PROGNOSIS OF THE DIABETES TYPE II USING MACHINE LEARNING TECHNOLOGY

by

Amanzhol Shungeyev

Submitted to the School of Medicine

in partial fulfillment of the requirements for the degree of

Master of Sport Medicine and Rehabilitation

at the

NAZARBAYEV UNIVERSITY

April 2022

Author . . . . . .

School of Medicine

Apr 20, 2022

Certified by . . . . . .

Syed Ali

Associate Professor of Biomedical Science Department

Thesis Supervisor

Accepted by . . . . . .

Prashant Kumar Jamwal

Associate Professor of Electrical and Computer Engineering Department

TABLE OF CONTENT

ABSTRACT

The annual statistics of the incidence of diabetes mellitus shows a stable growth. Most diabetics are not aware of their diagnosis and the risks that this disease carries. Also, an important problem is playing sports in diabetes, it worries both people with newly diagnosed diabetes and those who live with this disease for a long time. Thus, early diagnosis and treatment are critical to prevent morbidity. The main goal of the study is to assess the risk of diabetes among people depending on their lifestyle and marital status. This study proposes several data mining models for predicting type 2 diabetes mellitus (T2DM). Diabetes prediction models are very accurate with rates above 95%, which is why their application is very relevant to healthcare applications. Two sets of data were used to conduct machine learning, one is the main one, the second was used for validation.

1. BACKGROUND

Type 2 diabetes mellitus (T2DM) is a chronic violation of carbohydrate metabolism that is characterized by a disruption in the interaction of the hormone insulin, produced by pancreatic β-cells, with tissue cells, because of which insulin cannot penetrate into the cells, but accumulates in the blood. In type 2 diabetes, insulin is produced in a normal amount, and often in more quantities than necessary. But at the same time, due to a violation of its effect on cells or a violation of the susceptibility of tissue cells to insulin (insulin resistance), the cells starve without receiving insulin, and insulin itself accumulates in the blood [1].

According to the global diabetes community, a healthy person's fasting blood glucose is below 5.5 mmol/L. The blood glucose range for people with prediabetes and diabetes is 5.5 - 6.9 mmol/L and above 7.0 mmol/L respectively [2].

Exercising with diabetes is an issue that worries both people with newly diagnosed diabetes and those who have been living with this disease for a long time. Physical activity plays a special role in the lives of diabetics and everyone who strives to lead a healthy lifestyle. If you have type 1 diabetes, exercise will help increase your insulin sensitivity, which means you can reduce your insulin dosage by the amount of carbohydrates you consume. At the same time, exercise can reduce the risk of developing type 2 diabetes [3].

As a preliminary step before choosing a training program, diabetics should be screened for complications associated with the disease, and based on the data obtained, appropriate adjustments should be made to the exercise program. When insulin or oral insulin secretaries are used by athletes, it may contribute to immediate or delayed exercise-induced hypoglycemia. Athletes diagnosed with diabetes are recommended to combine regular aerobic exercise with resistance exercise. Insulin-dependent athletes should closely monitor their blood sugar levels before, during, and after exercise. It is important to note that it is

categorically not recommended to make physical exercise when significant hyperglycemia is detected, since physical activity can paradoxically aggravate hyperglycemia and lead to ketoacidosis. Awareness by athletes of the symptoms and actions to be taken (e.g., use of fast-absorbing glucose) is essential. Thus, early diagnosis of diabetes in athletes is extremely important [4].

For timely diagnosing and even predicting the risk of developing diabetes, it is necessary to use advanced information technologies, namely data mining technology. Data mining is the process of discovering hidden patterns or relationships between variables in large arrays of processed or raw data. It is subdivided into tasks of classification, modeling and forecasting, and others using methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [5].

The incidence of diabetes is growing every year. The world statistics of diabetes in 2019 was estimated at 9.3% among adults (20–79 years), which is almost half a billion people (463 million). According to researchers' forecasts, this figure will increase to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. Every second, namely 50.1% of people living with diabetes, do not know about this diagnosis [6].

2. RELATED WORKS

There are many different studies that are engaged in predicting the occurrence of diabetes using different methods: machine learning, data mining, hybrid methods, and a small part uses a neural network and a genetic algorithm. In the course of a preliminary study it was found that there is a pattern that the vast majority of studies are based on the Pima Indian Diabetes Dataset (PIDD) from the University of California, Irvine (UCI) Machine Learning Database [7].

Han W. et al. proposed a model consists of double-level algorithms. In the first level the improved K-means algorithm was used to remove incorrectly clustered data. Then, they applied the logistic regression algorithm to classify the remaining data. This approach allowed to achieve a predictive accuracy of up to 95.42% [8]. Roshan B. compared the prediction accuracy of three types of algorithms: Gradient Boosting, Logistic Regression and Naive Bayes. The results showed that the Gradient Boosting algorithm perfomed with a highest accuracy (86%) [9]. Boshra F. et al.designed a system for diabetes prediction based on the six types of classifiers: logistic regression, decision tree, adaboost, support vector machine (SVM), xgboost and random forest (RF). The adaboost algorithm demonstarted the best result 83.76% [10]. The list of studies that use Pima Indian Dataset also includes:

- In the course of work, Sisodia D. et al. used Decision Tree (DT), support vector machine (SVM) and Naive Bayes (NB) algorithms. However, the most accurate result of 76.3% was obtained using the Naive Bayes algorithm [11].
- Ram D. et al. created the models utilizing a logistic regression and decision tree algorithms. Model based on the logistic regression has the highest prediction accuracy of 78.26% [12].

- Sidong W. et al. used the Logistic Regression, Deep Neural Network (DNN), Support Vector Machine, Decision Tree and Naive Bayes algorithms for developing T2DM prediction model. The highest accuraccy (77.86%) was achived by DNN based model [13].
- Aiswarya I. et al. in their pridictive model applied the decision tree algorithm and Naïve Bayes algorithms. The most accurate model was based on the Naïve Bayes with result 79.57% [14].
- Quan Z. et al used the decision tree, random forest algorithms and neural network. The highest achieved the accuracy is 76.67% with neural network, however the results of random forest based model approximately same (76.04%) [15].
- Vijayan V. et al. used the Decision Tree, Support Vector Machine, Naive Bayes, Decision Stump and AdaBoost Algorithm for predictive model creation. The highest accuracy (80.72%) was shown by AdaBoost algorithm [16].

