ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Rough Set-Based Feature Selection for Predicting Diabetes Using Logistic Regression with Stochastic Gradient Decent Algorithm

Kanaan M. Kaka-Khan[1], Hoger Mahmud[2], Aras Ahmed Ali[3]

[1]Department of Information Technology, University of Human Development, Iraq, [2]Department of Information Technology, the American University of Iraq, Sulaimani, [3]University College of Goizha, Sulaymaniyah

ABSTRACT

Disease prediction and decision-making plays an important role in medical diagnosis. Research has shown that cost of disease prediction and diagnosis can be reduced by applying interdisciplinary approaches. Machine learning and data mining techniques in computer science are proven to have high potentials by interdisciplinary researchers in the field of disease prediction and diagnosis. In this research, a new approach is proposed to predict diabetes in patients. The approach utilizes stochastic gradient descent which is a machine learning technique to perform logistic regression on a dataset. The dataset is populated with eight original variables (features) collected from patients before being diagnosed with diabetes. The features are used as input values in the proposed approach to predict diabetes in the patients. To examine the effect of having the right variable in the process of making predictions, five variables are selected from the dataset based on rough set theory (RST). The proposed approach is applied again but this time on the selected features to predict diabetes in the patients. The results obtained from both applications have been documented and compared as part of the approach evaluations. The results show that the proposed approach improves the accuracy of predicting diabetes when RST is used to select variables for making the prediction. This paper contributes toward the ongoing efforts to find innovative ways to improve the prediction of diabetes in patients.

Index Terms: Logistic Regression, Stochastic Gradient Descent, Rough Set Theory, K-fold Cross-validation, Diabetes Prediction

1. INTRODUCTION

Changes in human lifestyle and the deterioration of the environment have left a negative impact on human health. For that reason, human health has always been the subject of research with the aim to improve it. Diabetes is a group of metabolic diseases which result in high blood sugar levels for a prolonged period. As stated by International Diabetes Federation, 537 million adults (20–79 years) are living with diabetes which is 1 in 10 of adult population. This number is predicted to rise to 643 million by 2030 and 783 million by 2045 [1]. Diabetes has been the subject of research for some times by multidisciplinary scientists with the aim to find and improve methods that lead to effective prevention, diagnosis, and treatment of the disease. For instance, in a similar approach, in 2013, Anouncia *et al.* proposed a diagnosis system for diabetes. The system is implemented to diagnose the type of diabetes based on symptoms provided by patients. They have used rough set-based knowledge representation in developing their system and the results showed improvements in terms of accuracy of diabetes type diagnosis and the time it takes for the diagnosis [2]. Despite all the efforts invested into researching diagnostic techniques for diabetes, research

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp85-93	E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Kanaan M. Kaka-Khan, Department of Information Technology, University of Human Development, Iraq. E-mail: kanaan.mikael@uhd.edu.iq

Received: 21-08-2022 Accepted: 02-10-2022 Published: 18-10-2022

shows that there is still room for improvement, especially in areas related to the level of accurately in predicting the disease in a patient. Rough set theory (RST) has been used by researchers to predict a wide array of topics such as time series prediction [3], crop prediction [4], currency crisis prediction [5], and stock market trends prediction [6]. In this research, we use RST to select variables in a dataset with the aim to improve the level of accuracy in predicting diabetes in a patient. Stochastic gradient descent algorithm is used to process the variables selected to make diabetes prediction based on computed logistic regression values from the dataset. The dataset used for all experiments in this study is made available by the Pima Indian Diabetes [7]. This paper contributes toward the ongoing efforts to find innovative ways to improve the prediction of diabetes in patients by proposing a new approach to predict diabetes in patients using machine learning techniques. The results presented in Sections 5.1 and 5.2 show that the approach improves accuracy in making diabetes predictions compared to other available approaches.

The rest of this paper is organized as follows: Section 2 provides the theoretical background needed to understand the selected techniques and Section 3 provides a survey of related literatures. Section 4 provides the description of the methodology used in this study. Experimental results and discussion are provided in Section 5. Finally, conclusions are drawn in Section 6.

2. BACKGROUND

This section provides a basic background on the theories used in the study.

