RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES

Social Science Journal

Using Machine Learning, Predict Type-2 Diabetes Classification Approaches

Mr. Somendra Tripathi, Hari Om Sharan,C. S.Raghuvanshi

Faculty of Engineering & Technology Rama University Uttar Pradesh, Kanpur, India
[a]somendra.tripathi@gmail.com, [b]drsharan.hariom@gmail.com, [c]drcsraghuvanshi@gmail.com

Keywords: AUC, MCC, Naive Bayes, DT, artificial neural network (ann, SVM, LR, and Randomized Forrest

Abstract:

In India, and over 30 million individuals possess diabetes, and many more are at risk. In order to prevent diabetes and the health issues it is connected with, early detection and treatment are necessary. This study tries to evaluate a person's risk of developing diabetes based on their lifestyle and family history. Different machine learning algorithms were used to predict the risk of Type 2 diabetes since these algorithms is quite accurate, which is crucial in the medical field. People can self-assess potential risk of diabetes that once model has been trained correctly accurately. 752 instances have been collected for the investigation through an offline and online assessment with questions pertaining concerning health, lifestyle, and family medical history. The Apache Indian Diabetes database also utilized the same algorithms. For both datasets, Random Forest Classifier is determined to perform at the highest level of correctness.

1. Introduction:

Diabetes, often called diabetes mellitus (DM), is indeed a collection of diabetic complications that are characterized by chronically elevated levels of blood glucose. High glucose levels also cause excessive urination, constant thirst, and reduced appetite [1].

Diabetes can progress to diabetic ketoacidosis, a hyperosmolar hyperglycemic condition, or even death if it is not treated in a reasonable timeframe. This may have long-term impacts. Heart disease, central nervous system stroke, nephritic syndrome, pressure sores, and eye consequences, etc. [2]. Absolute insulin deficiency occurs when the body's adrenal glands can still produce a sufficient amount of insulin or when the body's tissues and organs are unable to use the insulin that's also produced. There really are three different types of diabetes mellitus [3]:

- **Diabetes Mellitus Type-1:** Diabetes Mellitus Type-1, commonly called as "insulin-subordinate insulin resistance," is characterised by the pancreas producing less insulin than the body requires (IDDM). To compensate for the impaired insulin production by the pancreas in type-1 DM patients, exogenous insulin treatment is recommended.
- **Diabetes Mellitus Type-2:** Diabetic patients type-2 is indicated by the body's insulin resistance because of body's cells behave to the hormone differently since then ordinarily would. In the conclusion, the body might not generate any insulin. Alternative synonyms for this includes "adult commencing diabetes" (ASD) and "non-insulin subordinate diabetes mellitus" (NIDDM). People with high BMIs or those who lead sedentary lifestyles are more likely to have this kind of diabetes.
- **Gestational diabetes type-3:** The last underlying basis shown during pregnancy is gestational diabetes.

Normally, a human's levels of blood glucose fall between 80 to 100 milli-grammes per deciliter. Only when a person's fasting blood glucose level is proven to be high than 126 mg/ dl is a person diagnosed as having diabetes. Pre-diabetic status is characterized in clinical practice as having a glucose level between 100 and 125 mg/dl [4]. A person like this is more likely to acquire type-2 diabetes. Over time, it has been shown that those with the following health conditions are much more susceptible to developing diabetes:

- A Body Mass Index of at least 25
- Those in the family who have diabetes
- People whose systems have HDL cholesterol concentrations < 45 mg/dl
- Experiencing gestational diabetes and chronic hypertension
- Patients who have previously suffered from polycystic ovarian syndrome
- Over-45-year-old representatives of ethnic communities like African Americans, Native Americans, Latin Americans, or Asian-Pacific

When a doctor determines that a patient has pre-diabetes, they advise them to change their lifestyle. A healthy diet and exercise routine can help avoid diabetes [5].

The purpose of this research is to estimate a human's likelihood of developing diabetes. The article's subsequent parts incorporate section 2's relevant research. Section 3 provides a brief description of the machine learning algorithms used. The findings are discussed in part 5, while the approach is covered in section 4. The conclusion is outlined in Section 6.

2. Related work

Almost over 70% of the adult Indian population suffering from hyperglycemia, rendering it a key issue. By combining various techniques including machine learning and data mining, several researchers have attempted to anticipate the signs of diabetes [11]. A select few of them have also used genetic algorithms and neural networks. Since the topic of diabetes prediction is supervised in nature, several have employed machine learning, data mining, and ANN supervised approaches.