3. PROBLEM STATEMENT

3.1. RESEARCH QUESTION

Are BMI, regular medicine, blood pressure and stress level the most important parameters for T2DM risk prediction?

3.2. HYPOTHESIS

BMI, regular medicine, blood pressure and stress level are the most important parameters for T2DM risk prediction.

3.3. OBJECTIVE

The main goal of this study is to develop a machine learning model for predicting the risk of type 2 diabetes with an accuracy above 95%.

3.4. SPECIFIC AIMS

1) Approval from the NU Institutional Ethics Committee
2) Dataset search:
    a) Searching a dataset with clinical indicators of patients and the presence of a diagnosis of type 2 diabetes
    b) General initial assessment of the found dataset
3) Dataset preprocessing:
    a) Statistical analysis, distribution checking, patterns identification and assumptions testing
    b) Rejection of the outliers, restoring the null values and make normalization of data frame
4) Final model and validation
    a) Building the model and identification of best algorithm and feature importance
    b) Validation by Pima Indian Dataset

4. EXPERIMENTAL PLAN

4.1. THE DATASET DESCRIPTION

This study is based on Diabetes Dataset 2019 (DD2019) from the Department of Computer Science and Engineering, Birla Institute of Technology - Mesra, collected by Neha Prerna Tigga and Dr. Shruti Garg [17] The dataset includes 952 participants aged 18 years and older (580 men and 372 women) and consists of 18 different parameters: age, gender, family diabetes, high BP, physical activity, BMI, smoking, alcohol, sleep, sound sleep, regular medicine, junk food, stress, BP level, pregnancies, Pdiabetes (gestation diabetes), urination frequency and a parameter that identifies the presence of diabetes diagnosis on patient (class 1) or not (class 0) (Table 1-1.1.). The validity verification process of the developed models utilizes the Pima Indian Dataset on which the testing of proposed algorithms was made.

Pima Indians Diabetes Dataset (PIDD) is a publicly available dataset from the University of California, Irvine (UCI) [7]. It consists of the medical analysis of 768 women at least 21 years old and contains 9 metrics: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function (a function which scores likelihood of diabetes based on family history), age and a parameter that determines the presence of a diagnosis of diabetes in a patient (class 1 – diabetic and class 0 – non-diabetic). Tables 2 and 2.1. describes the details of each feature (type and description).

Table 1. The DD2019 description.

Parameters			Options
Total			952 samples and 17 parameters
1	Age	Age in years	Less than 40
			40-49
			50-59
			60 or older
2	Gender	Gender	Male/ Female
3	Family diabetes	Family history with diabetes	Yes/ No
4	High BP	Diagnosed with high blood pressure	Yes/ No
5	Phisical activity	The amount of time spent in physical activity	None
			Less than half an hr
			More than half an hr
			One hr or more
6	BMI	Body Mass Index (kg/m^2)	Numeric
7	Smoking	Smoking	Yes/ No
8	Alcohol	Alcohol consumption	Yes/ No
9	Sleep	Sleep duration in hours	Numeric
10	Sound sleep	Sound sleep duration in hours	Numeric
11	Regular medicine	Regular medicine intake	Yes/ No
12	Junk food	Junk food consumption	Occasionally
			Often
			Very often
			Always
13	Stress	Stress level	Not at all
			Sometimes
			Very often

			Always
14	BP level	Blood pressure level	High/ Normal/ Low
15	Pregnancies	Number of pregnancies	Numeric
16	Pdiabetes	Pregnancy/Gestinational diabetes	Yes/ No
17	Urination frequency	Urination frequency	Not much/ Quite often
18	Diabetic	The presence of diabetes diagnosis on patient	Yes - 266
			No - 685

Table 1.1. The overview of numerical attributes of the DD2019.

	Attribute	Mean ± std	Range
1	BMI	25.763713 ± 5.402595	15 – 45
2	Sleep	6.949580 ± 1.273189	4 – 11
3	Sound Sleep	5.495798 ± 1.865618	0 – 11
4	Pregnancies	0.386813 ± 0.909455	0 – 4

Table 2. PIDD description.

Parameters			Options
Total			768 samples and 8 parameters
1	Age	Age in years	Numeric
2	Pregnancies	Number of pregnancies	Numeric
3	Glucose	Concentration of plasma glucose (mg/dL)	Numeric
4	Blood Pressure	Diastolic blood pressure (mm Hg)	Numeric
5	Skin Thickness	Triceps skin fold thickness (mm)	Numeric
6	Insulin	2-hour serum insulin (mu U/ml)	Numeric
7	BMI	Body Mass Index (kg/m^2)	Numeric

8	Diabetes Pedigree	A pedigree function for diabetes	Numeric
9	Outcome	The presence of diabetes diagnosis on patient	Yes – 268 (class 1)
			No – 500 (class 0)

Table 2.1. The overview of the PIDD.

	Attribute	Mean ± std	Range
1	Age	33.240885 ± 11.760232	21 – 81
2	Pregnancies	3.845052 ± 3.369578	0 – 17
3	Glucose	120.894531 ± 31.972618	0 – 199
4	Blood Pressure	69.105469 ± 19.355807	0 – 122
5	Skin Thickness	20.536458 ± 15.952218	0 – 99
6	Insulin	79.799479 ± 115.244002	0 – 846
7	BMI	31.992578 ± 7.884160	0 – 67.1
8	Diabetes Pedigree	0.471876 ± 0.331329	0.078 – 2.42

4.2. PROPOSED FRAMEWORK

This section consists of a description of the data preprocessing procedure and classification algorithms. The figure 1 illustrates the proposed framework.

