Name: Dhrumil Savalia Date of submission: 30/11/2021 Submitted to: Exposys Data labs

Data science

ABSTRACT

Diabetes is a disease given rise due to escalating level of blood glucose. Diabetes is a chronic disease with the possibility to cause a worldwide health care catastrophe. According to IDF(International Diabetes Federation) 382 million people are suffering from diabetes across the whole globe. By 2035, this will be doubled upto 592 million. Various conventional methods, established on physical and chemical tests, are available for identifying diabetes. However, early prediction of diabetes is quite demanding task for medical doctors due to complex relationships of numerous factors as diabetes influences human organs such as kidney, eye, heart, nerves, foot etc. Data science models are prospective to benefit other scientific fields by radiating new light on usual questions. One such job is to assist predictions on medical data. Machine learning is an emanating scientific field in data science tackling with the ways in which machines assimilate from experience. The aim of this project is to develop a system which can carry out early prediction of diabetes for a patient with a higher accuracy by the results of machine learning technique. This project aims to predict diabetes via supervised machine learning methods Logistic regression. This project also aims to propose an efficient technique for proper detection of the diabetes disease.

TABLE OF CONTENTS

Introduction

Diabetes mellitus is an boundless infection rendered by hyperglycemia. It might cause various problems. As per the developing desolation as of late, in 2040, the world's diabetic patients will attain 642 million, which implies that one of the ten grown-ups later on is suffering from diabetes. There is no unpredictability that this stupefying figure needs much consideration. World Health Organization has assessed 12 million deaths happen around the world, consistently due to Heart maladies. A significant portion of the deaths in the United States and other created nations are anticipated to cardio vascular maladies. The early analysis of cardiovascular sicknesses can help in settling on options on way of life changes in high hazard patients and thus decrease the complexities. This exploration means to focus on most notable hazard elements of cardiovascular illness just as expected the general hazard using calculated deterioration. Machine Learning has been associated to various parts of medicinal wellbeing. In this project, we have used Logistic regression to predict diabetes mellitus.

Figure 1: Predicting Diabetes Using Machine Learning

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The aim of the dataset is to anticipate whether or not a patient has diabetes, based on established measurements included in the dataset.

Many parameters were placed on the selection of these examples from a larger database.

The dataset composes of various medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age etc.

Proposed model is to predict diabetes that specialists can be valuable as a model to help envisage diabetes. In this investigation, examined the connection between problems in diabetic patients and their properties, for example, blood glucose, skin thickness, haemoglobin and blood pressure of the patients. The point of this examination is to anticipate confusions dependent on their demonstrations.

The below figure 2 is the diabetes indicator which gives the alert upon increase in the level of diabetes

Figure 2 : Representation of Diabetes Level Check

.

Existing system

The healthcare industry collects big amounts of healthcare data which, unluckily, are not “mined” to search concealed information. Sensible decisions are usually made based totally on doctor’s instinct instead of analyzing statistics concealed in the database. This exercise terminates in unwanted biases and errors. The existing process is very slow to give the result. It is very problematic to predict diabetes.

Proposed system

Diabetes prediction model is created by training the PIMA dataset to anticipate diabetes of a person. Diabetes prediction is done by inputting clinical information in the model. The set of rules will compute the opportunity of existence of diabetes. This thus saves time and money instead of conducting various tests. Format of statistics plays crucial role on this software. At the time of uploading the user information utility will take a look at its right record format and if it is no longer as consistent then ERROR dialog box may appear. Our device will implement the logistic regression algorithm. 80% of the entries in the statistics set can be used for training and the last 20% for testing the accuracy of the set of rules. Furthermore, a few steps can be taken for optimizing the algorithms thereby enhancing the accuracy, precision and rec.

Our proposed System has the following benefits:

- Powerful, flexible, and easy to use.
- Increased efficiency of doctor.
- Improved patient satisfaction.
- Reduce the use of papers.
- Simple and Quick.
- More accurate result.