2.1. RST

Rough set [8] is proposed by Pawlak to deal with uncertainty and incompleteness. It offers mathematical tools to discover patterns hidden in datasets and identifies partial or total dependencies in a dataset based on indiscernibility relation. The technique calculates a selection of features to determine the relevant feature. The general procedures in rough set are as follows:

The Lower Approximation of set *D* is the set of objects in a table of information which certainly belongs to the class *X*:

$$AX = \{xi \in U \mid [xi]_{nd(A)} \subset X\}, \mathrm{X} \in \mathrm{Att} \tag{1}$$

The Upper Approximation of a set *X* includes all objects in a table of information which possibly belongs to the class *X*:

$$AX = \{xi \in U \mid [xi]_{(A)} \cap A \neq \phi\} \tag{2}$$

Boundary Region is the difference between upper approximation set and lower approximation set that is referred to as Bnd (*X*)

$$\beta = AX - \underline{AX} \tag{3}$$

Positive Region is the set of all objects that belong to lower approximation, which means, the union of the lower approximation consist of the union of all the lower approximation sets:

$$\rho = \cup \underline{A} \text{ (Union of all lower sets)} \tag{4}$$

Indiscernibility of positive reign for any G $\subseteq$ *Att* is the associated equivalence relation:

$$\mathrm{IND}\ (\mathrm{G}) = \{(x, y) \in p\times : \forall a \in \mathrm{G}, \alpha\ (x) = (y)\} \tag{5}$$

Reducts are the minimum range representation of the original data without loss of information:

$$\text{reducts } \delta = \min IND \tag{6}$$

2.2. Stochastic Gradient Descent

According to [9], stochastic gradient descent is a function's minimizing process, following the slope or gradient of that function. In general, in machine learning, stochastic gradient descent can be considered as a technique to evaluate and update the weights every iteration, which minimizes the error in training data models. While training, this optimization technique tries to show each and every training sample to the model one by one. For each training sample, the model produces an output (prediction), calculates the error, and updates to minimize the error for the next output, and this process is repeated for a fixed number of epochs or iterations. Equation-7 describes the way of finding and updating the set of weights (coefficients) in a model from the training data.

$$\mathrm{B} = \text{b-learning rate} \times \text{error} \times x \tag{7}$$

Here, b is the coefficient (weight) being estimated, learning rate is a learning value that can be configured between (0.01 and 10), error is the model's predicted error, and x is the input value. The accuracy of the prediction can be calculated simply by dividing the number of corrected predictions by the actual values produced in formula 3.

$$Accuracy = \frac{\sum Correct\ \ predictions}{\sum Actual\ values} \tag{8}$$

2.3. Logistic Regression

Logistic regression [10] is a two-class problems linear classification algorithm. Equation 9 represents the logistic regression algorithm. In this algorithm, to make a prediction (*y*), using coefficient (weight) values, the input values (*X*) are combined in a linear form. Logistic regression produces an output of binary value (0 or 1).

$$yhat = \frac{1.0}{1.0 + e^{-(b0+b1\times x1)}} \tag{9}$$

The foundation of logistic regression algorithm is Euler's number, the estimated output is represented as *yhat*, the algorithm's bias is $b_{0,}$ and the coefficient (weight) for the single input value (x_1) is represented as $b_{1.}$ The logistic regression produces a real value as an output (yhat) which is between 0 and1. To be mapped to an estimated class value, the output needs to be converted (rounded) to an integer value. Each column (attribute) of the dataset has an associate value (b) that should be estimated from the training data and it is the actual model's representation that can be saved for further use.

3. RELATED WORK

Prediction is a widely used approach in many fields of science including healthcare to foresee possible outcomes of a cause. Disease prediction is certainly an area, where researchers have been working by applying a number of different theories including machine learning theories with the aim to find methods to make the most accurate prediction possible. RST is one of the theories used to classify and predict diseases. For instances, the authors of [11] have used the theory to classify medical diagnosis, the authors of [12] and [13] have modified and used the theory to improve disease prediction. Type 1 and 2 diabetes were the focus of the authors of [14], in which they developed a hybrid reasoning model to address prediction accuracy issues. Based on their results, they claim that their approach raises diabetes prediction accuracy to 95% compared to other existing approaches. In 2017, RST was used by the authors of [15] to develop a model for patient clustering in a dataset. The authors considered average values calculated from diabetes indicators in a dataset to cluster the patients in it. In the same year, deep learning was utilized by the authors of [16] to establish an intelligent diabetes prediction model, in which patients' risk factors collected in a dataset were considered to make the prediction.