This study describes this few works that are very closely connected. For the goal of predicting diabetes, several research papers have made use of the Pima Indians Diabetes Dataset (PIDD). Weka tool and machine learning techniques were used by [13, 14, 16, 17, 20, 21, 23]. Researcher methods may be generically categorised into four categories: machine learning techniques, data mining techniques, hybrid methods, and genetic or neural network algorithms.

employed electrocardiogram (ECG) readings and deep learning techniques to identify diabetes in [12]. They specifically employed convolution neural networks and long short-term memory, and subsequently support vector machines were used to extract features. As a consequence, they discovered an extremely high accuracy of 95.7%. In the goal of predicting diabetes, [13] utilized three machine learning techniques to PIDD: decision tree (DT), naive based (NB), and support vector machine (SVM). The accuracy of the Naive Bayes classifier was measured to be 76.30%.

In [14], data mining methods were used to reliably predict up to 95.42% of a person's chance of acquiring type 2 diabetes. These approaches included enhanced kNN and logistic regression. The adjustment was carried out by empirically choosing the first seed point's value. By doing 100 runs and choosing the least value of the "inside clustering sum of squared errors," the first seed point was established.

In [15] especially in comparison regression analysis, the artificial neural network (ANN), and decision tree (DT) for assessing the likelihood of diabetes and pre-diabetes premised on 12 risk variables, which would include education level, work stress, BMI, age, sleep duration, gender, marital status, family history of diabetes, coffee consumption, preference for salty foods, physical activity, and fish consumption. Among the three approaches, DT was discovered to deliver the greatest outcomes.

Applied a hybrid method in [16] which also incorporates the genetic algorithm (GA) for feature selection and the radial basis function neural network (RBFNN) for classification. They discovered that the hybrid approach outperformed RBFNN on its own.

The number of pregnancies, BMI, and glucose level were revealed to be the most significant factors for diabetes prediction among all parameters in [17] when logistic regression was used to PIDD for diabetic prediction.

[18] used naive bayes, IB1, and C4.5 algorithms to do feature selection and diabetes identification. The research has come to the conclusion that the most essential variables for blood sugar administration are patient age, the frequency of the diagnosis, the requirement for insulin, and

food control. The style of care, home monitoring, and significance of smoking are some additional factors that also have an impact on the results.

Raw data was gathered in [19] in the form of Electronic Reports (EHR) from a variety of sources, including clinical reports, prescriptions written by doctors, reports from diagnostic centers, pharmacy-related data, and data requested by insurance personnel. All of this data combined with a map reveals precise traits that are closely associated to diabetes.

[20] investigated logistic regression, ANN, logistic regression, and naive bays as four classification techniques. All were processed to further bagging and boosting, and random forest was also used. All groups managed a maximum accuracy of between 84% and 86%.

Going to follow extracting the features using untrained methodologies including such Principle Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) approaches, Random Forest, J48, and ANN were employed in [21] for classification. It is discovered that mRMR accuracy is superior to PCA with all characteristics.

The danger of metabolic syndrome and diabetes was a concern in [22]. Naive Bays and the J48 (C4.5) decision tree model were used for prediction, and k-medoids sampling was used to balance the training set. NB did better than the others in their investigation.

The influence of several data mining algorithms for diabetes diagnosis is summarised in [23]. The following classification techniques were used: Multilayer Perceptron (MLP), Bayes Classification, J48graft, JRip (RIP-PER), and Fuzzy Lattice Reasoning (FLR). Most accurate was determined to be J-48 surgical intervention.

Machine learning algorithms were used on individuals in [24] who had a history of cardiovascular risk but did not have diabetes. From Rama University Hospital, five years' collection of data has been acquired in the form of either an EMR. Then, 10-fold classification algorithm was used utilizing machine learning techniques. Logistic regression model used to have the maximum accuracy.

3.0 An Introduction of Machine Learning Classification Models

3.1 Method of Logistic Regression

A version of reinforcement methods called logistic regression utilises the sigmoid function to estimate probabilities in order to compute the relationship between someone binary explanatory variables and at least one independent variable. Contrary to its name, logistic regression is not used to solve regression concerns but instead is a variety of classification problem within which the independent variable might be binominal, ordinal, interval, or ratio-level and the dependent variable is dichotomous (0/1, -1/1, true/false). The following is the sigmoid/logistic function: [6]

$$\text{푦} = \frac{1}{1+\text{푒}^{-\text{푥}}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where y is the consequence acquired by weighing the total of the input variables x. The output is 1 if it is more than 0.5; else, it is 0.