Logistic Regression

Logistic regression is a statistical method for interpreting a dataset in which there are one or more independent variables that determine a result. The result is computed with a dichotomous/binary variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use categorical variables. In layman terms , it evaluates the probability of occurrence of an event by fitting data to a logit function.

Logistic regression was developed by statistician David Cox in 1958. This binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the probability of a given outcome by a specific percentage.
Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

It’s being used in Healthcare, Social Sciences & various ML for advanced research & analytics. Trauma & Injury Severity Score, which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression.

Binary logistic regression is estimated using Maximum Likelihood Estimation (MLE), unlike linear regression which uses the Ordinary Least Squares (OLS) approach. MLE is an iterative procedure, meaning that it starts with a guess as to the best weight for each predictor variable (that is, each coefficient in the model) and then adjusts these coefficients repeatedly until there is no additional improvement in the ability to predict the value of the outcome variable (either 0 or 1) for each case.

Life insurance actuaries use logistic regression to predict, based on given data on a policy holder (e.g. age, gender, results from a physical examination) the chances that the policy holder wilbefore the term of the policy expires.

Political campaigns try to predict the chances that a voter will vote for their candidate (or do something else desirable, such as donate to the campaign

Pregnancies	Glucose	BloodPres	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
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
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0

Figure 3: PIMA dataset

Steps for building model

Preprocess Data

Data Preprocessing is a method that is used to convert the raw data into a clean data set. In other words, whenever the data is collected from various sources it is collected in raw format which is not appropriate for the analysis.

Therefore, certain steps are performed to transform the data into a small clean data set. This method is performed before the implementation of Iterative Analysis. The set of steps is known as Data Preprocessing. It includes

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Data Preprocessing is obligatory due to the presence of unformatted real-world data. Mostly real-world data is composed of –

Inaccurate data (missing data) - There are many causes for missing data such as data is not appropriately collected, a mistake in data entry, technical issues with biometrics and much more.

The presence of noisy data (erroneous data and outliers) - The reasons for the existence of noisy data could be a technological difficulty of gadget that gathers data, a human error during data entry and much more.

Inconsistent data - The presence of inconsistencies are because of the existence of data duplication, human data entry, containing mistakes in codes or names, i.e., contravention of data constraints etc.

Here we have not put much effort on data preprocessing as the data was cleansed already.

Train and Test Data Creation

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

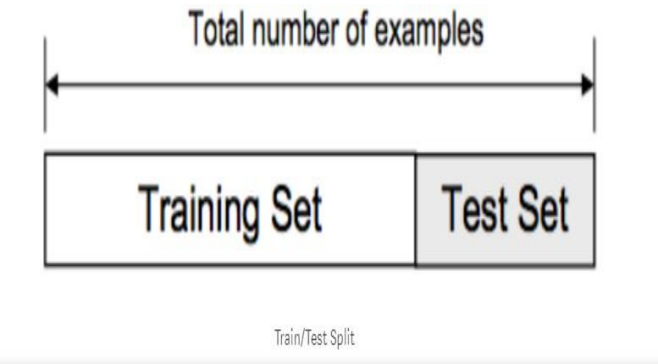

Figure 4: Training and test set

The above figure shows splitting of data into training set and test set. The size of training set will be larger than that of test set. The training set will be trained tested against the test set.

Data standardization -It is very important to standardize data to bring it in a given range. As the values in the dataset corresponding to different parameters may vary and hence are not in a common range which urges the need to standardize the data in the common range. This is done by defining a scaler function and then simultaneously fitting and transforming the data.
This is done before splitting training and testing data.

Hyperparameter tuning: This is done to select an appropriate model for the data. As in our case logistic regression is the best method for binary classification as a part of supervised learning model. We can use Grid search CV method to identify which method is suitable in our case, with respect to the data set we compare SVM (Support vector machine) & logistic regression using 5 cross fold validation. According to this grid search cv suggested logistic regression for better accuracy.