In 2018, Fuzzy RST is applied first to select specific features in a dataset, later in the process, to improve prediction performance, save processing time, and better diagnosis accuracy that the Optimized Generic Algorithm (OGA) is applied. The results obtained from the study shows that the approach has achieved the objectives of the study [17]. In 2020, Vamsidhar Talasila and Kotakonda Madhubabu proposed the use of RST technique to select the most relevant features to be inputted to the Recurrent Neural Network (RNN) technique for disease prediction. They claimed that the RST-RNN method achieved accuracy of 98.57% [18]. In the same year, Gao and Cheng proposed an improved neighborhood rough set attribute reduction algorithm (INRS) to increase the dependence of conditional attributes based on considering the importance of individual features for diabetes prediction [14]. In 2021, Gadekallu and Gao proposed a model using an approach based on rough sets to reduce the attributes needed in heart disease and diabetes prediction [19]. The main limitation of these studies is the fact that none has considered the quantity and quality of viables used to make diagnostic predictions.

The approach used in this study is similar to the ones used in the surveyed literatures but differs in objectives. We use RST to select the best features in a dataset and use stochastic gradient decent algorithm to compute the logistic regression values from the selected features in the dataset with the aim to improve the prediction accuracy of diabetes in a patient.

4. METHODOLOGY

This section provides insights on the methodology used to achieve the objectives of the study. The methodology is comprised six major steps:

4.1. Step 1

A dataset is selected, examined for suitability and reliability based on a number of characteristics, and uploaded to be analyzed. The dataset selected and uploaded for the purpose of this research is provided by Pima Indians Diabetes [7]. The selected dataset involves predicting diabetes within 5 years in Pima Indians given medical details. The dataset is a 2-class classification problem and consists of 76 samples with 8 input and 1 output variable. The variable names are as follows: Number of Times Pregnant, Plasma Glucose concentration a 2 h in an oral glucose tolerance test, Diastolic Blood Pressure (mm Hg), Triceps Skinfold Thickness (mm), 2-h Serum Insulin (mu U/ml), Body Mass Index (weight in kg/[height in m]^2), Diabetes Pedigree Function, Age, and Class Variable (0 or 1). Before implementing the model, it is highly preferred to do preprocessing due to some deficiencies. Usually, the dataset contains features highly varying in magnitudes, units, and range which may results in inaccurate output [20]. In this work due to use of stochastic gradient descent algorithm, the dataset has been normalized using min-max scaling to bring all values to between 0 and 1. Table 1 shows a sample of the selected dataset.

4.2. Step 2

The selected diabetes dataset is preprocessed and normalized. To increase the efficiency and accuracy of the model, the dataset needs to be pre-processed before applying the proposed model since the data may contain null values, incorrect, and redundant information. In general, data processing involves two major steps: data cleaning and data normalization. Data cleaning means removing incorrect information or filling out missing values to increases the validity and quality of a dataset though applying a number of different methods [21]. In this study, in case of any tuple containing missing values, the missed attribute value assumed to be 0 (this is achieved using the fill_mising_values () function from the python script developed for the implementation phase of this study). Redundant or unnecessary columns are deleted to have a high quality dataset (this is achieved using the remove_duplicate_columns () function from the python script). To let all features have equal weight and contribution to the model, the range of each feature needs to be scaled, for this purpose, the dataset is normalized to a range of [0,1] by the following processes: *String columns converting:* the string columns are converted to float through *str* column using the float() function. *Min max finding:* min and max values of each column of the dataset are found through using the dataset minmax() function. Finally, the dataset is normalized by the min-max normalization method using the following equation adapted form [22].

$$X' = \frac{x - min(x)}{max(x) - min(x)} \tag{10}$$

4.3. Step 3

In this step, RST is applied to select the features which might produce a better prediction. There are *nine* variables in total in the dataset, as shown in Table 1. The class variable is considered as a dependent variable and the other eight variables are assumed as predictors or independent variables. Table 2 presents the regression calculation summary for diabetes classification of the dataset. The result of the calculation clearly shows that the accuracy of diabetes prediction is *30.32%* if all variables in the dataset are considered in the calculation. The low accuracy result is an indication that there might be one or more variables which are not fit to be used for prediction. The regression calculation also shows that the un-standardized regression coefficient (*b*) is 0.06 for pregnancies, which indicates that if all other predictors are controlled then an increment of one unit in pregnancies increases the accuracy by 0.06. The same statement can be made for the other variables. To flitter the features that might produce a better diabetes prediction, the dataset is grouped together into nine elementary sets based on indiscernibility relation level between the data elements. Table 3 shows the details of the groups. To further process the groups, the discernibility matrix has been developed for the elementary sets and the result is shown in Table 4. From the discernibility matrix, a discernibility function has been developed, as shown in equation 11.

$$f(A) = f(A1) \times f(A2) \times \ldots \times f(An) \tag{11}$$

As the result of discernibility function of all elementary sets for the entire dataset, we found that:

$f(A) = a1 \vee a2 \vee a5 \vee a6 \vee a8$ where a1 is Pregnancies; a2 is Plasma glucose; a5 is Insulin; a6 is DPF; and a8 is age attribute. Table 5 shows the reduct matrix for the elementary sets. From the reduct matrix, all reducts and core attributes have been found:

TABLE 1: The first ten records of the diabetes dataset used in this study

Pregnancies	Plasma glucose	Blood pressure	Skinfold thickness	Insulin	BMI	DPF	Age	Class variable
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

BMI: Body mass index

$f(R1) = a1 \vee a2 \vee a6$; $f(R2) = a1 \vee a2 \vee a5 \vee a8$; $f(R3) = a2 \vee a5 \vee a8$; $f(R4) = a1 \vee a2 \vee a8$; $f(R5) = a2 \vee a6 \vee a8$; $f(R6) = a1 \vee a2 \vee a6 \vee a8$; $f(R7) = a2 \vee a5 \vee a6$; $f(R8) = a1 \vee a2 \vee a5$; $f(R9) = \vee a2 \vee a5 \vee a6 \vee a8$. Finally, Table 6 shows the features that are selected to be used for making diabetes prediction.

TABLE 2: Linear regression statistics of diabetes dataset

Multiple R	0.550684207
R Square	0.303253096
Adjusted R Square	0.295909255
Standard Error	0.400210451

	Coefficients	Standard Error	t Stat	P-value	Unstandardized regression coefficient (b)
Intercept	0.853894266	0.085484958	-9.98882	0.00	0.066
Pregnancies	0.020591872	0.00512998	4.014026	0.00	1.863
Plasma glucose	0.005920273	0.000515123	11.49294	0.00	0.022
blood pressure	0.002331879	0.000811639	2.87305	0.00	0.081
Skinfold thickness	0.00015452	0.001112215	0.13893	0.89	0.247
Insulin	0.000180535	0.000149819	-1.20502	0.23	0.004
MI	0.013244031	0.00208776	6.343656	0.00	0.000
DPF	0.147237439	0.045053885	3.26803	0.00	0.686
Age	0.002621394	0.00154864	1.692707	0.09	0.001

TABLE 3: Elementary sets

Samples	Pregnancies	Plasma glucose	Blood pressure	Skinfold thickness	Insulin	BMI	DPF	Age
Group 1	0–1	0–22	0–13	0–10	0–94	0–6	0–0.25	21–26
Group 2	2–3	23–46	14–28	11–22	95–190	7–14	0.26–0.51	27–33
Group 3	4–5	47–70	29–43	23–34	191–286	15–22	0.52–0.77	34–41
Group 4	6–7	71–94	44–58	35–46	287–382	23–30	0.78–1.03	42–49
Group 5	8–9	95–118	59–73	47–58	383–478	31–38	1.04–1.29	50–57
Group 6	10–11	119–142	74–88	59–70	479–574	39–46	1.3–1.55	58–63
Group 7	12–13	143–166	89–103	71–82	575–670	47–54	1.56–1.81	64–69
Group 8	14–15	167–190	104–118	83–94	671–766	55–62	1.82–2.03	70–75
Group 9	16–17	191–199	119–122	95–99	767–846	63–67	2.04–2.42	76–81

TABLE 4: Discernibility matrix

Samples	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9
Group 1	-								
Group 2	a1a2a4a7a8	-							
Group 3	a2a3a4a8	a1a3a4a8	-						
Group 4	a1a2a4a6a7	a2a3a4a7a8	a1a2a7a8	-					
Group 5	a2a3a5a7a8	a1a3a4a8	a1a2a4a6a7	a2a3a5a7a8	-				
Group 6	a1a3a5a6a8	a3a4a6a8	a1a3a5a7a8	a2a4a5a7a8	a2a3a5a7a8	-			
Group 7	a1a2a4a6a8	a2a4a5a7	a1a2a4a6	a2a3a5a7a8	a2a3a5a8	a5a6a7	-		
Group 8	a1a2a4a6a7	a1a3a4a8	a1a2a7a8	a2a3a5a7a8	a2a4a5a7	a3a4a5	a2a4a5	-	
Group 9	a2a4a5a7	a1a2a4a7	a3a5a8	a2a5a7a8	a3a4a6	a2a3a8	a2a4a5	a3a4a5	-