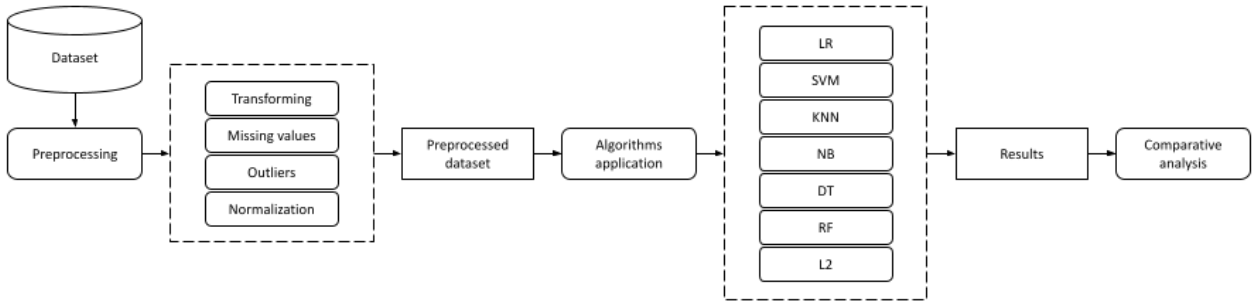


Figure 1. The block diagram of the proposed framework

4.2.1. DATA PREPROCESSING

The predictive result of the model in significant degree depends on the quality of the data used. For this reason, data preprocessing is a key factor in model creation [8]. In this study, several suitable methods were selected to optimize the original dataset:

- Conversion of categorical features to a numerical form
- Missing or null value rejection/imputation
- Outlier rejection
- Data normalization

4.2.1.1. CONVERSION OF CATEGORICAL FEATURES TO A NUMERICAL

This study uses machine learning algorithms that cannot handle categorical variables. For this reason, there is a need to convert the column of labels or categorical columns into separate columns of 0s and 1s. This process is called getting dummies pandas columns [18]. This method was used only on DD2019, as it contains categorical attributes (Table 1).

4.2.1.2. MISSING VALUE REJECTION OR IMPUTATION

The problem of missing values is common to most datasets and their elimination is an important step in preprocessing, as it can significantly affect the final result of the classification model. For these purposes, a high-quality and reliable method for handling missing values is needed. In this study, two options depending on the dataset were used: iterative imputation of missing values and complete elimination of empty values.

Iterative imputation is a process in which each feature is modeled as a function of other features, i.e. a regression problem in which missing values are predicted. The imputation process is carried out sequentially, thereby providing the ability to use previously imputed values to predict subsequent attributes. Based on the name of the method, it can be

understood that this process is repeated multiple times. This feature of the method makes it possible to calculate more and more accurate values of missing data with each new iteration. [19]

4.2.1.3. OUTLIER REJECTION

An outlier is a noticeable deviation of a value from other values. The need to exclude it from the dataset is due to the high sensitivity of classifiers to the distribution of attributes and the range of data. The outlier definition formula is as follows [20]:

$$P(x) = \begin{cases} x, & \text{if } Q_1 - 1.5 * IQR \leq x \leq Q_3 + 1.5 * IQR \\ \text{reject}, & \text{otherwise} \end{cases} \quad (1)$$

where x are instances of the feature vector lying in the n-dimensional space, $x \in R^n$. Q1, Q3 and IQR are the first quartile, third quartile and interquartile range of attributes, respectively, where Q1, Q3, IQR $\in R^n$.

4.2.1.4. DATA NORMALIZATION

In this study, the normalization process brings the data to a certain range (from 0 to 1). Normalization is applied when the ranges of various features have significant differences, it helps to reduce data redundancy and improve integrity of the data. The normalization definition formula is as follows [21]:

$$x' = \frac{x - \min}{\max - \min} \quad (2)$$

where x' is value after the normalization, x is value before normalization, min and max is the maximum and minimum values in the feature range respectively.

4.2.2. BRIEF DESCRIPTION OF ALGORITHMS USED

4.2.2.1. LOGISTIC REGRESSION METHOD

Logistic regression is a type of multiple regression general purpose of which is to analyze the relationship between multiple independent variables (also called regressors or predictors) and a dependent variable. Binary logistic regression is used when the dependent variable is binary (that is, it can only take two values). Logistic regression can be used to estimate the probability that an event will occur for a particular subject (diabetic/non-diabetic) [22].

In multiple linear regression, the dependent variable is assumed to be a linear function of the independent variables, i.e.:

$$y(w, x) = w_0 + w_1 * x_1 + \cdots + w_n * x_n \quad (3)$$

where y is the estimated continuous outcome; $w_0 + w_1 * x_1 + \cdots + w_n * x_n$ is the linear regression equation for the independent variables in the model; w_0 is the intercept (constant value), or the point at which the regression line touches the vertical Y axis.

4.2.2.2. SUPPORT VECTOR MACHINE

Support Vector Machines are a family of supervised binary classification algorithms that use a linear division of the feature space using a hyperplane. The main idea of the method is to map feature space vectors representing the objects being classified into a space of higher dimension. This is because in a space of higher dimension, the linear separability of a set turns out to be higher than in a space of lower dimension. The reasons for this are intuitive: the more features are used to recognize objects, the higher the expected quality of recognition. After transferring to a space of higher dimension, a separating hyperplane is built in it. In this case, all vectors located on one “side” of the hyperplane belong to one class, and those located on the other, to the second [23].

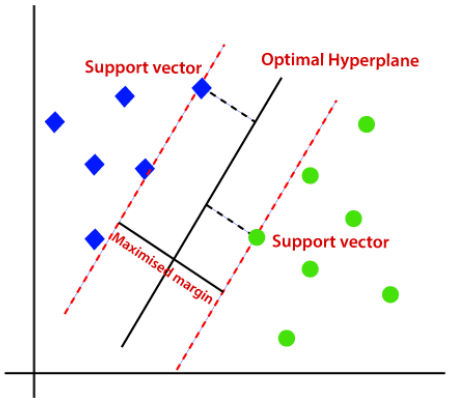


Figure 2. Support Vector Machine.

4.2.2.3. K-NEAREST NEIGHBOR CLASSIFIER

The k-nearest neighbor method is used to solve the classification problem. It assigns objects to the class that has the majority of its k nearest neighbors in the multidimensional feature space. This is one of the simplest algorithms for training classification models. The number k is the number of neighboring objects in the feature space that are compared with the object being classified. In other words, if k = 10, then each object is compared with 10 neighbors. During the learning process, the algorithm simply remembers all feature vectors and their corresponding class labels. When working with real data, i.e. observations whose class labels are unknown, the distance between the vector of the new observation and the previously stored ones is calculated. Then the k vectors closest to it are selected, and the new object belongs to the class that most of them belong to [24].