3.2 Classifier K-Nearest Neighbor

Although the K-Nearest Neighbor (KNN) approach is typically employed in business to address classification issues, it may be utilized to handle regression-related difficulties as well. The ease of translation and quick computation are its main benefits. The points (2.5, 7) and (5.5, 4.5) in figure 1 will be assigned to any of the clusters. To determine the distances between every new data point and any old data point, the KNN employs the Euclidean distance function. As a result, the numbers (2.5, 7) will be in the green cluster and the numbers (5.5, 4.5) will be in the red cluster [7].

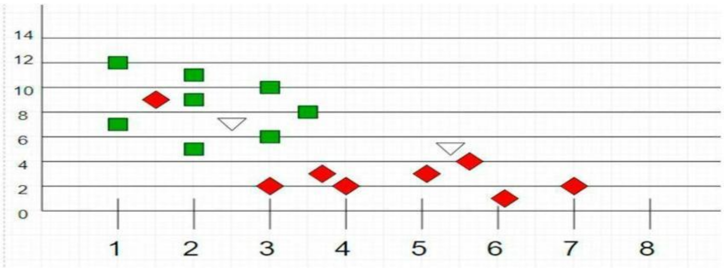


Fig. 1 K-Nearest Neighbor

3.3 Static Vector Machine (SVM)

In machine learning algorithms, SVM is a supervised estimator which may be applied to both classification and regression. It is mostly used to address classification-related issues. SVM attempts to categories data points in a multidimensional space using the proper hyper-plane. A decision boundary for classifying data points is a hyper-plane. With the widest possible gap between the classes and the hyper-plane, the hyper-plane classifies the data points. Support vector machine categorization is shown in Figure 2 [8].

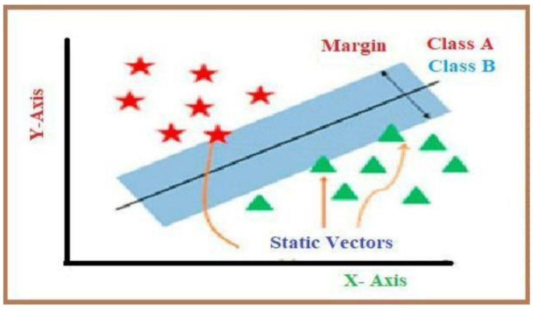


Fig. 2 Static Vector Machine

3.4 Naive Bayes Method to Categorization

A probabilistic machine learning approach based on the probability theory's Bayes theorem is known as the naive bayes classification method. It is one of the finest classifiers because, despite its simplicity, it performs better than other classifiers. Below [9] is the Bayes theorem for computing posterior probability:

폭(ੑ°|폠) ੑ°(ੑ°|폠)ੑ৪(ಆ ੑ৪(ಆ) .... (2)

Where: =)

P(c|x): Posterior Probability

P(x|c): Likelihood

P(c): Class Prior Probability

P(x): Predictor Prior Probability

3.5 Clustering Algorithm Method for Identification

The choice process forms the basis of a tree structure. It has exceptional precision and stability and may be regarded as a tree. A decision tree is shown in Figure 3 [10].

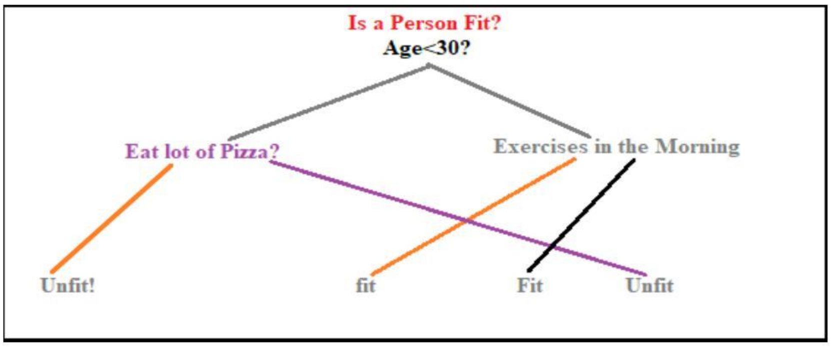


Fig.3 Clustering Algorithm

3.6 Classification with Random Forests

From a chosen at random portion of the training dataset displayed in figure 4, the random forest classifier builds numerous decision trees. The final class of test items is then determined by averaging the votes from several decision trees [29].

