Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Diabetes disease prediction using significant attribute selection and classification approach

To cite this article: P Tiwari and V Singh 2021 *J. Phys.: Conf. Ser.* **1714** 012013

View the article online for updates and enhancements.

Diabetes disease prediction using significant attribute selection and classification approach

P Tiwari[1] **and V Singh**[2]

[1]Computer Science Department, Rewa Institute of Technology, Rewa, India,
priyanka8634@gmail.com
[2]Computer Science Department, Rewa Institute of Technology, Rewa, India,
singhvarun1@gmail.com

Abstract. Data Mining performs a major role in healthcare services because disease recognition and investigation contains a vast amount of data. These conditions generate several data managing problems, and to operate efficiently. The healthcare datasets are undefined and influential and it is extremely monotonous to manage and to operate. To get better of the exceeding problems, numerous analyses present various ML algorithms for different disease examination and prediction. The undertaking of disease identification and prediction is an element of classification and forecasting. In this paper, diabetes is estimated by major characteristics and the relation of contradictory characteristics is also categorized. Significant features selection was done via the recursive feature elimination with random forest. The estimation of our system specifies a powerful alliance of diabetes with (BMI) and with glucose level was drawing out using the Apriori approach. XGBoost has examined for the estimation of diabetes. The XGBoost gives better accuracy of 78.91% compared to the ANN approach and might help support medicinal professionals through treatment decisions.

1. Introduction

Data mining (DM) is the procedure of withdrawing important information from a vast dataset and it performs a major task in healthcare areas. DM approaches intend to exist an important benefit for diabetes examiners because this reveals concealed facts from a large amount of diabetes allied information [1]. DM is the approach of rectifying helpful data and designs from a huge dataset. Mainly in healthcare, the size of the dataset is large and effective and difficult to forecast depends on figures. With the factors of medical areas, several DM approaches were introduced. The trendy and well-accepted approaches in the area of medicine are statistics-based analysis, forecasting & classification. The huge number of healthcare data composed of different origins; it is steadily significant that the approach is required to be urbanized with powerful attributes for study, explanation, and judgment procedure. This procedure referred to the ordinary withdrawn of secreted and potentially helpful data in a database [2].

Diabetes is a syndrome [3] which happens whenever the insulin-making in the body is insufficient or is incapable to employ the formed insulin appropriately, thus, this brings about high glucose in the blood. The cells of the body collapse the food in glucose and this glucose desires to be related to all the cells of the body. The insulin is the hormone that conducts the glucose formed by collapsing food in body cells. Some changes in the creation of insulin promote a boost in blood sugar and this prompts to injure the tissues and fall of organs. Normally a person considers being suffered from diabetes when the sugar level is exceeded to normal. There were three types. Type 1, Type 2, and Gestational. Type 1 and type 2 diabetes is a general type of disease, and other gestational diabetes, which happens in pregnancy, in addition to other forms.

Diabetes is a class of metabolic diseases distinguished by hyperglycemia results after deficiency in both insulin secretions, insulin action. The chronic hyperglycemia of diabetes is related to lasting injure, dysfunction, and breakdown of several organs [4]. The consequence of diabetes is exclusive in mothers' case because the disease is transferred to their unborn children. Diabetes is the main reason for blindness, amputations, heart failure and stroke are a few of the difficulties arouses with this disease [5].

Diabetes prediction is measured as a demanding issue for significant research. Many criteria like fructose amine, WBC count, fibrinogen, and hematological index were revealed to be unproductive because of few restrictions. Many estimation analyses use these criteria for the diagnosis of diabetes [6]. The prognostic analysis uses ML approaches to take a calculation of indefinite results [7]. With the prognostic analysis in diabetes, diagnosis, prediction, self-management & prevention can attain.

The remaining paper is formed as follows: Section II introduces the literature survey of diabetes estimation. Section III discusses a detailed method. Experiments and outcomes were presented in section IV while the conclusion of the paper with a future direction was given in section V.

2. Literature survey

Data mining has been engaged as a necessary task in the t healthcare approaches stated in [8] [9][10]. The relations of disease and source of disease & the consequences of signs are spontaneously measured in patients and assessed by users through DM approaches. The healthcare schemes employ huge datasets as input to the scheme to discover relationships among attributes. The relation outcomes were not assessed sufficiently in previous works. This explores the relation of secreted facts positioned amongst the huge medicinal databases and has been exploring related attributes by ruling frequent items via candidate generation. Knowledge of risk components linked by healthcare experts to recognize patients at elevated risk. A statistical analysis and DM approach [11] assist healthcare experts in the recognition of heart disease. The alike examination has recognized the disease of blood vessels and heart by statistical ethics and includes stroke, heart diseases, and peripheral artery disease, etc.

[12] Introduced an efficient method for estimation of heart disease from the dataset through the clustering method. Originally the dataset is grouped via the K-means and this with draws attributes and related information of heart disease. This permits the dataset to partitioned into the k segment. The introduced system excavates the recurrent patterns consequently with mined data relates to heart disease.

The author presents a maximum frequent Items using MAFIA which is the ML approach accomplished with the chosen model [13]. This mainly approximates heart disease and risk factors. Moreover, little practice 14] determines the estimation of accurateness slanting issues. This method uses the ID3 algorithm [15] for the procedure of training and uses in a huge quantity of disease datasets. The outcomes of analysis determine that the intended estimation approach is accomplished in estimating the heart disease efficiently. But estimation of diabetes is somewhat dissimilar with others.

An analysis of the estimation of heart disease risk using Bayes algorithms [16] conceded out in the literature. This utilizes the essential DM classification approach with some necessary attributes and many occurrences; this method is efficient because of the bagging method, which is a repeated procedure.

With the outcomes of [17], the bagging method is precise and competent than the J48 decision tree & Bayesian classification methods for the prediction of heart disease. The estimated model utilized in various medical datasets, and also utilized in the diabetes data, the scores of risk have been designed to predict diabetes risk.