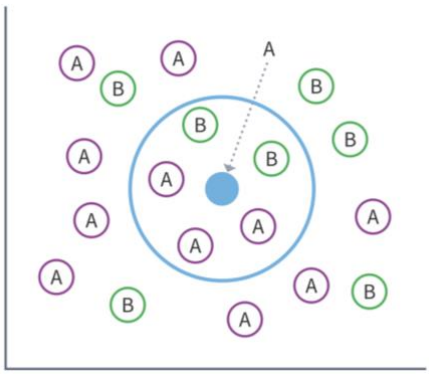


Figure 3. K-Nearest Neighbor Classifier.

4.2.2.4. NAIVE BAYES CLASSIFICATION

Like any classifier, Bayesian assigns class labels to observations represented by feature vectors. It is assumed that each feature independently affects the probability that an observation belongs to a class. For example, an object can be considered an apple if it is round, red, and about 10 cm in diameter. A Naive Bayes classifier "thinks" that each of these features independently contributes to the likelihood that the object is an apple, regardless of any possible correlations. between the characteristics of color, shape and size. An additional advantage of the method is the small number of examples needed for training [25]. The NB classifier is a probabilistic model that can be written as:

$$p(c|x) = \frac{p(c) * p(x|c)}{p(x)} \quad (4)$$

where $p(c|x)$ – posterior probability, $p(c)$ – class prior probability, $p(x|c)$ – likelihood, $p(x)$ – predictor prior probability.

4.2.2.5. DECISION TREE CLASSIFIER

A decision tree is a classifier built on the basis of “if, then” decision rules arranged in a tree-like hierarchical structure. The decision tree is based on the process of recursively splitting the initial set of objects into subsets associated with predefined classes. Partitioning is performed using decision rules, in which attribute values are checked according to a given condition. Structurally, a decision tree consists of two types of objects - nodes and leaf. The nodes contain decision rules and subsets of observations that satisfy them. The leaves contain observations classified by the tree: each leaf is associated with one of the classes, and the object that is distributed into the leaf is assigned the corresponding class label [26].

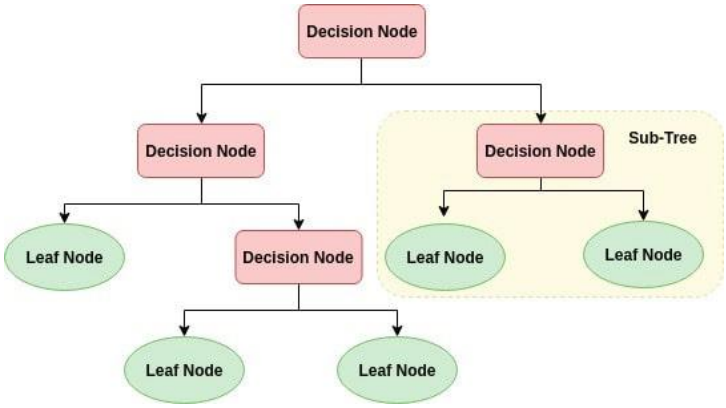


Figure 4. Decision Tree.

4.2.2.6. RANDOM FOREST CLASSIFICATION

Random forest is a set of decision trees. In the regression problem, their answers are averaged; in the classification problem, the decision is made by majority voting. All trees are built independently of each other [27].

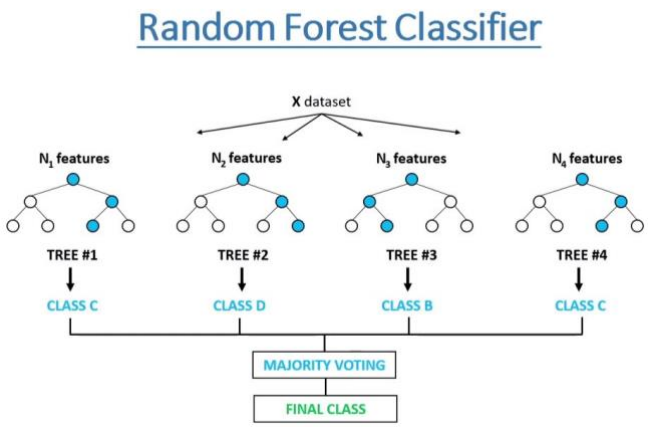


Figure 5. Random Forest.

4.2.2.7. L2 REGULARIZATION

L2 regularization, or ridge regression, for integral equations allows you to balance between data compliance and a small solution norm. Thus, the function underestimates the peaks by adding the sum of the weights squared with the lambda factor. The term λ is called the regularization parameter. It balances the cross-entropy error function and the regularization

penalty. If the value of λ is large, the weights will tend to zero, if the value of λ is small or equal to zero, then the weights will simply tend to minimize the cross-entropy error function. As a rule, the value of the parameter λ is set to 0.1 or 1, or in the region between these values, but in general its value depends on the specific data [28].

$$L(x,y) = \sum_{i=1}^{n} \left(y_i - h_\theta(x_i)\right)^2 + \lambda \sum_{i=1}^{n} \theta_i^2 \quad (5)$$

where n is the dimension of y_i and x_i. λ is the penalty term or regularization parameter.

4.2.3. THE CROSS VALIDATION

Cross validation is a statistical technique used to evaluate a machine learning model on independent data. The procedure has one parameter, called k, which refers to the number of groups into which the given data sample should be divided. Thus, the procedure is often referred to as k-fold cross-validation. When choosing a specific value for k, in this study k=10, a 10-fold cross-validation is performed. This is a popular method because it results in a less biased assessment of model performance than other methods such as simple train/test split. The Figure 6 demonstrates the process of the data division into train and test [29]