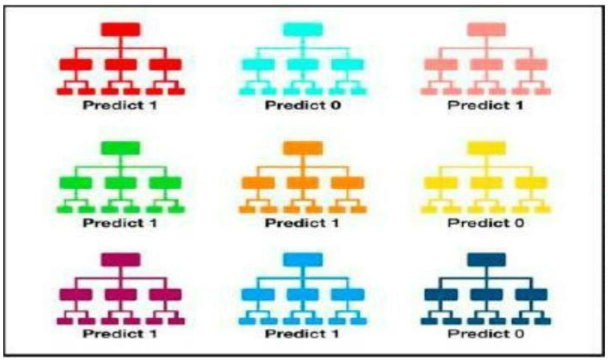


Fig.4 Random Forests

4. Methods

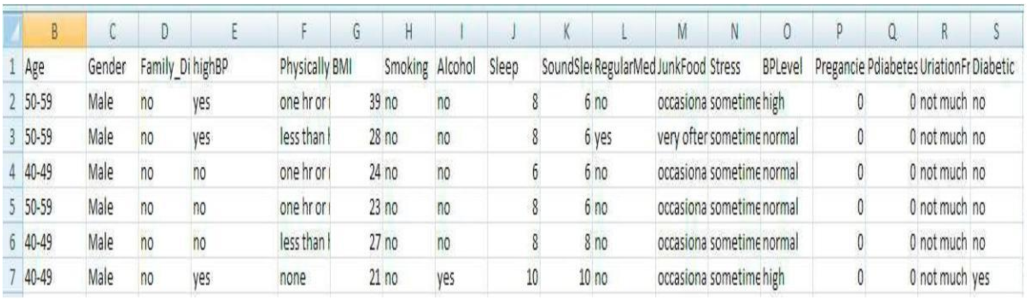
4.1 Data Description

A total of 950 people, including 376 women and 584 men, are chosen for this study who are 19 years old or older. A questionnaire that was self-prepared based on the factors that might cause diabetes was given to the participants and is provided in Table 1. The PIMA Indian

Diabetes database was used in the same studies to confirm the model's validity [28], as shown in Table1. A sample dataset obtained using a questionnaire is shown in Figure 6.

S. No.	Our Set Data		Pima Data Set	
	Parameters	**Instances**	**Parameters**	**Instances**
	Total Participants	**950**	**Total Participants**	**760**
1	Age	18 or above	Age	21 or above
2	Gender		Gender	
	Male	567	All Female	760
	Female	383		
3	famil history with diabetes	Yes/ No	Pregancies	Numeric
4	Diagnosed with high blood pressure	Yes/ No	Glucose	Plasma glucose concentration a 1.55 hours in an oral glucose tolerance test.
5	Walk/run/physically active	• None • Less than half an hour • More than half an hour • One hour or more	Blood pressure	Diastolic blood pressure (mm Hg)
6	BMI	Numeric	Skin thickness	Triceps skin fold thickness (mm)
7	Smoking	Yes/No	Insulin	2-Hour serum insulin (mu U/ml)
8	Alcohol consumption	Yes/No	BMI	Body mass index (weight in kg/(height in m))
9	Hours of sleep	Numeric	Diabetes pedigree function	Diabetes pedigree Function
10	Hours of sound sleep	Numeric	Outcome	Diabetic – 254 Non- diabetic - 515
11	Regular intake of medicine?	Yes/No		
12	Junk food consumption	Yes/No		
13	Stress	• Not at all • Sometimes • Often • Always		
14	Blood pressure level	High/normal/low		
15	Number of pregnancies	Numeric		
16	Gestation diabetes	Yes/No		
17	Frequency of urination	•Not much • Quite much		

