparameters a maximum accuracy of 0.99 was achieved by the KNN among the other algorithms.

In EDA part, outlier removal, replacing zeros with mean and min-max scaling are used to get better accuracy but, when these methods are applied in the dataset accuracy drops to around 0.65 in optimal. So, outlier removal and replacing zeros with mean decrease the efficiency of learning process. Therefore, data is just normalized as indicated before.

REFERENCES

[1] Karamanou, M., Protogerou, A., Tsoucalas, G., Androutsos, G., Poulakou-Rebelakou, E. (2016). Milestones in the history of diabetes mellitus: The main contributors. World journal of diabetes, 7(1), 1–7. https://doi.org/10.4239/wjd.v7.i1.1.

[2] Villegas-Valverde CC, Kokuina E, Breff-Fonseca MC. Strengthening National Health Priorities for Diabetes Prevention and Management. MEDICC Rev. 2018 Oct;20(4):5.

[3] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x.

[4] Umpierrez GE, Isaacs SD, Bazargan N, You X, Thaler LM, Kitabchi AE. Hyperglycemia: An independent marker of in-hospital mortality in patients with undiagnosed diabetes. Journal of Clinical Endocrinology Metabolism. 2002;87:978–982.

[5] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578 – 1585, 2018, international Conference on Computational Intelligence and Data Science. [Online].

[6] Mujumdar, A., Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science.

[7] T. M. Alam, M. A. Iqbal, Y. Ali, W.Wahab, S.Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, and Z. Abbas, "A model for early prediction of diabetes". Informatics in medicine unlocked Volume 16, 2019, 100204.

TABLE IV
SVM

Box Constraint	Kernel Scale
922.31	3.8662

TABLE V
DECISION TREE

Min Leaf Size
1

TABLE VI
NAIVE BAYES

Distribution Names	Width
kernel	0.00089506

TABLE VII
KNN

Num Neighbors	Distance
6	jaccard

TABLE VIII
ENSEMBLE

Method	Num Learning Cycles	Learn Rate	Min Leaf Size
AdaBoostM1	400	0.95152	5