TABLE 5: Reducts matrix

Samples	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9
Group 1	-	a1a2a4a7a8	a2a3a4a8	a1a2a4a6a7	a2a3a5a7a8	a1a3a5a6a8	a1a2a4a6a8	a1a2a4a6a7	a2a4a5a7
Group 2	a1a2a4a7a8	-	a1a3a4a8	a2a3a4a7a8	a1a3a4a8	a3a4a6a8	a2a4a5a7	a1a3a4a8	a1a2a4a7
Group 3	a2a3a4a8	a1a3a4a8	-	a1a2a7a8	a1a2a4a6a7	a1a3a5a7a8	a1a2a4a6	a1a2a7a8	a3a5a8
Group 4	a1a2a4a6a7	a2a3a4a7a8	a1a2a7a8	-	a2a3a5a7a8	a2a4a5a7a8	a2a3a5a7a8	a2a3a5a7a8	a2a5a7a8
Group 5	a2a3a5a7a8	a1a3a4a8	a1a2a4a6a7	a2a3a5a7a8	-	a2a3a5a7a8	a2a3a5a8	a2a4a5a7	a3a4a6
Group 6	a1a3a5a6a8	a3a4a6a8	a1a3a5a7a8	a2a4a5a7a8	a2a3a5a7a8	-	a5a6a7	a3a4a5	a2a3a8
Group 7	a1a2a4a6a8	a2a4a5a7	a1a2a4a6	a2a3a5a7a8	a2a3a5a8	a5a6a7	-	a2a4a5	a2a4a5
Group 8	a1a2a4a6a7	a1a3a4a8	a1a2a7a8	a2a3a5a7a8	a2a4a5a7	a3a4a5	a2a4a5	-	a3a4a5
Group 9	a2a4a5a7	a1a2a4a7	a3a5a8	a2a5a7a8	a3a4a6	a2a3a8	a2a4a5	a3a4a5	-

TABLE 6: Indiscernibility table

Samples	Pregnancies	Plasma glucose	Insulin	DPF	Age
Group 1	0–1	0–22	*	0–0.25	*
Group 2	2–3	23–46	95–190	*	27–33
Group 3	*	47–70	191–286	0.52–0.77	34–41
Group 4	*	71–94	287–382	*	42–49
Group 5	8–9	95–118	*	*	50–57
Group 6	*	119–142	*	1.3–1.55	58–63
Group 7	12–13	143–166	*	1.56–1.81	64–69
Group 8	14–15	167–190	671–766	*	*
Group 9	*	191–199	767–846	2.04–2.42	76–81

Table 3 shows the indiscernibility level of the relation between the patients.

Table 6 represents the last step of RST process, in which the data are simplified, and the indiscernibility relations are stated. The * symbol means that a certain variable has no impact in a certain case, for example, if the patient's pregnancy is (0–1) and plasma glucose is (0–22) and DPF is (0-0.25), then the patient has diabetes regardless of the value of other attributes, and so on.

4.4. Step 4

In this step, the logistic regression algorithm with stochastic gradient descent technique is applied on the selected features in the previous step. The major steps of the application are as follows:

4.4.1. Dataset loading

The dataset is loaded into the model through load_dataset() function.

4.4.2. Dataset preprocessing

The dataset is preprocessed through str column to float(), dataset minmax(), and normalize dataset() functions accordingly.

4.4.3. Dataset splitting into k folds

The dataset is split into k-folds and trainset. Test set creation for training the model is achieved through cross validation split() function.

4.4.4. Coefficients estimating

Coefficients or weights are the values that determine the model accuracy and can be estimated for training data using stochastic gradient descent. The algorithm uses two parameters to estimate the weights (coefficient), the first one is learning rate to specify the amount of each weight, and it is corrected continuously, while it is updated. The second one is Epochs which is the loop through the training process while updating the coefficient. The Coefficients Estimating *is achieved through* coefficients sgd() *function.*

4.4.5. Coefficients updating

For each instance in the training data, each coefficient is updated throughout all epochs. The error that the model makes is the criteria for updating the coefficients. The simple equation can be used to calculate the error (equation-12).

$$\text{Error} = (\text{Expected output value}) - (\text{Prediction made with the candidate coefficients}) \quad (12)$$

4.5. Step 5

Predictions are generated; equation 7 describes the prediction process which is the most important part of the model. Prediction process will be needed twice: first in stochastic gradient descent to evaluate candidate coefficient values and second in the model when it is finalized to produce outputs (predictions) on test data. The prediction process is achieved through predict() function. Fig. 1 shows the execution flow of the proposed approach.

4.6. Step 6

Finally, the results obtained are compared. Fig. 1 shows the proposed diabetes prediction method.