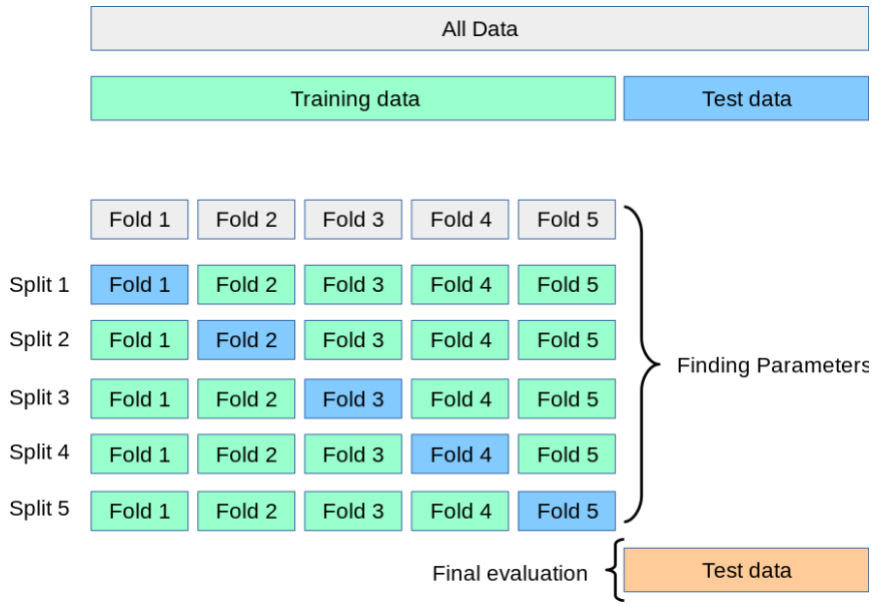


Figure 6. Grid search k-fold cross validation [30].

4.2.4. THE EVALUATION METRICS

The performance score of the prediction model is evaluated by Accuracy rate, Error rate, Sensitivity, Specificity and Precision. From the confusion matrices the measures used in equations 6-10 can be obtained. These matrices report True Negative (TN), False Positive (FP), False Negative (FN) and True Positive (TP). In order to calculate the model performance representing indicators the following formulas are used [31]:

$$\text{Accuracy rate} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \quad (6)$$

$$\text{Error rate} = \frac{\text{FN} + \text{FP}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \quad (7)$$

$$\text{Sencitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (9)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

5. RESULTS AND DISCUSSION

5.1. PREPROCESSING RESULT

5.1.1. MISSING VALUES REJECTION OR IMPUTATION

The analysis of the DD2019 represents that the 4.9% (47 rows) of the samples have at least one null attribute. Among the 18 attributes four of them (BMI, Pregnancies, Pdiabetes and Diabetic) have a null value (Figure 7). It was decided to remove empty values, since their content is less than 5%, that will contribute to the improvement the dataset quality.

```
Age False
Gender False
Family_Diabetes False
highBP False
PhysicallyActive False
BMI True
Smoking False
Alcohol False
Sleep False
SoundSleep False
RegularMedicine False
JunkFood False
Stress False
BPLevel False
Pregancies True
Pdiabetes True
UriationFreq False
Diabetic True
```

Figure 7. The DD2019 attribute analysis on a null value.

In the case of PIDD, the previously used method, namely the removal of missing values, cannot be applied as the only correct one, since 49% of the rows contained missing values. In most of the previously described works, researchers used the mean value imputation technique [32]. However, in this dataset, the use of this method was rejected, since equating the attributes of such many samples (patients) to common/mean value without considering their basic features (such as age, BMI, gender, etc.) introduces huge deviations and inaccuracies in final result. Figure 8 shows the distribution of missing values with respect to attributes. Also, it is important to note that a null value for the pregnancy attribute is not treated as a missing value.

The primary action was to remove missing values for features such as: Glucose, Blood Pressure and BMI. The number of deleted rows was 5.7% (44 rows). The next step was to fill in the empty values using the Iterative imputation method, but this was done after the outlier rejection part (Section 3.1.2).

```
Zero values in Pregnancies : 111
Zero values in Glucose : 5
Zero values in BloodPressure : 35
Zero values in SkinThickness : 227
Zero values in Insulin : 374
Zero values in BMI : 11
Zero values in DiabetesPedigreeFunction : 0
Zero values in Age : 0
```

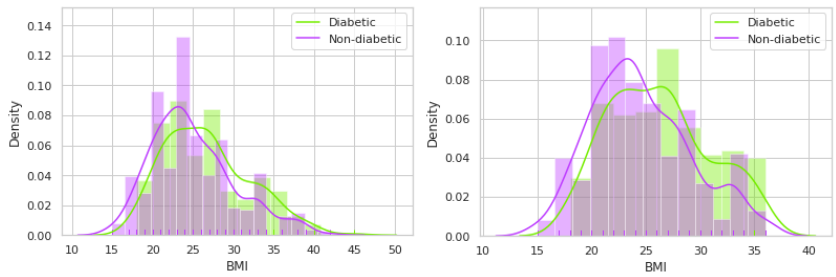
Figure 8. The PIDD attribute analysis on a missing value.

5.1.2. OUTLIER REJECTION

In DD2019 the presence of outliers is considered only for 4 attributes (BMI, Sleep and Sound Sleep), since all the rest are categorical values. However, in this case there is an exception in the form of a pregnancy attribute. It has a numerical value, but its range is from 0 to 4 (Table 1.1.), which is a normal indicator for a person and does not have any extreme values. Results of the outlier identification can be clearly seen on the figure 9. A comparative analysis of changes after the outlier rejection is made only for the BMI attribute, since only outliers were found in it (Figure 10).

```
Outliers in BMI: 33
Outliers in Sleep: 0
Outliers in SoundSleep: 0
```

Figure 9. The outlier identification in DD2019.



(a) (b)

Figure 10. The comparison of (a) initial BMI attribute's classes distribution and (b) after the outlier rejection.

The number of outliers in PIDD is represented on figure 11. Figure 12 shows the class distribution in the initial PIDD, one can notice the difficulty of differentiating diabetic and non-diabetic patients. Also, it is possible to identify outliers in a dataset by having a wider and sharper shape with flat tails, and positive and negative skewness. The method of outlier rejection has affected the distribution of classes, the shape has become narrower and flatter with shorter tails, also the symmetry is improved. This can be seen by comparing Figure 12 and 12.1.

```
Outliers in Pregnancies: 4
Outliers in Glucose: 0
Outliers in BloodPressure: 14
Outliers in SkinThickness: 1
Outliers in Insulin: 29
Outliers in BMI: 7
Outliers in DiabetesPedigreeFunction: 28
Outliers in Age: 7
```

Figure 11. The outlier identification in PIDD.