The requirement of diabetic index renowned [18], this conducts a review concerning the risk of diabetes. They discover that the majority indexes were preservative & no one indices had taken relations amongst the risky factors.

In [19] utilized association rule mining to analytically discover relations of various characteristics and features linked to disease. The produced rules do not set up diabetes index while the analysis doesn’t assign an exacting result of concern and doesn't estimate the risk in the database, but they found few principle relations among diagnosis codes.

In [20] utilized FP-Growth and Apriori methods for diabetes prediction. But forecasting rule experiences large repetitions. Few researchers utilized splitting and merging approaches with quantitative association rules for the diabetic. Only some researches succeeded in this field. Random forest techniques were utilized in medicinal approximation. This is a tree-based approach and is efficient for large databases.

The introduced procedure [21] is an arrangement of the tree predictors hence every tree based on the value of a randomly sampled vector & by the equal dispersal for all trees in the forest. The simplification fault of a tree forest classifiers based on the power of distinct forest trees & relation among them. Random forests, an operative method in estimation & classification procedures. The main benefit of the approach is it decreases the overfitting issue. The accurate type of injecting randomness makes precise classifier & regression. Briefly, the ordinary thinking was the forests may not strive with an arcing algorithm inaccuracy. The results disperse this but precede various issues. Thus, the variations per boosting and arcing algorithms can decrease bias and variance.

By selecting random features for the segment, individual node produces error rates compared to Adaboost [22] but this is more vigorous regarding noise. In the diagnosis procedure, ML methods are employed. There are wonderful examine on health was developed by DM methods, analyze the methods & issues. Lee & Wang introduced a fuzzy expert system (FES) for diabetes decision support applications [23] where a fuzzy ontology of five-layer involves various fuzzy layers to define the knowledge with unreliability. They established semantic decision-making procedures in diagnosis. However, the methods are dominant and have definite restrictions. The fuzzification method only applies in a fuzzy expert system is quite a more significant somewhat ontology model. In disease diagnosis, the method suffers from accuracy.

Wenxin Zhu and Ping Zhong [26] examined the significance & combination of the secret data in SVM, which is an SVM extension. The paper used the benefits provided by the SVM+ also emphasis the over-classification problem of a single class. Since the samples have somewhat restricted & inadequate, so for this, the hidden data recognized from unknown data. The researchers verify classification efficiency by inserting the extra info to the conforming optimization problem & their result in an m-SVM style SVM+ context for single-class classification. The novel approach gives improved performance. Nevertheless, there are a few disadvantages that are more compound & harder in comparison to singe class m-SVM. This requires extra tuning of parameters with deep analysis mainly. Comparable earlier approaches, this cannot be done by statistical analysis. The attribute selection method wants more awareness.

Giveki, Davar, et al [27] proposed a new standard mechanism depends on Modified Cuckoo Search (MCS) and Feature Weighted Support Vector Machines (FW-SVMs). In this paper, the model proposed comprises three phases: Originally, to select an optimal feature of the set of features PCA is used. Formerly, to paradigm the FW-SVM via weighting diverse features depending upon their weight of degree the Mutual Information is organized. Finally, to select parameter has a major role in the accurate classification of the SVMs; the MCS applies to choose parameter values. By this method, the author attained accuracy i.e. 93.58% on the UCI dataset. The proposed system was suitable for all restricted features, wherever the process considers four features.

3. Proposed methodology

The artificial neural network remained not efficiently capable to predict diabetes in the dataset in provisions of accurateness and properly and incorrectly categorized instances. So, to overcome of this problem proposed an approach which is discussed in this section: First exhibits the collection of diabetic data from standard UCI repositories, followed by pre-processing step; feature selection for XGboost classifier; afterward, we describe Hyperparameter selection; finally, we describe the functioning procedure of XGboost classifier for classification of data. Fig. 1 demonstrates the full flow of methodology.

3.1. Preprocessing of data

Data is in raw form thus we performed some pre-processing for the removal of missing values, noisy and incompatible data. This is compulsory to do pre-processing for achieving good quality outcomes. For pre-data processing, it is useful to clean, integrate, transform, reduce, and discretize data [28].

3.1.1. Data cleaning. Data cleaning contains separating noisy data and quilting missing values. Noisy data consists of outliers that are separated to determine inconsistency [29]. Our dataset includes several values of zero (0) in glucose, blood pressure, skin thickness, insulin & BMI. Each ‘0’ value has then been replaced with the average value for all attributes.

3.1.2. Data reduction. Data reduction acquires condensed demonstration of database which is lesser in quantity however generates the almost same outcomes. Dimensionality reduction is utilized to lessen the numeral of attributes [30]. The Recursive Feature Elimination with the Random Forest approach is used to withdraw important attributes from the whole database. Glucose, BMI, diastolic blood pressure, age & insulin are important attributes.

3.1.3. Data transformation. The transformation of data involves information smoothing, uniformity, and aggregation [31]. Here, a standard scalar is used for scaling data. For data smoothing, the binning approach is utilized. The age attributes were useful in classifying in five categories, demonstrated in fig. 11 (a) Blood glucose deliberation is different from non-diabetic patients other than diabetic patients. Glucose measures are separated into 5 parts [32] demonstrated in fig. 11 (b). The pervasiveness of diabetes & obesity rising simultaneously globally. Moreover, the earlier analysis demonstrated that BMI is a significant risk factor for type-2 diabetes [33]. BMI standards comprise five classes as depicted in fig. 11 (c). A strong association was initiated among healthy and diabetic patients concerning their blood sugar levels [34]. BP separated into five various classes as depicted in fig. 11 (d).

After data cleaning, essential attributes are achieved by pre-processing, gathering, and transforming into bins.

3.2. Hyperparameters selection

Parameters that describe the representation architecture is known as hyperparameters and so the procedure of penetrating for perfect model architecture is known as hyperparameter tuning. These approaches relay how to illustrate probable model structure from the space of probable hyperparameter values. In our proposed, Randomized Search CV is used for parameter tuning of XGboost.