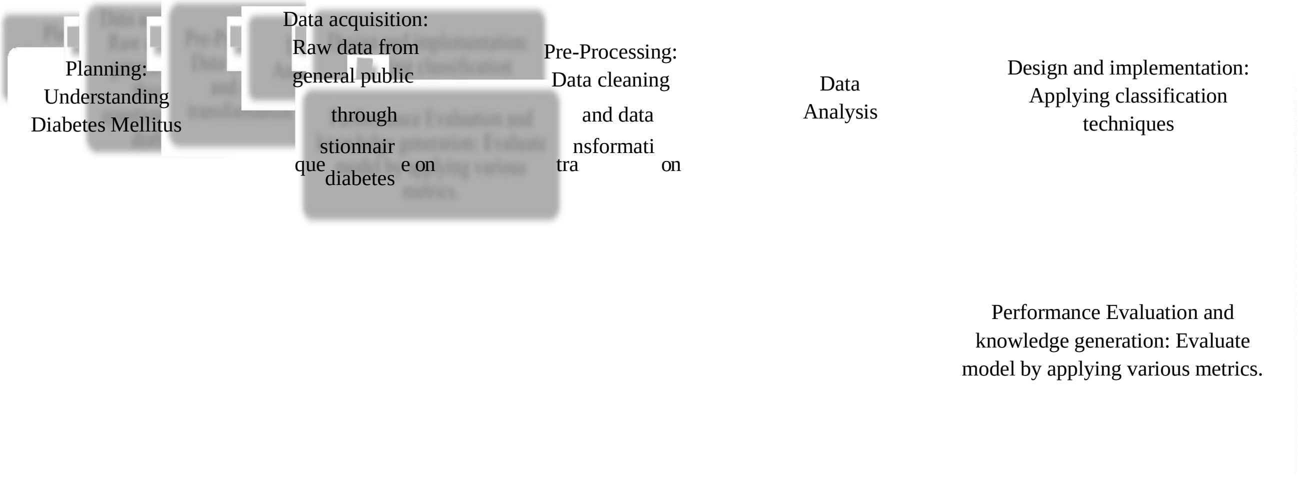
18	Diabetic?	• Diabetic - 383 •Non Diabetic - 567	



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Age	Gender	Family_Di	highBP	Physically	BMI	Smoking	Alcohol	Sleep	SoundSle	RegularMed	JunkFood	Stress	BPLevel	Pregancie	Pdiabetes	UriationFr	Diabetic
2	50-59	Male	no	yes	one hr or	39	no	no	8	6	no	occasiona	sometime	high	0	0	not much	no
3	50-59	Male	no	yes	less than	28	no	no	8	6	yes	very ofter	sometime	normal	0	0	not much	no
4	40-49	Male	no	no	one hr or	24	no	no	6	6	no	occasiona	sometime	normal	0	0	not much	no
5	50-59	Male	no	no	one hr or	23	no	no	8	6	no	occasiona	sometime	normal	0	0	not much	no
6	40-49	Male	no	no	less than	27	no	no	8	8	no	occasiona	sometime	normal	0	0	not much	no
7	40-49	Male	no	yes	none	21	no	yes	10	10	no	occasiona	sometime	high	0	0	not much	yes

Fig.5 Study of this Data set

R programming language with Visual studio were utilized for the report's implementation as well as coding, respectfully. Using the data collected and the Pima sample, algorithms for machine learning such as regression models, k-nearest neighborhood, the support vector machine, naive bays classifier, decision tree, and random forest classifications were applied in hopes of predicting diabetes. Then, all of these recommendations from each categorization are contrasted to each other. The stages to execute the machine learning algorithm are specified below and are seen in figure 5.



5. Discussion and Results

The figure 5 demonstrates the data set which used identify diabetes. Here, the response variable was the diabetes factor, while the other parameters were treated as independent factors. Only binary values can be used for the diabetic parameter, where 0 denotes non-diabetes and 1 denotes diabetes. The total sample is divided into a training set and a test set in the proportions of 75:25 in order to train the dataset. The training set was then used to predict the test set outcomes using all six classification approaches, including Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Decision Tree, and Random Forest, which produced the confusion matrix in Table 2.

Table : 2 Confusion matrix generated utilizing multiple methods.

	Logistic Regression	K Nearest Neighbour	Support Vector Machine	Naive Bayes	Decision Tree	Random Forest
Our dataset	0 1 0 157 14 1 21 46	0 1 0 147 27 1 31 36	0 1 0 157 16 1 17 50	0 1 0 144 27 1 19 48	0 1 0 156 15 1 23 44	0 1 0 167 4 1 10 57
Pima Dataset	0 1 0 107 18 1 31 36	0 1 0 104 21 1 35 32	0 1 0 107 18 1 31 36	0 1 0 105 20 1 23 44	0 1 0 93 32 1 26 41	0 1 0 105 20 1 28 39