Model Creation

- The procedure of training an ML model includes providing an ML algorithm with training data to learn from. The term ML model refers to the model that is created by the training process.
- The training data must contain a target or target attribute. The learning algorithm searches patterns in the training data that performs input data mapping and sticks to the target (the answer that you want to predict), and it outputs an ML model that seizes these patterns.
- ML model can be used to get predictions on new data i.e. suppose testing data for which you do not know the target.
- In our project we are using Logistic Regression for building our Model on Pima Indian Dataset.

Result Analysis*:*

In this final phase, we will test our model on our prepared dataset and also measure the Diabetes prediction performance on our dataset. To evaluate the performance of our created classification and make it comparable to current approaches, we use Accuracy to measure the effectiveness of classifiers.

.

Technical Requirements of the System

Hardware Requirements

- **System Processor** : Core i3.
- **Hard Disk** : 500 GB.
- **Ram** : 4 GB.

✓ Any desktop / Laptop system with above configuration or higher level.

Software Requirements

- **Operating system** : Windows 8 / 10
- **Programming Language** : Python
- **DL Libraries** : Numpy, Pandas, Sci kit learn

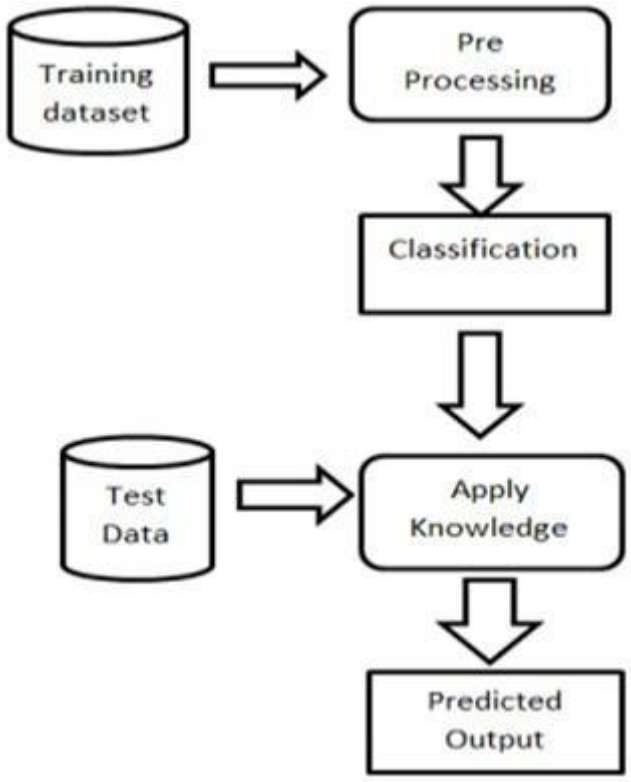

Fig 5: Architecture of Diabetes Prediction System

System Architecture

The above figure shows the architecture of Diabetes Prediction System. The training data set is fed to the system as input which will be initially pre processed. Data pre processing is the phase where the raw data will be transformed into meaningful and understandable format.

The pre - processed data will be on behind classified using the best classification model. Then the classified model will be compared with the test data in order to classify it accurately using some distance measures. The final classified data will be converted to data patterns using intelligent methods. The gained patterns will be examined for accuracy and precision. The recognized patterns will be presented as knowledge in the required form as output.