3.3. Association rule mining

Association rule mining is a significant aspect of DM. Association rule mining [35] methods are extensively utilized in finding secret correlation and relation among items set in a transaction. Mining association rules from a huge number of data in repositories is concerned for various industries that assist in several directive procedures. A kind of mining defines a scrupulous local pattern that would be simply estimated and interfaced. The two significant basic events of association rules are support and confidence. When two items are thereafter that support is described as the fraction of incidence of 2-items & sums of all transactions & prospect of considering rules resultant below situation that transaction to enclose precursor known as confidence.

3.3.1. Apriori. Apriori [36] is a mining approach that when given the input of a transactional database it mines all frequently occurring items in the transaction. In this, when input the PID dataset to Apriori it then produces a set of risk factors that arise recurrently and signifies those to be measured for mounting diabetes. Apriori works in the standards of support and confidence. The approach recurrently develops candidate itemsets for every item. Apriori is proficient in developing a huge set of risk factors and these can be used in prediction.

3.4. Classification

Classification is a significant task in DM. The scheme following this organizes the known data records into particular probable cases that were identified previously. The task of Classification tasks formulates the use of any single approach. Classification [37] requires estimating a definite result depends on a known input. Orderly to estimate the outcomes, this wants to obtain the previously available data. Depending on this information records are categorized Classification is separating the given data conforming to their alike attributes to one another. In this paper, classification is done by XGBoost Classifier.

XGBoost [38] is the execution method of gradient boosting models which give elevated exultation in the presentation and speed to the model. It is practicable with adjustable criteria. XGBoost is developed below the Gradient Boosting structure, which is planned to be extremely proficient, supple, and transferable. The boosting scheme is to merge the sequence of fragile classifiers by less accurateness to construct a tough classifier through enhanced classification execution. When the weak learner at every phase depends upon the loss function in gradient direction then this will know as Gradient Boosting Machines [39].

3.5. Proposed algorithm

Input: The Pima Indian Diabetes (PID) dataset
Output: Classification accuracy.

Strategy:

```
    Step 1    Start
    Step 2    Collect the (PID) dataset from the UCI ML repository.
    Step 3    Pre-processed the collected dataset.
    Step 4    To clean the dataset every zero value was restored with the average value of that
              aspect.
    Step 5    The Recursive Feature Elimination with Random Forest method is utilized to take
              out important features from a whole dataset in data reduction
    Step 6    Scale the data by a standard scalar
    Step 7    Smoothing of data using the binning method and categorized all features in bins
    Step 8    Apply the Apriori algorithm to obtain frequent itemset and finally obtain association
              rules
    Step 9    Randomized Search CV for parameter tuning of XGboost
    Step 10   Use XGboost classifier for classification of a dataset
    Step 11   Obtain confusion matrix and various performance measures
    Step 12   End
```

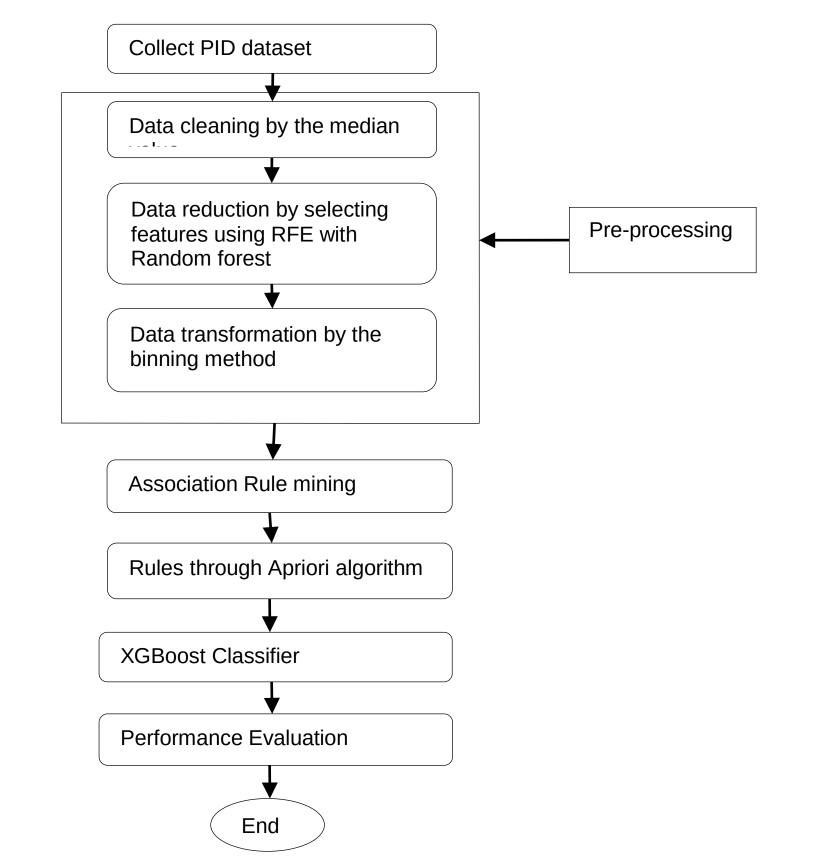

Collect PID dataset
Data cleaning by the median value
Data reduction by selecting features using RFE with Random forest
Data transformation by the binning method
Pre-processing
Association Rule mining
Rules through Apriori algorithm
XGBoost Classifier
Performance Evaluation
End

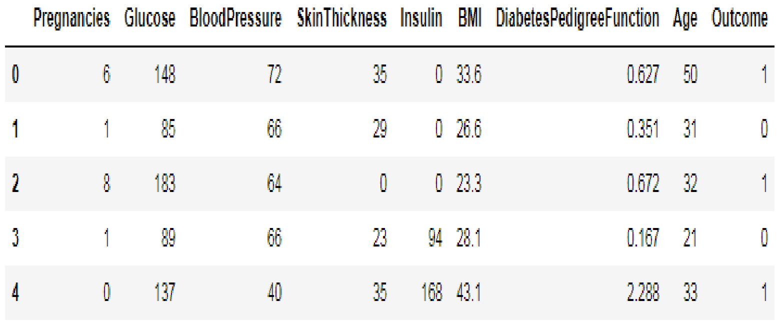
Figure 1. The workflow of proposed work.