The measure indicated in equation 3-9 may be derived through using confusion matrices which were collected. True Negative (TN), False Positive (FP), False Negative (FN), and True Positive were the conclusions of these matrices (TP). Due to the higher number of non-diabetic cases than diabetic ones in both datasets, the TN is greater than the TP. As a result, all strategies provide worthwhile outcomes. The following measurements have been determined using the following equations [25–27] in hopes of determining the precise precision of each method:

$$\text{퐴푐푐푢푟푎푐푦} \frac{\text{푇푃}+\text{푇푁}}{\text{푇푃}+\text{푇푁}+\text{퐹푁}+\text{퐹푃}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

$$\text{퐸푟푟표푟} = \frac{\text{퐹푃}+\text{퐹푁}}{\text{푇푃}+\text{푇푁}+\text{퐹푁}+\text{퐹푃}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

$$\text{Sensitivity} = \frac{\text{푇푃}}{\text{푇푃}+\text{퐹푁}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

$$\text{푆푝푒푐푖}\ \text{Specification} = \frac{\text{푇푁}}{\text{푇푁}+\text{퐹푃}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

$$\text{푃푟푒푐}\ \text{Precision} = \frac{\text{푇푃}}{\text{푇푃}+\text{퐹푃}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

$$\text{F} - \text{Measure} = \frac{2*(\text{Precision}*\text{Sensitivity})}{\text{Precision}*\text{Sensitivity}} \ldots\ldots\ldots\ldots\ldots\ldots(8)$$

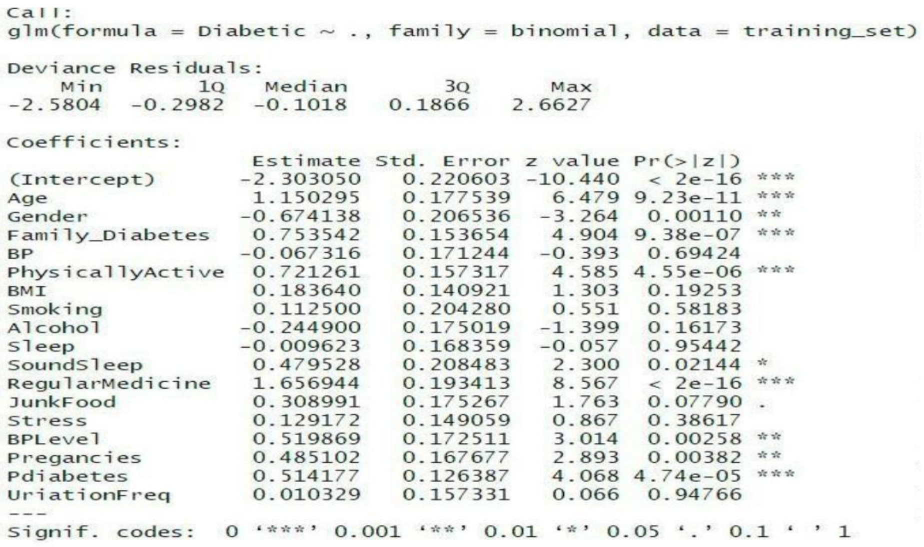
$$\text{푀퐶퐶} = \frac{(\text{푇푃}*\text{푇푁})-(\text{퐹푁}+\text{퐹푃})}{\sqrt{(\text{푇푃}+\text{퐹푃})(\text{푇푃}+\text{퐹푁})(\text{푇푁}+\text{퐹푃})(\text{푇푁}+\text{퐹푁})}} \ldots\ldots\ldots\ldots\ldots(9)$$

The accuracy rate, error rate, sensitivity, specificity, and precision of each model are all analyzed, and the outcomes are displayed in Table 3 along with the Matthew's correlation coefficient (MCC) as well as area under the curve (AUC).

Table.3 values for various metrics using multiple categorization methodologies.