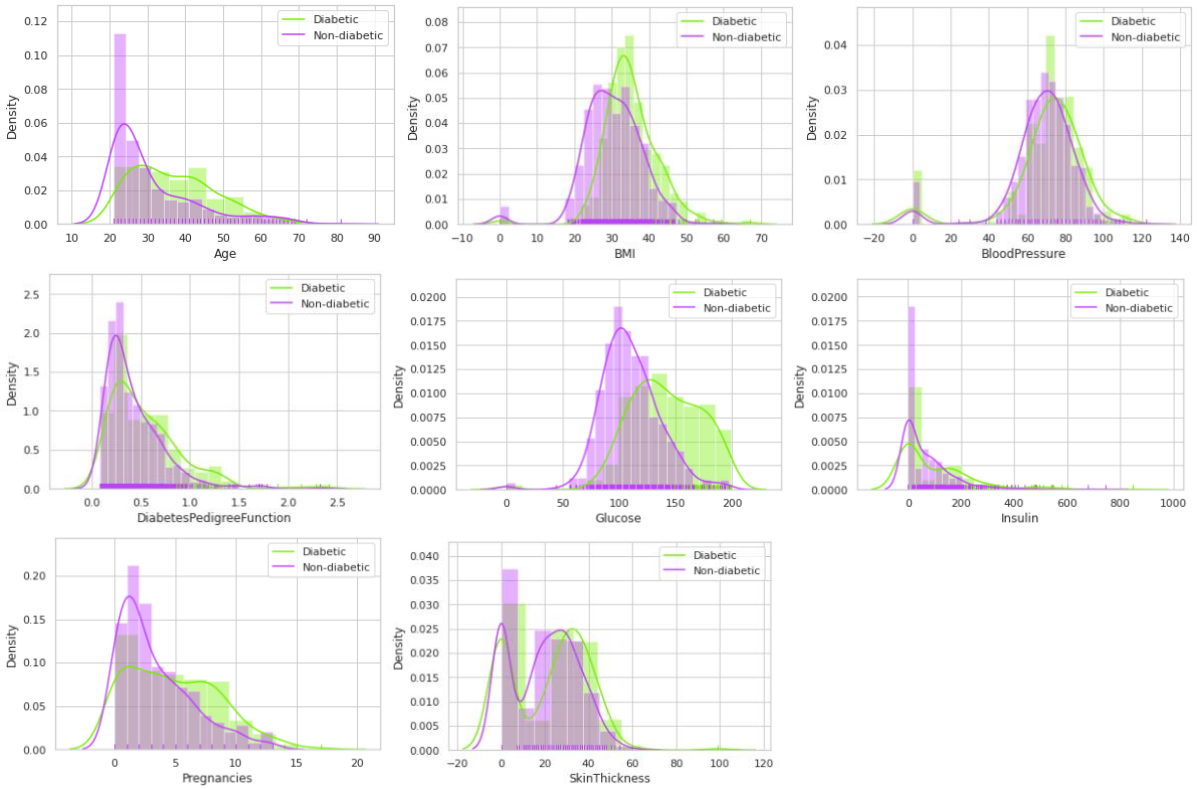


Figure 12. The initial PIDD attributes overview.

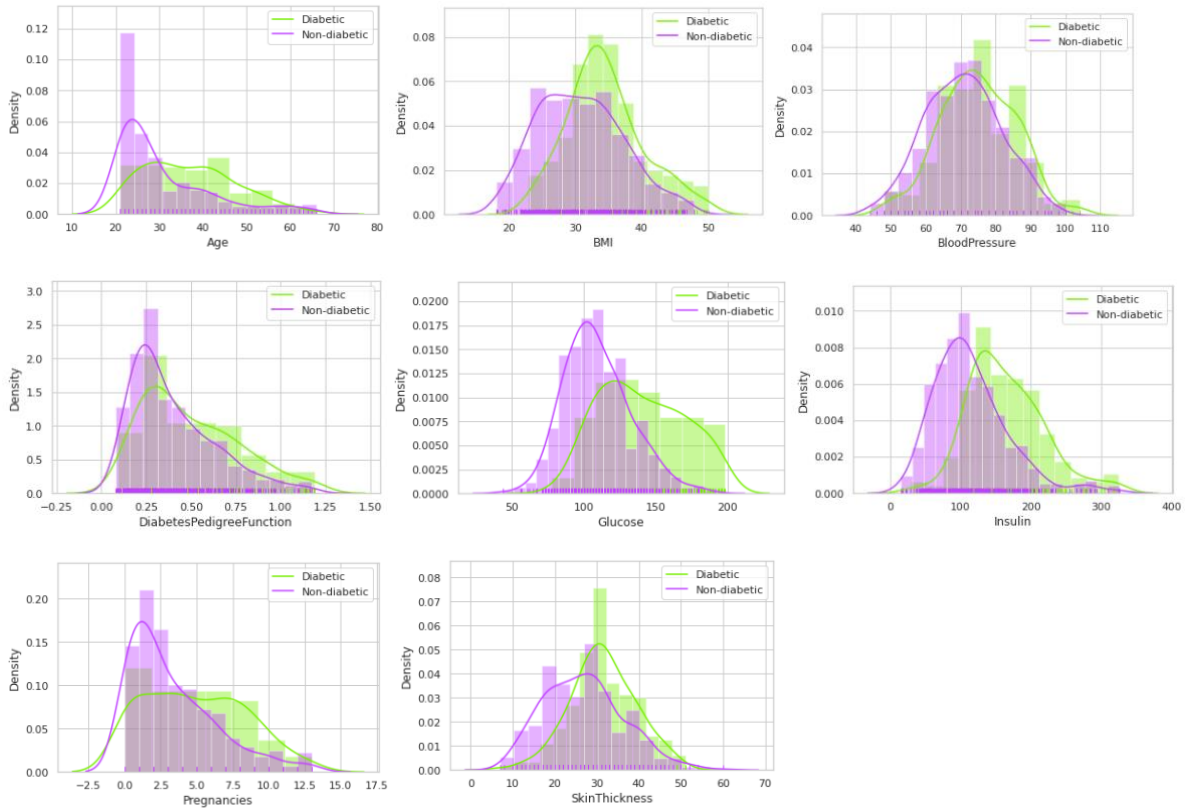


Figure 12.1. The PIDD attributes overview without outliers.