4. Result analysis

In the result analysis, the experiment of the proposed work was performed by using the Jupyter Notebook of Python programming.

4.1. Dataset

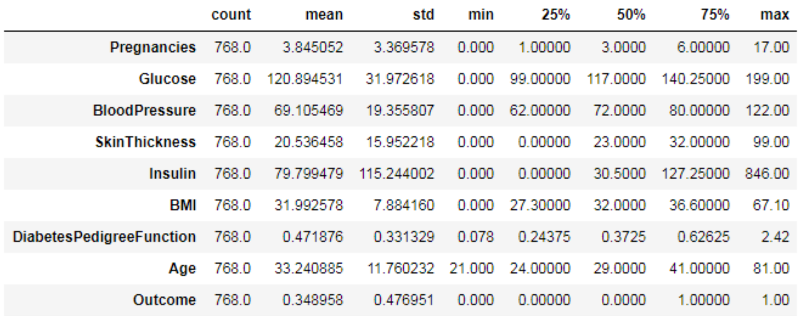
The dataset utilized is formerly occupied from the (NIDDK) [40]. The purpose of employing this dataset was to forecast throughout the analysis in case a patient has diabetes, depend on definite analytical dimensions incorporated in the dataset. Some restrictions were adjoined through the choice of the phenomenon from the greater dataset. The data type is a standard supervised binary classification. The (PID) database has: 9 = 8 + 1 attributes, 768 records relating female patients with 500 negative (65.1%) and 268 positive instances (34.9%)).



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 2. Dataset information.

4.2. Result analysis



	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Figure 3. Mean & standard deviation of dataset.

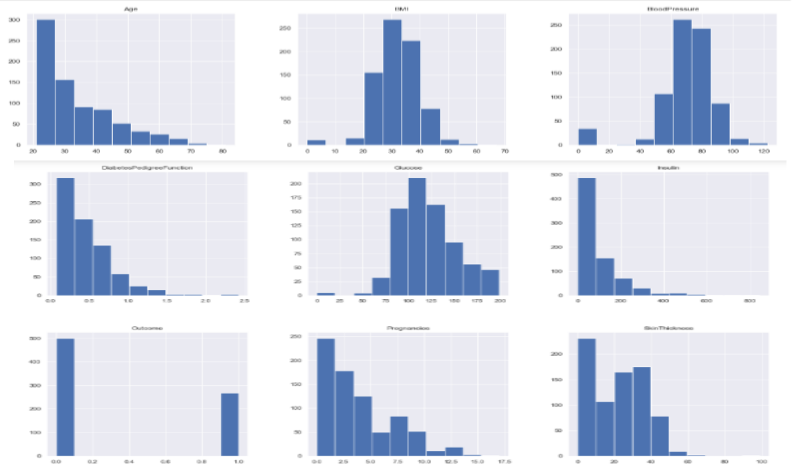


Figure 4. Before cleaning the histogram.

The dataset visualization before preprocessed is shown in Figure 4.

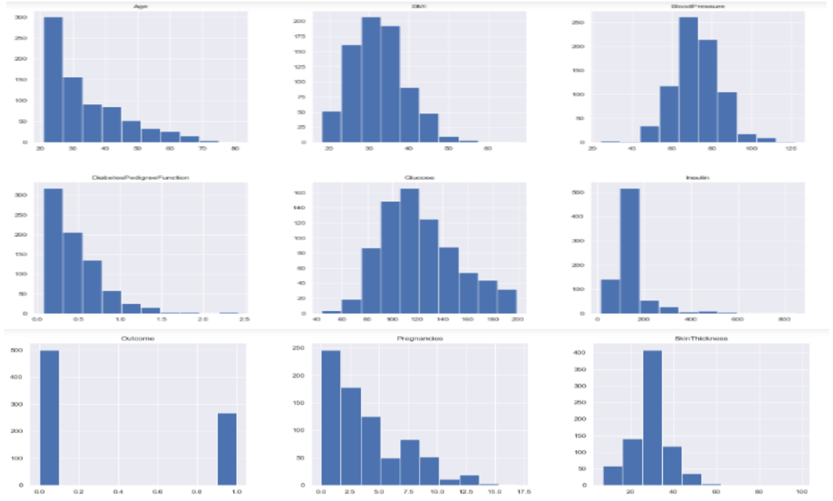
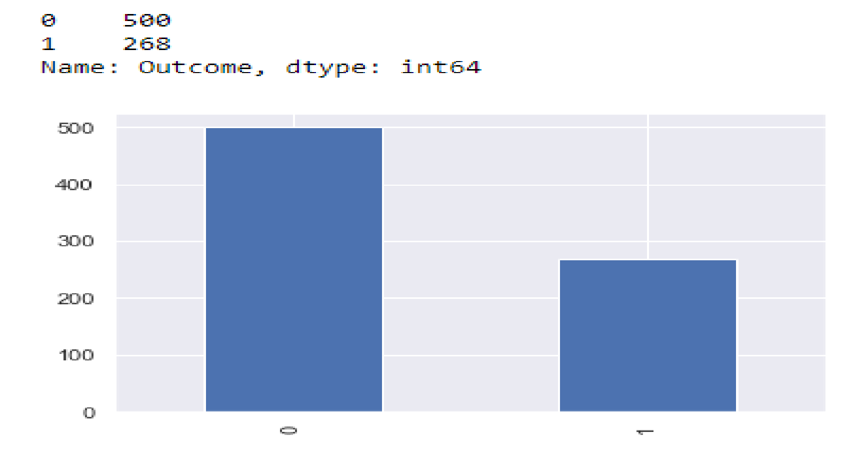


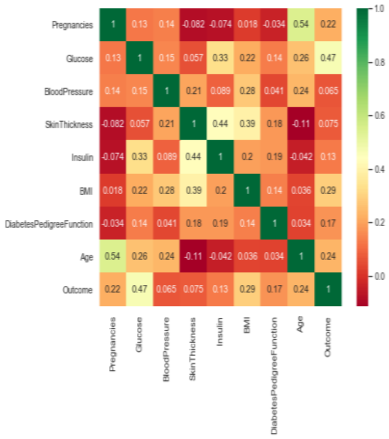
Figure 5. After cleaning the histogram.