DFD (Data flow diagram) for Classification of Data

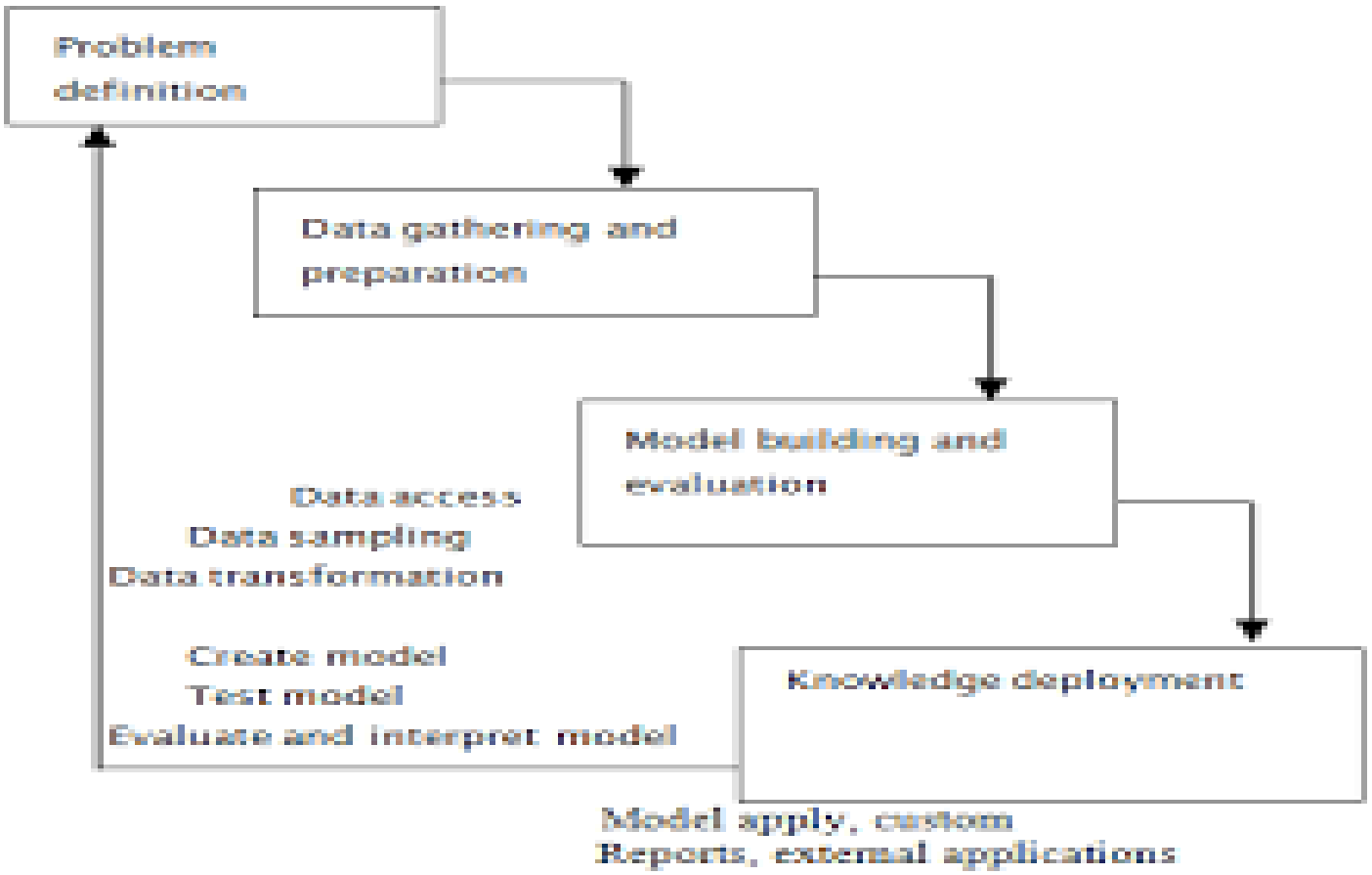

Fig 6: DFD for Classification

The above figure 4.5 shows the DFD for Classification of Data of the proposed system. For the problem definition data will be collected and pre- processed. Model will be generated for the acquired data and data access, data sampling, data transformation will be done. After generating the model compute and analyze it. Generated model is applied for various applications.

Values for each attribute is assigned based on significance for the input dataset. Process and generate value. If the value is equal to one then the patient is diabetic otherwise the patient is non diabetic.