5.1.3. DATA NORMALIZATION

In the process of testing models using PIDD at various stages of preprocessing, a pattern was revealed according to which the accuracy of models decreases on average from 1% to 4% with the data normalization method. For this reason, it was decided not to use this method on PIDD. However, in the case of DD2019, the use of data normalization technique had a positive effect on the predictive accuracy of the models. For DD2019, the data normalization technique was applied only on BMI, Sleep, Sound Sleep and Pregnancy numerical attributes (Table 1).

5.1.4. FINAL PREPROCESSED DATA

After the preprocessing part, the 8.4% of the initial DD2019 were removed by missing values and outlier rejection. The number of samples changed from 952 to 872. The implementation

of the conversion of the categorical features to numerical method changed the number of attributes from 17 to 32.

The 16.7% of the original PIDD has been removed after the missing values deletion of the Glucose, Blood Pressure and BMI attributes, and the outlier rejection. Also, the 12.2% of data was changed after the iterative imputation technique. The overall number of rows reduced to 640 from the initial 768.

5.2. EXPERIMENTAL RESULT

In the datasets, the diabetic parameter ('Diabetic' for DD2019 and 'Outcome' for PIDD) was considered as the dependent variable, all other attributes were taken as independent variables. The diabetic parameter has a division into 2 classes: class 1 - diabetic and class 0 - non-diabetic. To train the models, the dataset was divided into training and test sets in a ratio of 80:20, respectively. The evaluated performance of all seven algorithms using Confusion Matrix is as follows:

Table 3. Confusion matrix of algorithms used

	LR		KNN		SVM		NB		DT		RF		L2	
DD2019	144	11	150	5	150	5	129	26	149	6	150	5	141	14
	13	50	5	58	4	59	12	51	5	58	5	58	12	51
PIDD	95	14	101	8	102	7	92	17	87	22	96	13	97	12
	26	24	28	22	30	20	22	28	23	27	29	21	27	23

Table 4 shows algorithms used with their hypermeters. Hyperparameters and its optimization or tuning are key to managing the machine learning process. The selection of the necessary hyperparameters is one of the main tasks to achieve the best performance.

Table 4. The ML models and preprocessed tuned hyperparameters

	Algorithm	Best hyperparameters
1	Logistic Regression	multi_class = 'multinomial' others default
2	K-Nearest Neighbor	n_neighbors = 5 weights = 'distance' others default
3	Support Vector Machine	kernel = 'poly' C = 3 others default
4	Naïve Bayes	var_smoothing = 0.01 others default
5	Decision Tree	max_features = 'log2' criterion = 'entropy' others default
6	Random Forest	n_estimators = 100 max_features = 'sqrt' others default
7	L2 regularization	alpha = i * 0.01

The performance of all seven models according to the selected main indicators (Accuracy, Error, Sensitivity, Specificity, Precision, 10-fold CV) is shown in Table 5.

Table 5. The performance of the proposed algorithms.

	LR		**KNN**		**SVM**		**Naive Bayes**		**Decision Tree**		**Random Forest**		**L2 regularization**	
	DD2019	PID	DD2019	PID	DD2019	PID	DD2019	PID	DD2019	PID	DD2019	PID	DD2019	PID
Accuracy	0.890	0.748	0.954	0.774	0.959	0.767	0.826	0.755	0.954	0.712	0.954	0.761	0.881	0.755
Error	0.110	0.252	0.046	0.226	0.041	0.233	0.174	0.245	0.046	0.289	0.046	0.239	0.119	0.245
Sensitivity	0.794	0.480	0.921	0.440	0.937	0.400	0.810	0.560	0.937	0.540	0.921	0.440	0.810	0.460
Specificity	0.929	0.872	0.968	0.927	0.968	0.936	0.832	0.844	0.955	0.789	0.968	0.908	0.910	0.890
Precision	0.818	0.632	0.921	0.733	0.922	0.741	0.662	0.622	0.894	0.540	0.921	0.688	0.785	0.657
10-fold CV	0.901	0.778	0.965	0.761	0.931	0.784	0.856	0.771	0.963	0.736	0.953	0.771	0.895	0.788

According to Table 5, the main observation that can be made is that the accuracy rate of all models based on DD2019 is higher than PIDD models. This may be due to 3 main

factors: the number of rows in the data set, the quality of the data used, and the number of attributes relevant to diabetes prediction. Four out of seven algorithms demonstrate accuracy above 90%. However, the K-Nearest Neighbor and Support Vector Machine algorithms have the best result, this is proved by the Accuracy, Error, Sensitivity, Specificity and Precision indicators, which means that the classification of these algorithms is highly accurate.

The analysis of the obtained results also includes an assessment of the feature importance for each of the models. Table 6 shows the three most important attributes for the two models with the highest accuracy. The attributes are listed in descending order of their importance.

Table 6. The feature importance.

	K-Nearest Neighbour	Support Vector Machine
DD2019	Gender	Regular Medicine
	Stress sometimes	Family Diabetes
	Regular Medicine	Gender
PIDD	Glucose	Insulin
	Insulin	BMI
	Age	Glucose

6. CONCLUSION AND FURTHER WORK

The increase in the incidence of diabetes worldwide is a global health problem. The purpose of this study is to attempt to create a system for predicting the risk of diabetes. In this study, seven machine learning classification algorithms are implemented and their comparative analysis is carried out according to various statistical indicators. The system was trained on two independent datasets DD2019 and PIDD. There are two algorithms that showed the best accuracy when using both datasets, these are K-Nearest Neighbor and Support Vector Machine for both DD2019 and PIDD (95.4%, 77.4% and 95.9%, 76.7% respectively). After identifying models with the highest accuracy value, an analysis was made of the importance of specific attributes for these algorithms. At the machine learning process on DD2019, for the K-Nearest Neighbor algorithm the most important attributes are Gender, Stress sometimes, Regular Medicine, while for Support Vector Machine these attributes are Regular Medicine, Family Diabetes, Gender. The analysis of algorithms based on PIDD shows that the attributes Glucose, Insulin, Age have the highest values for the K-Nearest Neighbor algorithm. For the model on the Support Vector Machine, the most important attributes are Insulin, BMI, Glucose. The results obtained can be used in predicting other diseases. This study still requires further improvement and expansion of the field of study. The main direction at the moment is the use of other machine learning algorithms, building multilayer structures and creating the neural network. Also, a possible area for further work is the collection of our own database with as many attributes and participants as possible.