The preprocessed dataset visualization is shown in Figure 5.



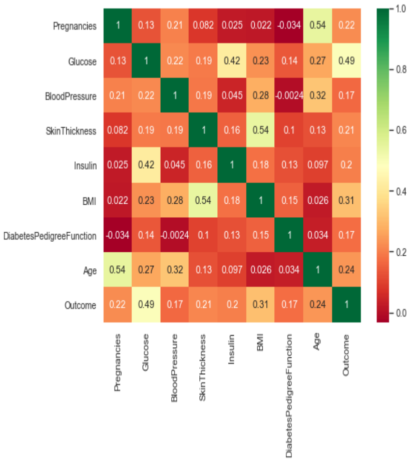
```
0     500
1     268
Name: Outcome, dtype: int64
```

Figure 6. Count of the outcome.



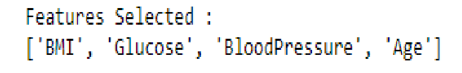
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.14	-0.082	-0.074	0.018	-0.034	0.54	0.22
Glucose	0.13	1	0.15	0.057	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1	0.21	0.089	0.28	0.041	0.24	0.065
SkinThickness	-0.082	0.057	0.21	1	0.44	0.39	0.18	-0.11	0.075
Insulin	-0.074	0.33	0.089	0.44	1	0.2	0.19	-0.042	0.13
BMI	0.018	0.22	0.28	0.39	0.2	1	0.14	0.036	0.29
DiabetesPedigreeFunction	-0.034	0.14	0.041	0.18	0.19	0.14	1	0.034	0.17
Age	0.54	0.26	0.24	-0.11	-0.042	0.036	0.034	1	0.24
Outcome	0.22	0.47	0.065	0.075	0.13	0.29	0.17	0.24	1

Figure 7. Heat map of unclean data.



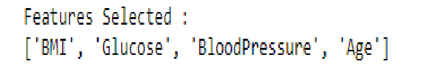
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.21	0.082	0.025	0.022	-0.034	0.54	0.22
Glucose	0.13	1	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1	0.19	0.045	0.28	-0.0024	0.32	0.17
SkinThickness	0.082	0.19	0.19	1	0.16	0.54	0.1	0.13	0.21
Insulin	0.025	0.42	0.045	0.16	1	0.18	0.13	0.097	0.2
BMI	0.022	0.23	0.28	0.54	0.18	1	0.15	0.026	0.31
DiabetesPedigreeFunction	-0.034	0.14	-0.0024	0.1	0.13	0.15	1	0.034	0.17
Age	0.54	0.27	0.32	0.13	0.097	0.026	0.034	1	0.24
Outcome	0.22	0.49	0.17	0.21	0.2	0.31	0.17	0.24	1

Figure 8. Heat map of clean data.



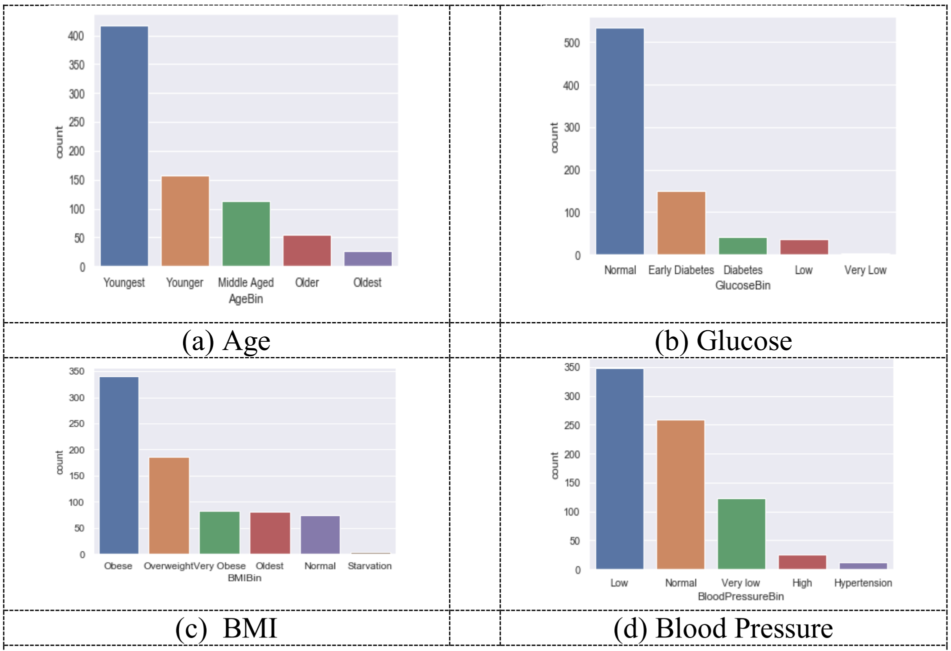
```
Features Selected :
['BMI', 'Glucose', 'BloodPressure', 'Age']
```

Figure 9. Feature selected for binning.



```
Features Selected :
['BMI', 'Glucose', 'BloodPressure', 'Age']
```

Figure 10. Features after binning.