4.7. Model Performance Evaluation

In this research, k-fold cross-validation technique has been used to evaluate the learned model's performance on unseen data. Cross-validation is a resampling procedure used to validate machine learning models on a limited data sample. Using k-fold, cross-validation means that k models will be construct, evaluated, and through using mean model error, the model's performance is estimated. After rounding the predicted value of each row which is a float number between 0 and 1, it will be compared to its actual value. If they are equal, the prediction is considered as a correct result. Simple error equation (equation 13) will be used to evaluate each model.

$$Accuracy = \frac{No.of\ correct\ results}{Total\ no.of\ samples} * 100 \quad (13)$$

The general procedure is as follows: (1) Shuffle the dataset randomly. (2) Split the dataset into k groups, (3) take a group as a test set and the remaining as a training set, the same procedure will be repeated for each and every group; (4) as usual, the model will be Fitting on the training set and evaluating on the test set, and (5) retain the result (evaluation score) the model can be discarded [17], [23]. For this work,

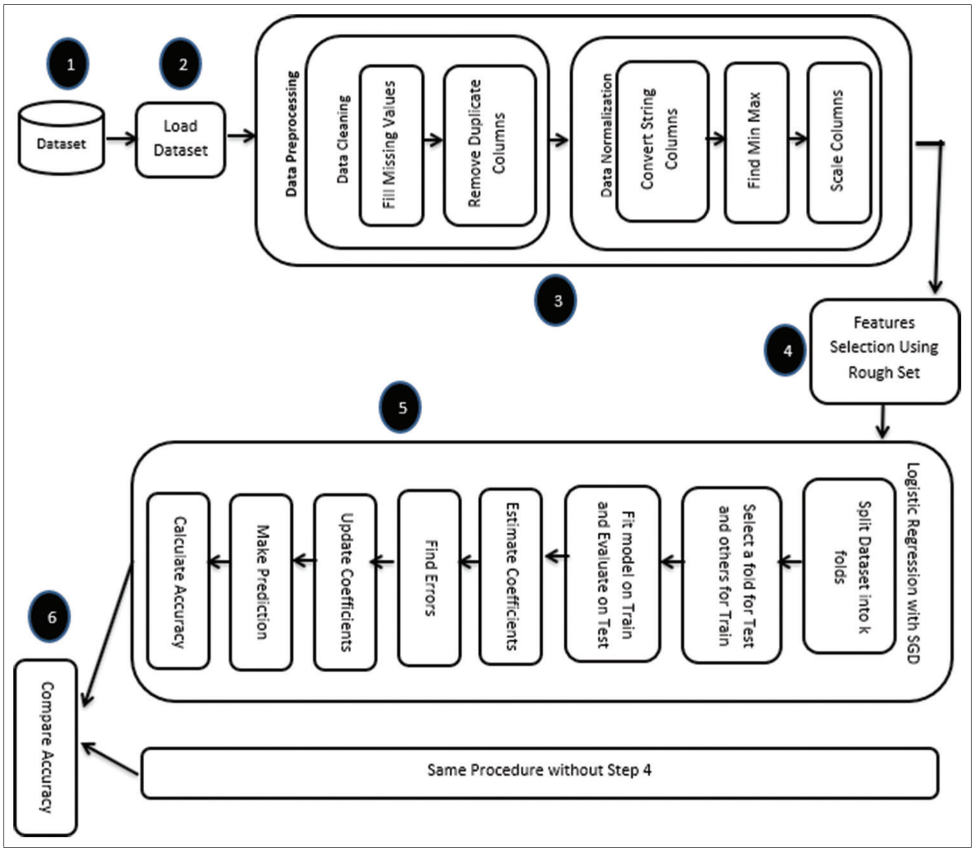

Fig. 1. Proposed diabetes prediction method.

a learning rate, training epochs, and k value are (0.1, 100, 5) subsequently.

After implementing the model twice; first on the dataset with all features, and second with features selected by applying RST, the results can be discussed as follows:

4.8. Making Prediction on Dataset with all Features

The aim of using logistic regression is predicting the dependent variable (output variable) based on equation 7, and the aim of using stochastic gradient descent technique is minimizing the error of predicted coefficient values while training the model on the dataset. For model training, k-fold cross-validation technique is used to split out the dataset to 5 folds (groups), a fold is used as a test set and the others as train sets, for example:

- Mode l: Fold1 for test and fold2, fold3, fold4, and fold5 for train
- Mode 2: Fold2 for test and fold1, fold3, fold5, and fold5 for train
- Mode 3: Fold3 for test and fold1, fold2, fold4, and fold5 for train
- Mode 4: Fold4 for test and fold1, fold2, fold3, and fold5 for train
- Mode 5: Fold5 for test and fold1, fold2, fold3, and fold4 for train.