	Logistic Regression		K Nearest Neighbour		Support Vector Machine		Naive Bayes		Decision Tree		Random Forest	
	Our dataset	Pima dataset	Our dataset	Pima dataset	Our dataset	Pima dataset	Our dataset	Pima dataset	Our dataset	Pima dataset	Our dataset	Pima dataset
Accuracy	0.857	0.744	0.773	0.708	0.865	0.744	0.806	0.689	0.840	0.697	**0.941**	0.750
Error	0.142	0.255	0.226	0.291	0.134	0.255	0.193	0.310	0.159	0.312	**0.058**	0.250
Sensitivity	0.882	0.775	0.826	0.748	0.901	0.775	0.883	0.820	0.871	0.781	**0.943**	0.789
Specificity	0.779	0.666	0.610	0.603	0.769	0.666	0.640	0.687	0.745	0.561	**0.934**	0.661
Precision	0.923	0.856	0.865	0.832	0.912	0.856	0.842	0.840	0.912	0.744	**0.976**	0.840
F-Measure	0.902	0.813	0.845	0.787	0.906	0.813	0.862	0.830	0.891	0.762	**0.959**	0.813
MCC	0.685	0.416	0.419	0.331	0.664	0.416	0.540	0.502	0.592	0.349	**0.852**	0.436
10-fold CV	0.890	0.770	0.830	0.742	0.885	0.770	0.849	0.756	0.854	0.749	**0.969**	0.774
Kappa	0.727	0.470	0.516	0.419	0.713	0.466	0.638	0.447	0.646	0.422	**0.922**	0.488
AUC	0.908	0.765	0.916	0.815	0.893	0.771	0.857	0.760	0.916	0.842	**1.000**	**1.000**