(a) Age

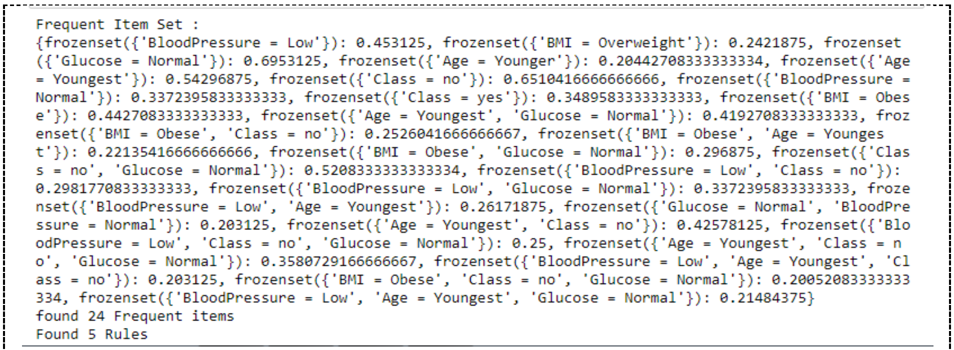
(b) Glucose

(c) BMI

(d) Blood Pressure

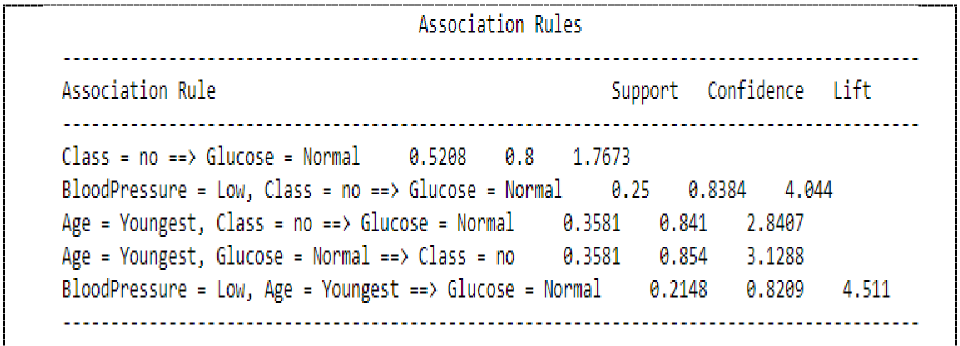
Figure 11. Graph of all binned features.

Figure 11 shows features obtained by the binning method. The age attributes helpful in classifying into five categories, as depicted in figure 11 (a). Glucose measures were separated into 5 classes [34] depicted in figure 11 (b). The relation between BMI and diabetes pervasiveness is dependable.BMI measures were classified into five classes depicted in figure 11 (c). Blood pressure was separated into five various categories depicted in figure 11 (d).



```
Frequent Item Set :
{frozenset({'BloodPressure = Low'}): 0.453125, frozenset({'BMI = Overweight'}): 0.2421875, frozenset
({'Glucose = Normal'}): 0.6953125, frozenset({'Age = Younger'}): 0.20442708333333334, frozenset({'Age
= Youngest'}): 0.54296875, frozenset({'Class = no'}): 0.6510416666666666, frozenset({'BloodPressure =
Normal'}): 0.3372395833333333, frozenset({'Class = yes'}): 0.3489583333333333, frozenset({'BMI = Obes
e'}): 0.4427083333333333, frozenset({'Age = Youngest', 'Glucose = Normal'}): 0.4192708333333333, froz
enset({'BMI = Obese', 'Class = no'}): 0.2526041666666667, frozenset({'BMI = Obese', 'Age = Younges
t'}): 0.2213541666666666, frozenset({'BMI = Obese', 'Glucose = Normal'}): 0.296875, frozenset({'Clas
s = no', 'Glucose = Normal'}): 0.5208333333333334, frozenset({'BloodPressure = Low', 'Class = no'}):
0.2981770833333333, frozenset({'BloodPressure = Low', 'Glucose = Normal'}): 0.3372395833333333, froze
nset({'BloodPressure = Low', 'Age = Youngest'}): 0.26171875, frozenset({'Glucose = Normal', 'BloodPre
ssure = Normal'}): 0.203125, frozenset({'Age = Youngest', 'Class = no'}): 0.42578125, frozenset({'Blo
odPressure = Low', 'Class = no', 'Glucose = Normal'}): 0.25, frozenset({'Age = Youngest', 'Class = n
o', 'Glucose = Normal'}): 0.3580729166666667, frozenset({'BloodPressure = Low', 'Age = Youngest', 'Cl
ass = no'}): 0.203125, frozenset({'BMI = Obese', 'Class = no', 'Glucose = Normal'}): 0.20052083333333
334, frozenset({'BloodPressure = Low', 'Age = Youngest', 'Glucose = Normal'}): 0.21484375}
found 24 Frequent items
Found 5 Rules
```

Figure 12. Frequent itemset.

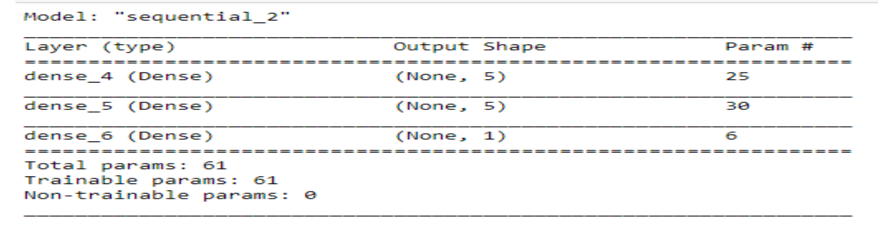


```
                              Association Rules

---------------------------------------------------------------------------

Association Rule                                    Support   Confidence   Lift

---------------------------------------------------------------------------

Class = no ==> Glucose = Normal     0.5208    0.8    1.7673
BloodPressure = Low, Class = no ==> Glucose = Normal    0.25    0.8384    4.044
Age = Youngest, Class = no ==> Glucose = Normal     0.3581    0.841    2.8407
Age = Youngest, Glucose = Normal ==> Class = no     0.3581    0.854    3.1288
BloodPressure = Low, Age = Youngest ==> Glucose = Normal     0.2148    0.8209    4.511
---------------------------------------------------------------------------
```

Figure 13. Association rules.