TABLE 7: Accuracy score of each model used

Model No.	Accuracy
Model 1	73.857
Model 2	78.431
Model 3	81.699
Model 4	75.816
Model 5	75.816
Score	77.124%

For each model, after training for 100 epochs (iterations) and minimizing the errors to a desired results and calculate the accuracy using equation 11, the score can be calculated using equation 14.

$$Score = \frac{Sum\ of\ all\ model\ acuracy\ results}{Total\ no.of\ models} \qquad (14)$$

The total number of models used is five. Table 7 summarizes the models result and the overall score. The overall score is 77.12% for the model on the dataset with all features.

4.9. Making Prediction on Dataset with RST-Based Selected Feature

The same process applied on the dataset with selected features based on RST, the result is presented in Table 8.

TABLE 8: Accuracy and score for all five models for selected features

Model No.	Accuracy
Model 1	77.342
Model 2	81.013
Model 3	83.874
Model 4	78.394
Model 5	79.628
Score	80.215%

TABLE 9: Accuracy and score for all five models using all features, RST-based selected features

Model No.	All features (Accuracy)	RST-based selected features (Accuracy)
Model 1	73.856	77.342
Model 2	78.431	81.013
Model 3	81.699	83.874
Model 4	75.816	78.394
Model 5	75.816	79.628
Score	77.124%	80.215%

RST: Rough set theory

TABLE 10: Accuracy summery of baseline and proposed algorithm for diabetes

Model name	Prediction accuracy (%)
Baseline score	65
Logistic regression with SGD algorithm	77.124
RST-based logistic regression with SGD algorithm	80.215

TABLE 11: Dataset classification comparison

Works	Data size	Methods	Accuracy (%)
[24]	768 samples with 9 attributes	Logistic Regression	77
[25]	768 samples with 9 attributes	Modified PSO Naïve Bayes	78.6
[26]	768 samples with 9 attributes	Modified Weighted knn (SDKNN)	83.76
[27]	768 samples with 9 attributes	random forest classifier	79.57
Our proposed method	768 samples with 9 attributes	Logistic regression with SGD algorithm	77
	768 samples with 6 attributes	RST-based logistic regression with SGD algorithm	80.215

RST: Rough set theory

Table 9 shows the comparison between the results obtained from both implementations; implementing the model on the dataset with all features and the RST-based selected features. The results show that RST-based selected features for machine learning compared to the data set with all features give more accurate predictions.

The baseline score for the selected dataset is 65% our experiment results which indicated that the proposed approach increased the prediction accuracy for diabetes dataset with all features from 65% to 77% and 80% for RST-based features dataset, as shown in Table 10.

Finally, it can be summarized that implementing the logistic regression algorithm with stochastic gradient descent technique is one of the suitable choices for diabetes predictions on the basis of the results. At the same time, rather than using all features, more precise predictions can be made by feature selection based on rough set for neural network. Table 11 summarizes a comparison between our works with some of the most recently published works.

5. CONCLUSION AND FUTURE WORK

In the health-care sector predicting, the presence or non-presence of diseases is important to help people know their health status so that they take the necessary steps to control the disease.

This paper explores the use of stochastic gradient descent algorithm to apply logistic regression on datasets to make predictions on the presence of diabetes. The Pima Indian Diabetes dataset is used to produce results using the proposed technique. The experiments results show that diabetes can be predicted more accurately using logistic regression with stochastic gradient descent algorithm when RST is used to select the important features on a normalized dataset. This is paper makes a real contribution in the use of interdisciplinary techniques to improve prediction mechanisms in health-care sector in general diabetes prediction in specific. The main purpose of this work is showing the significance of using RST with machine learning algorithms, hence in the future; the same theory can be applied with other algorithms to have a better result.

REFERENCES

[1] "Diabetesatlas". Available from: https://www.diabetesatlas.org [Last accessed on 2022 Aug 08].

[2] M. Anouncia, C. Maddona, P. Jeevitha and R. Nandhini. "Design of a diabetic diagnosis system using rough sets". *Cybernetics and Information Technologies*, vol. 13, no. 3, pp. 124-169, 2013.

[3] F. E. Gmati, S. Chakhar, W. L. Chaari and H. Chen. "A rough set approach to events prediction in multiple time series". In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, vol. 10868, pp. 796-807, 2018.

[4] H. Patel and D. Patel. "Crop prediction framework using rough set theory". *International Journal of Engineering and Technology*, vol. 9, pp. 2505-2513, 2017.