Another finding from table 3 is that, because our dataset has more fields that are pertinent to determining the risk of diabetes, all of the algorithms are more accurate on it than the PIMA database. For our dataset, the Random Forest classifier outperforms all others in terms of accuracy (at 90%), sensitivity, specificity, precision, and F-measure. Also, the random forest model's AUC score is 1, indicating that it performs outstanding categorization.



```
Call:
glm(formula = Diabetic ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5804  -0.2982  -0.1018   0.1866   2.6627

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.303050   0.220603 -10.440  < 2e-16 ***
Age               1.150295   0.177539   6.479 9.23e-11 ***
Gender           -0.674138   0.206536  -3.264  0.00110 **
Family_Diabetes   0.753542   0.153654   4.904 9.38e-07 ***
BP               -0.067316   0.171244  -0.393  0.69424
PhysicallyActive  0.721261   0.157317   4.585 4.55e-06 ***
BMI               0.183640   0.140921   1.303  0.19253
Smoking           0.112500   0.204280   0.551  0.58183
Alcohol          -0.244900   0.175019  -1.399  0.16173
Sleep            -0.009623   0.168359  -0.057  0.95442
SoundSleep        0.479528   0.208483   2.300  0.02144 *
RegularMedicine   1.656944   0.193413   8.567  < 2e-16 ***
JunkFood          0.308991   0.175267   1.763  0.07790 .
Stress            0.129172   0.149059   0.867  0.38617
BPLevel           0.519869   0.172511   3.014  0.00258 **
Pregancies        0.485102   0.167677   2.893  0.00382 **
Pdiabetes         0.514177   0.126387   4.068 4.74e-05 ***
UriationFreq      0.010329   0.157331   0.066  0.94766
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig.7. Variable Importance

The 10-fold cross validation method was also used to evaluate the efficacy of various models. A portion of the data is set away during the cross validation procedure, and the remaining data is used to train the model. Moreover, the procedure is repeated for a different data chunk. The sections are determined using the k-value. Data is separated into 10 sections since 10-fold cross validation was used in this instance. Cross validation for random forests has the highest accuracy. The random forest's kappa statistics are superb, exceeding.8

The significance of each parameter in the dataset is displayed in Figure 7. On the classifier model construction, a R function named "summary" is used to conduct this analysis. The star next to each parameter indicates the significance of the associated variable. The ratings are "***," "**," "*," and "." in that sequence. Where "***" denotes the greatest priority and "." the lowest priority. The variable without a rating is essentially irrelevant. To determine which parameter has a stronger influence on the forecast, variable importance is analyzed.

6. Conclusion

Detecting diabetes risk at an early stage is one of the worldwide health challenges. This study aims to develop a framework that predicts the risk associated with type 2 diabetes. Six machine learning classification techniques were used in this study, and the outcomes were evaluated using several statistical metrics. Tests were run on a dataset comprised of 18 diabetes-related questions that were compiled through online and offline surveys. On the PIMA database, the same methods were also used.

According to the experimental findings, Random Forest performed the best overall on our dataset, with an accuracy rate of 94.10%. With the PIMA dataset, random forest also provides the maximum accuracy. All of the models achieved good results for various parameters, such accuracy; recall sensitivity, etc., across the six distinct machine learning techniques used. Figure 7 shows the finding that among all factors, "Age," "Family diabetes," "Physically active," "Regular Medication," and "Diabetes" or gestational diabetes had the highest significance. These factors influence diabetes prediction more than the others.

This outcome can be applied in the future to forecast any other illness. This study has room for improvement, including the use of additional machine learning algorithms to forecast diabetes or any other disease.

References

1. http://diabetesindia.com/

2. Anjana, R. M., Pradeepa, R., Deepa, M., Datta, M., Sudha, V., Unnikrishnan, R., Bhansali, A., Joshi, S. R., Joshi, P. P., Yajnik, C. S., Dhandhania, V. K. (2011) "Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: Phase I results of the Indian Council of Medical Research–INdiaDIABetes (ICMR–INDIAB) study." Diabetologia **54 (12)**: 3022- 3027.

3. https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview

4. https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

5. Kaveeshwar, S. A., Cornwall, J. (2014) "The current state of diabetes mellitus in India." The Australasian medical journal **7(1)**: 45.

6. https://www.statisticssolutions.com/what-is-logistic-regression/

7. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

8. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

9. https://www.saedsayad.com/naive_bayesian.htm

10.https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

11. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I, Chouvarda, I. (2017) "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal **15**: 104-116.

12. Swapna, G., Vinayakumar R., Soman K. P. (2018) "Diabetes detection using deep learning algorithms." ICT Express **4 (4)**: 243-246.

13. Sisodia, D., Sisodia, D. S. (2018) "Prediction of diabetes using classification algorithms." Procedia computer science **132**: 1578-1585.

14. Wu, H., Yang S., Huang, Z., He, J., Wang, X. (2018) “Type 2 diabetes mellitus prediction model based on data mining.” Informatics in Medicine Unlocked **10**: 100-107.

15. Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q., Liu, Q. (2013) “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors.” The Kaohsiung journal of medical sciences **29 (2)**: 93-9.

16. Choubey, D.K., Paul, S. (2017) “GA_RBF NN: a classification system for diabetes.” International Journal of Biomedical Engineering and Technology **23 (1)**: 71-93.

17. Tigga, N. P. and Garg S. “Predicting type 2 Diabetes using Logistic Regression” accepted to publish in Lecture Notes of Electrical Engineering, Springer.

18. Huang, Y., McCullagh, P., Black, N., Harper, R. (2007) “Feature selection and classification model construction on type 2 diabetic patients’ data.” Artificial intelligence in medicine **41 (3)**: 251-262.

19. Eswari, T., Sampath, P., Lavanya, S. (2015) “Predictive methodology for diabetic data analysis in big data.” Procedia Computer Science **50**: 203-208.

20. Nai-arun, N., Moungmai, R. (2015) “Comparison of classifiers for the risk of diabetes prediction.” Procedia Computer Science. **69**: 132-142.

21. 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

22. Perveen, S., Shahbaz, M., Keshavjee, K., Guergachi, A. (2019) “Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques.” IEEE Access **7**: 1365-1375.

23. Rahman, R. M., Afroz, F. (2013) “Comparison of various classification techniques using different data mining tools for diabetes diagnosis.” Journal of Software Engineering and Applications **6 (03)**: 85.

24. Choi, B.G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., Noh, Y. K. (2019) "Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks." Yonsei medical journal **60 (2)**: 191-9.

2**5.**Käräjämäki, A.J., Bloigu, R., Kauma, H., Kesäniemi, Y. A., Koivurova, O. P., Perkiömäki, J., Huikuri, H., Ukkola, O. (2017) "Non- alcoholic fatty liver disease with and without metabolic syndrome: different long-term outcomes." Metabolism **66**: 55-63.

2**6.**Gurka, M.J., Golden, S. H., Musani, S. K., Sim, M., Vishnu, A., Guo, Y., Cardel, M., Pearson, T.A., DeBoer, M.D. (2017) "Independent associations between a metabolic syndrome severity score and future diabetes by sex and race: the Atherosclerosis Risk In Communities Study and Jackson Heart Study." Diabetologia **60 (7)**: 1261-1270.

27. Laakso, M., Kuusisto, J., Stančáková, A., Kuulasmaa, T., Pajukanta, P., Lusis, A. J., Collins, F. S., Mohlke, K. L., Boehnke, M. (2017) "The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases." Journal of lipid research **58 (3)**: 481-493.

28.https://www.kaggle.com/uciml/pima-indians-diabetes-database

2**9.**Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P. (2012) "An assessment of the effectiveness of a random forest classifier for land-cover classification." ISPRS Journal of Photogrammetry and Remote Sensing **67**: 93-104.