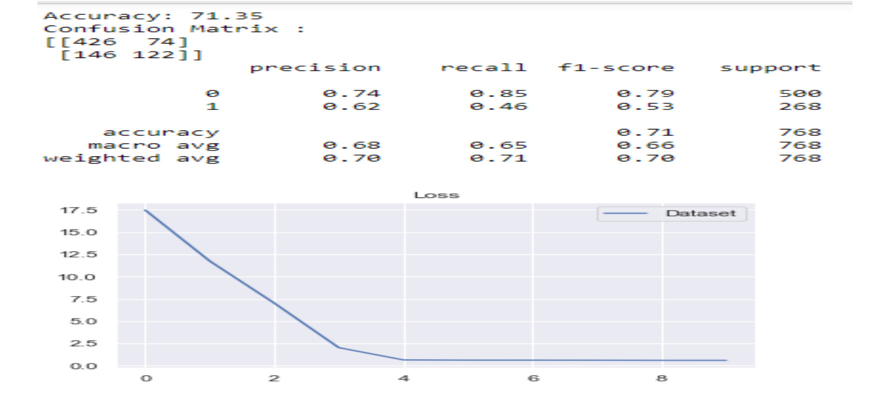
In the dataset, there were 268 patients with diabetes, thus simply that instance was utilized to produce rules between them. To produce association rules from a known dataset requires setting the least support value is 0.25 & the least confidence value is 0.9. The finest rules are depicted in figure 13.

4.3. Base result



```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 5)                 25
_________________________________________________________________
dense_5 (Dense)              (None, 5)                 30
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 6
=================================================================
Total params: 61
Trainable params: 61
Non-trainable params: 0
_________________________________________________________________
```

Figure 14. ANN layer specification.



```
Accuracy:  71.35
Confusion Matrix :
[[426  74]
 [146 122]]
              precision    recall  f1-score   support

           0       0.74      0.85      0.79       500
           1       0.62      0.46      0.53       268

    accuracy                           0.71       768
   macro avg       0.68      0.65      0.66       768
weighted avg       0.70      0.71      0.70       768
```

Figure 15. Confusion matrix for ANN & loss value graph.

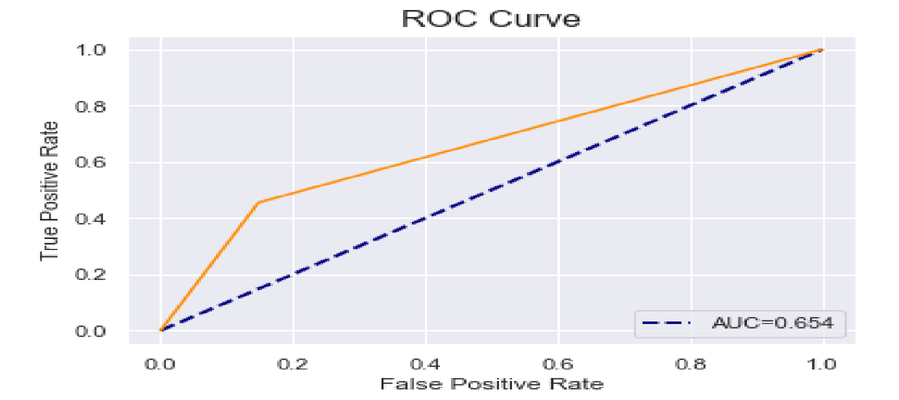
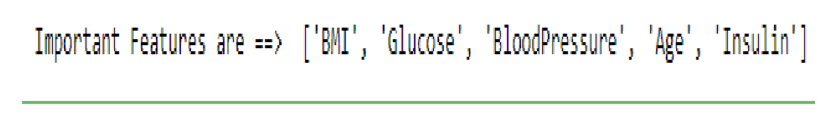


Figure 16. ROC curve of ANN.

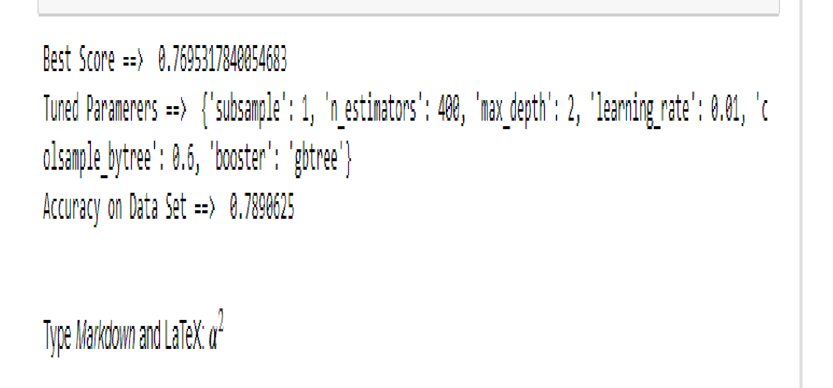
4.4. Proposed results

We have used the median for missing values, Feature selection using RFE (Random Forest), standard scalar for scaling data, and Randomized Search CV for parameter tuning of XGboost. We evaluated our XGboost algorithm to better see the accurate classification. To estimate the presentation, we select four metrics expansively used in the area of boost algorithm: Sensitivity, Specificity, AUC, and Accuracy. We have collated our proposed performance with the existing method.



```
Important Features are ==> ['BMI', 'Glucose', 'BloodPressure', 'Age', 'Insulin']
```

Figure 17. Feature selection using recursive feature elimination with random forest.



```
Best Score ==>  0.7695317840054683
Tuned Paramerers ==>  {'subsample': 1, 'n_estimators': 400, 'max_depth': 2, 'learning_rate': 0.01, 'c
olsample_bytree': 0.6, 'booster': 'gbtree'}
Accuracy on Data Set ==>  0.7890625
```

Type *Markdown* and LaTeX: α^2

Figure 18. Randomized search CV for selection of hyperparameters.