7. REFERENCE LIST

[1] DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., Hu, F. B., Kahn, C. R., Raz, I., Shulman, G. I., Simonson, D. C., Testa, M. A., & Weiss, R. (2015). Type 2 diabetes mellitus. Nature reviews. Disease primers, 1, 15019. https://doi.org/10.1038/nrdp.2015.19

[2] Normal blood sugar ranges and blood sugar ranges for adults and children with type 1 diabetes, type 2 diabetes and blood sugar ranges to determine people with diabetes. Diabetes.co.uk. (2019, January 15). Retrieved February 20, 2022, from https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

[3] Shugart, C., Jackson, J., & Fields, K. B. (2010). Diabetes in sports. Sports health, 2(1), 29–38. https://doi.org/10.1177/1941738109347974

[4] Albright, A., Franz, M., Hornsby, G., Kriska, A., Marrero, D., Ullrich, I., & Verity, L. S. (2000). American College of Sports Medicine position stand. Exercise and type 2 diabetes. Medicine and science in sports and exercise, 32(7), 1345–1360. https://doi.org/10.1097/00005768-200007000-00024

[5] Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. 2013 International Conference on Machine Intelligence and Research Advancement, 203-207. doi: 10.1109/ICMIRA.2013.45

[6] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, AA, Ogurtsova, K ., Shaw, JE, Bright, D., Williams, R., & IDF Diabetes Atlas Committee (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International

Diabetes Federation Diabetes Atlas, 9th edition. Diabetes research and clinical practice, 157, 107843. https://doi.org/10.1016/j.diabres.2019.107843

[7] UCI Machine Learning Repository: Data Set. Retrieved February 20, 2022, from https://archive.ics.uci.edu/ml/datasets/pima%2bindians%2bdiabetes

[8] Han, W., Shengqi, Y., Zhangqin, H., Jian, H., Xiaoyi ,W. (2018). Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, volume 10, 100-107. https://doi.org/10.1016/j.imu.2017.12.006.

[9] Roshan, B., Ashish, K., Ritu, C., Harleen, K. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach, sci. 1, 1112. https://doi.org/10.1007/s42452-019-1117-9

[10] Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F., & Sayadi, M. (2021). Diabetes Diagnosis Using Machine Learning. Frontiers in Health Informatics, 10(1), 65. doi:http://dx.doi.org/10.30699/fhi.v10i1.267

[11] Deepti S., Dilip S. (2018). Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, volume 132, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122.

[12] Joshi, R. D., & Dhakal, C. K. (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. International journal of environmental research and public health, 18(14), 7346. https://doi.org/10.3390/ijerph18147346

[13] Sidong, W., Zhao, X., Chunyan, M. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification, 291-295. DOI: 10.1109/WF-IoT.2018.8355130

[14] Aiswarya, I., Jeyalatha, S., Ronak, S. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process, 5. 1-14. DOI: 10.5121/ijdkp.2015.5101

[15] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in genetics, 9, 515. https://doi.org/10.3389/fgene.2018.00515

[16] Vijayan, V. & Anjali, C.. (2015). Prediction and diagnosis of diabetes mellitus — A machine learning approach, 122-127. DOI: 10.1109/RAICS.2015.7488400

[17] Neha, P. T., Shruti, G. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods, Procedia Computer Science, volume 167, 706-716. https://doi.org/10.1016/j.procs.2020.03.336.

[18] K. P. N. V. Satya, S., Karthik, J., Niharika, C., Srinivas, P., Ravinder, N., Prasad, C. (2021). Optimized Conversion of Categorical and Numerical Features in Machine Learning Models, 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 294-299, doi: 10.1109/I-SMAC52330.2021.9640967.

[19] Zhang, S., Wu, X., Zhu, M. (2010). Efficient missing data imputation for supervised learning, 9th IEEE International Conference on Cognitive Informatics (ICCI'10), 672-679, doi: 10.1109/COGINF.2010.5599826.

[20] Hasan, M., Alam, M., Das, D., Hossain, E., Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, in IEEE Access, 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[21] Muhammad, A., Peshawa, F. (2014). Data Normalization and Standardization: A Technical Report. DOI: 10.13140/RG.2.2.28948.04489

[22] Stoltzfus, C. (2021). Logistic Regression: A Brief Primer, Academic Emergency Medicine 2011; 18:1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

[23] Osuna, Edgar & Freund, Robert & Girosi, Federico. (1970). Support Vector Machines: Training and Applications. Tech Rep A.I. Memo No. 1602.

[24] Cunningham, Padraig & Delany, Sarah. (2007). k-Nearest neighbour classifiers. Mult Classif Syst. 54. 10.1145/3459665.

[25] Szafron, Duane & Greiner, Russ & Lu, Paul & Wishart, David & Macdonell, Cam & Anvik, John & Poulin, Brett & lu, Zhiyong & Eisner, Roman & Ca, Eisner@cs. (2003). Explaining naïve Bayes classifications.

[26] Song, Yan-Yan & Lu, Ying. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 27. 130-5. 10.11919/j.issn.1002-0829.215044.

[27] Louppe, Gilles. (2014). Understanding Random Forests: From Theory to Practice. 10.13140/2.1.1570.5928.

[28] Demir-Kavuk, O., Kamada, M., Akutsu, T. Ernst-Walter K. (2011). Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features, BMC Bioinformatics 12, 412. https://doi.org/10.1186/1471-2105-12-412

[29] Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S. (2012). The 'K' in K-fold Cross Validation, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. http://www.i6doc.com/en/livre/?GCOI=28001100967420

[30] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[31] Karlijn, J., Viand, S., Johannes, B., Friedo, W., Carmine Zoccali, Kitty, J. (2009). Diagnostic methods I: sensitivity, specificity, and other measures of accuracy, Kidney International, 75 (12), 1257-1263. https://doi.org/10.1038/ki.2009.92.

[32] Hasan, K., Alam, M., Das, D., Hossain, E., Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, in IEEE Access, vol. 8, 76516-76531. doi: 10.1109/ACCESS.2020.2989857.

LIST OF TABLES

LIST OF FIGURES