[5] S. K. Manga. "Currency crisis prediction by using rough set theory". *International Journal of Computer Applications*, vol. 32, p. 48-52, 2011.

[6] B. B. Nair, V. Mohandas and N. Sakthivel. "A decision tree-rough set hybrid system for stock market trend prediction". *International Journal of Computer Applications*, vol. 6, no. 9, pp. 1-6, 2010.

[7] "Pima-Indians-Diabetes-Dataset". Available from: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Last accessed on 2022 May 04].

[8] Z. Pawlak. "Rough set theory and its applications to data analysis". *Cybernetics and Systems*, vol. 29, no. 7, pp. 661-688, 1998.

[9] P. Achlioptas. "*Stochastic Gradient Descent in Theory and Practice*". Stanford University, Stanford, CA, 2019.

[10] J. Brownlee. *Machine Learning Algorithms from Scratch with Python*. Machine Learning Mastery, 151 Calle de San Francisco, US, 2016.

[11] H. H. Inbarani and S. U. Kumar. "A novel neighborhood rough set based classification approach for medical diagnosis". *Procedia Computer Science*, vol. 47, pp. 351-359, 2015.

[12] E. S. Al-Shamery and A. A. R. Al-Obaidi. "Disease prediction improvement based on modified rough set and most common decision tree". *Journal of Engineering and Applied Sciences*, vol. 13, no. Special issue 5. pp. 4609-4615, 2018.

[13] R. Ghorbani and R. Ghousi. "Predictive data mining approaches in medical diagnosis: A review of some diseases prediction". *International Journal of Data and Network Science*, vol. 3, no. 2, pp. 47-70, 2019.

[14] R. Ali, J. Hussain, M. H. Siddiqi, M. Hussain and S. Lee. "H2RM: A hybrid rough set reasoning model for prediction and management of diabetes mellitus". *Sensors*, vol. 15, no. 7, pp. 15921-15951, 2015.

[15] S. Sawa, R. D. Caytiles and N. C. S. Iyengar. "A Rough Set Theory Approach to Diabetes". In: *Conference: Next Generation Computer and Information Technology*, 2017.

[16] S. Ramesh, H. Balaji, N. Iyengar and R. D. Caytiles. "Optimal predictive analytics of pima diabetics using deep learning". *International Journal of Database Theory and Application*, vol. 10, no. 9, pp. 47-62, 2017.

[17] K. Thangadurai and N. Nandhini. "Integration of rough set theory and genetic algorithm for optimal feature subset selection on diabetic diagnosis". *ICTACT Journal on Soft Computing*, vol. 8, no. 2, 2018.

[18] V. Talasila, K. Madhubabu, K. Madhubabu, M. Mahadasyam, N. Atchala and L. Kande. "The prediction of diseases using rough set theory with recurrent neural network in big data analytics". *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 10-18, 2020.

[19] T. R. Gadekallu and X. Z. Gao. "An efficient attribute reduction and fuzzy logic classifier for heart disease and diabetes prediction". *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 14, no. 1, pp. 158-165, 2021.

[20] "Medium". Available from: https://www.medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e [Last accessed on 2022 Jun 05].

[21] E. Rahm and H. H. Do. "Data cleaning: Problems and current approaches". *IEEE Data Engineering Bulletin*, vol. 23, no. 4, pp. 3-13, 2000.

[22] D. Borkin, A. Némethová, G. Michal'conok and K. Maiorov. "Impact of data normalization on classification model accuracy". *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, vol. 27, no. 45, pp. 79-84, 2019.

[23] "Machine Learning Mastery". Available from: https://www.machinelearningmastery.com/k-fold-cross-validation [Last accessed on 2022 Aug 06].

[24] G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta and S. K. Tayebati. "Comparative machine-learning approach: A follow-up study on Type 2 diabetes predictions by cross-validation methods". *Machines*, vol. 7, no. 4, pp. 74, 2019.

[25] D. K. Choubey, P. Kumar, S. Tripathi and S. Kumar. Performance evaluation of classification methods with PCA and PSO for diabetes. *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, p. 5, 2020.

[26] R. Patra and B. Khuntia. "Analysis and prediction of Pima Indian diabetes dataset using SDKNN classifier technique". *IOP Conference Series: Materials Science and Engineering*, vol. 1070, no. 1, p. 012059, 2021.

[27] V. Chang, J. Bailey, Q. A. Xu and Z. Sun. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms". *Neural Computing and Applications*. vol. 34, no. 10, pp. 1-7, 2022.