Implementation

Analysis:

In this final stage, prepared classification model is tested on our prepared image dataset and also measure the performance on our dataset (X_test).

After model building, knowing the power of model prediction on a new example, is very crucial issue. Once a predictive model is restored using the historical data, one would be inquisitive as to how the model will provide result on the data excluding from the given One might even try different model types for the same prediction question, and then, choose the better one by comparing their accuracies. To evaluate the performance of a predictor, there are commonly used performance metrics, such as accuracy, recall, precision etc. All the performance parameters are based on one matrix i.e. confusion matrix, which estimates the count of the model of its right and wrong prediction. Below figure shows a confusion matrix for a two-class classification problem.

Confusion matrix:

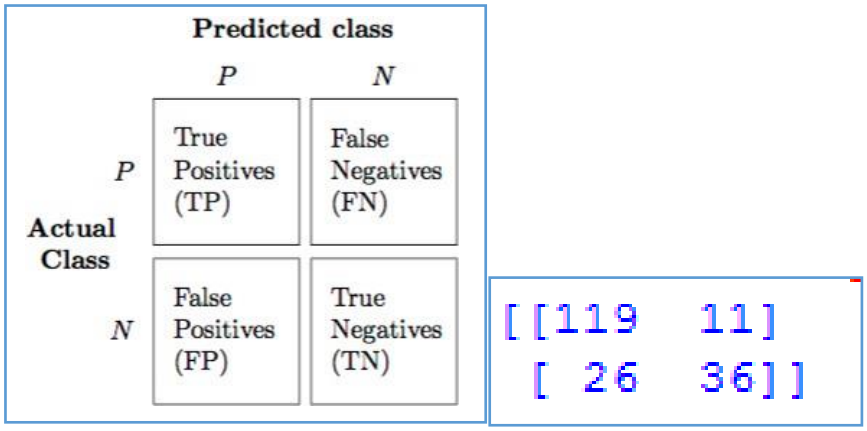

Figure 7: Confusion matrix

The above figure 5.1 shows the Confusion matrix for the proposed system. Precision is described as the ratio of true positive and sum of true positive and false positive, whereas recall describes the ratio of true positives and sum of true positive and false negative. F1 score is the harmonic mean of the precision and recall scores. For a model to be precise and accurate it requires to have better precision and recall, indirectly better F1 score.

Experimental Results

Model performance

Data distribution:

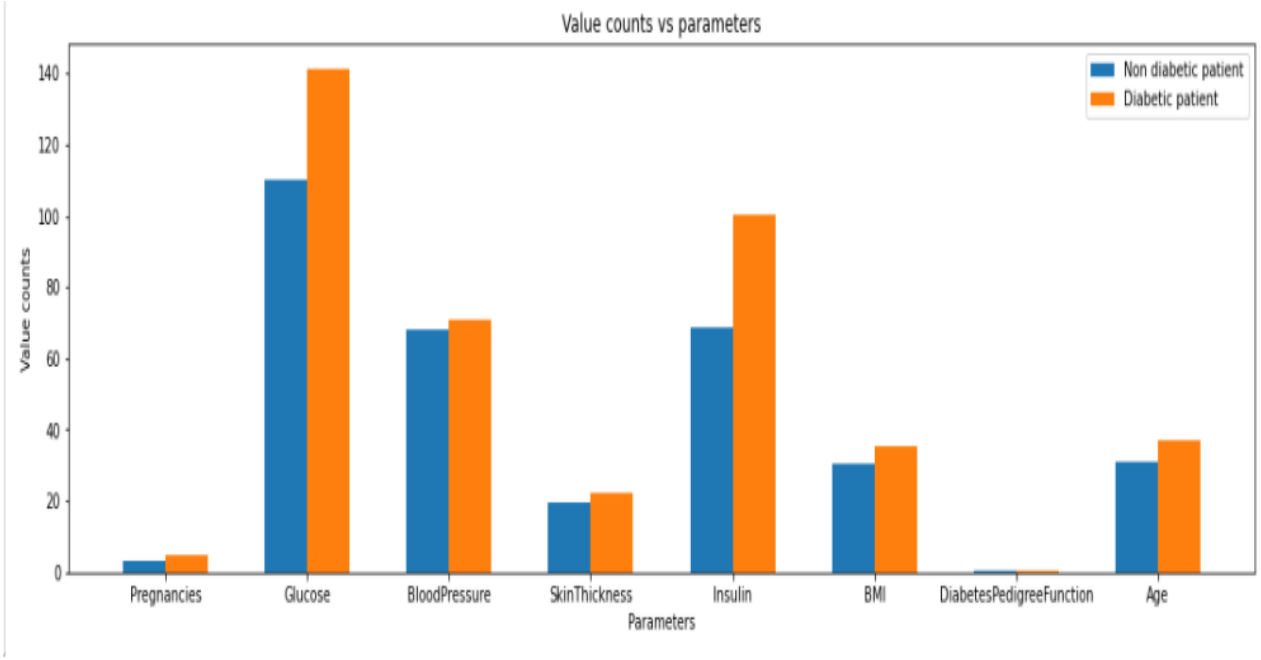

Figure 8: Data distribution

```
array([[356,  44],
       [ 89, 125]], dtype=int64)
```

Above figure represents the obtained confusion matrix

Accuracy score=76%

Precision score=74%

Recall score=58.4%

Conclusion and Future Enhancement

Conclusion

Diabetes is an important health inconvenience in human society. This paper has summed up techniques for prediction of this sickness. Studying deep we were able to identify the influence of machine learning techniques in the prediction of chronic diseases for immediate clinical diagnosis. It continues to be a bright light in the field of medical treatment and diagnosis. Some techniques of deep studying has been discussed which may be applied for Diabetes prediction, alongside spearhead machine learning algorithms. An inquisitive assessment has been completed for obtaining best available algorithm for clinical dataset. In future our aim is to carry ahead the work of varying scientific dataset, wherein dataset varies with time

Future Enhancement

The proposed diabetes prediction model can be used for prediction of various other diseases depending on the complexity of the data, and appropriate algorithms can be sued by proper analysis.

References

- https://www.academia.edu
- https://en.wikipedia.org/wiki/Clinical_data_management
- http://shodh.inflibnet.ac.in:8080/jspui/bitstream/123456789/4170/3/03_literature%20review.pdf
- World Health Organization. Available online: http://www.who.int (accessed on 14 September 2018).
- Baldwin, D. Wayfinding technology: A road map to the future. J. Vis. Impair. Blind. 2003, 97, 612–620.
- 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.
- Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics, 97, 120-127.
- Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017, February). Risk prediction of type II diabetes based on random forest model. In Advances in Electrical, Electronics, Information, Communication and BioInformatics (AEEICB), 2017 Third International Conference on (pp. 382-386). IEEE.
- Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbour regression. Neurocomputing, 251, 26-34.
- Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on (pp. 1006-1010). IEEE.
- Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez- Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2017). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. Expert Systems with Applications, 72, 335-343

- Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 1589). IEEE.
- Pradeep, K. R., & Naveen, N. C. (2016, December). Predictive analysis of diabetes using J48 algorithm of classification techniques. In Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on (pp. 347-352). IEEE.
- Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115-121.
- Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 47, 76-83.
- Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.
- Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An Efficient RuleBased Classification of Diabetes Using ID3, C4. 5, & CART Ensembles. In Frontiers of Information Technology (FIT), 2014 12th International Conference on (pp. 226-231). IEEE.
- 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), 93-99.
- Krati Saxena, D., Khan, Z., & Singh, S.(2014) Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm.
- Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43).
- Guo, Y., Bai, G., & Hu, Y. (2012, December). Using bayes network for prediction of type-2 diabetes. In Internet Technology And Secured Transactions, 2012 International Conference for (pp. 471-472). IEEE.