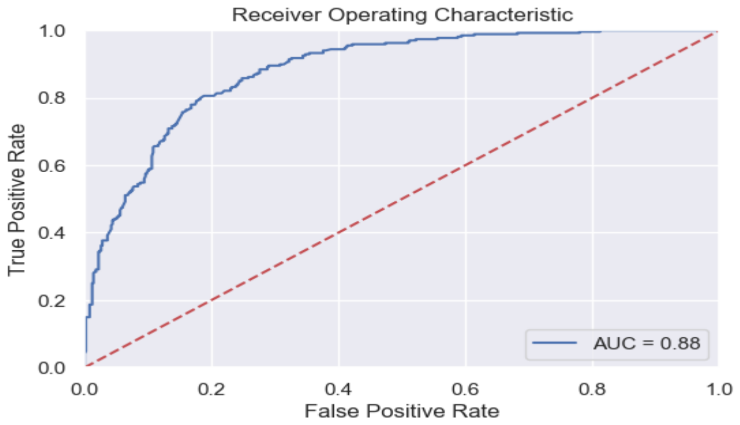
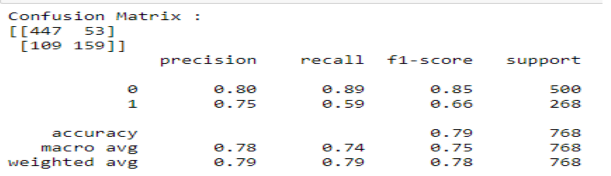


Figure 19. ROC curve of XGboost.



```
Confusion Matrix :
[[447  53]
 [109 159]]
              precision    recall  f1-score   support

           0       0.80      0.89      0.85       500
           1       0.75      0.59      0.66       268

    accuracy                           0.79       768
   macro avg       0.78      0.74      0.75       768
weighted avg       0.79      0.79      0.78       768
```

Figure 20. Confusion matrix for XGboost classification.

Table 1. Comparison of performance parameters.

Method	Accuracy	Sensitivity	Specificity	AUC
ANN	71.35	45.22	85.2	0.65
XGboost	78.91	59.33	89.4	0.88

5. Conclusion

Some of the motivating and necessary petitions of ML are perceived in a medicinal administration. The idea of ML has quickly become pleasing to healthcare organizations. Diabetes is a most growing disease in humanity and this needs regular monitoring. To ensure this we discover different ML algorithms that will help in before time estimation of disease. In this, the introduced model uses an Extreme gradient boosting (XGBoost) technique. This system utilizes the PID dataset from the UCI depository of ML. The accurateness of the proposed algorithm has enhanced in comparison to other existing artificial neural networks. Our introduced system gives an accuracy of 78.91% for estimating diabetes. In prospect, Light GBM can be utilized to enhance the performance to integrate large data and produce more accuracy.

Acknowledgments

We pay thanks to Prof. Varun Singh from Rewa Institute of Technology for assistance & his proper guidance especially technically purpose. I also thank our fellow Members for supporting and inspiring me.

References

[1] Koh, Chye H and Tan G 2011 Data mining applications in healthcare *Journal of healthcare information management* **19** 65

[2] Kumar B S and Gunavathi R 2016 A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis *International Journal of Advanced Research in Computer and Communication Engineering* **5** pp 463-67

[3] Iyer A, Jeyalatha S and Sumbaly R 2015 Diagnosis of Diabetes Using Classification Mining Techniques *International Journal of Data Mining & Knowledge Management Process (IJDKP)* **5** pp. 1-14

[4] American Diabetes Association 2010 Diagnosis and classification of diabetes mellitus. Diabetes care, 33 Suppl 1(Suppl 1), S62–S69. https://doi.org/10.2337/dc10-S062.

[5] Indoria P and Rathore Y K 2018 A Survey: Detection and Prediction of Diabetes Using Machine Learning Techniques *International Journal of Engineering Research & Technology (IJERT)* **7** pp. 287-91.

[6] Alam T M, Iqbal M A, Ali Y, Wahab A, Ijaz S, Baig T I and Abbas Z 2019 A model for early prediction of diabetes *Informatics in Medicine Unlocked* **16** 100204

[7] Jayanthi N, Babu B V and Rao N S 2017 Survey on clinical prediction models for diabetes prediction *J. Big Data* **4** pp 1-15

[8] Ishtake S H and Sanap S A 2013 Intelligent heart disease prediction system using data mining techniques *International Journal of Healthcare & Biomedical Research* **1** pp 94-101

[9] Tomar D 2013 A survey on Data Mining approaches for Healthcare *International Journal of Bio-Science and Bio-Technology* **5** pp 241-66

[10] Milovic B and Milovic M 2012 Prediction and decision making in health care using data mining *Kuwait Chapter of the Arabian Journal of Business and Management Review* **1** p 126.

[11] Illhoi Y et al. 2012 Data mining in healthcare and biomedicine: a survey of the literature *Journal of medical systems* **36** 2431-48

[12] Chaurasia V and Pal S 2013 Early prediction of heart diseases using data mining techniques *Carib. j. SciTech* **1** pp 208-17

[13] Methaila A et al. 2014 Early Heart Disease Prediction Using Data Mining Techniques *Computer Science & Information Technology (CS &IT)*

[14] Kavousi M et al. 2012 Evaluation of newer risk markers for coronary heart disease risk classification: a cohort study *Annals of Internal Medicine* **156** pp 438-44.

[15] Ranganatha S et al. 2013 Medical data mining and analysis for heart disease dataset using classification techniques *Research & Technology in the Coming Decades (CRT 2013), National Conference on Challenges in IET*

[16] Phattharat S and Sripanidkulchai K 2016 Improving type 2 diabetes mellitus risk prediction using classification *13th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE* pp 1-6

[17] Sajida P et al. 2016 Performance Analysis of Data Mining Classification Techniques to Predict Diabetes *Procedia Computer Science* **82** pp 115-121

[18] Cynthiya J L and Amanullah K M The Surveillance on Diabetes Diagnosis Using Data Mining Technique

[19] Patil B M, Joshi R C and Toshniwal D 2010 Association rule for classification of type-2 diabetic patients *Second International Conference on Machine Learning and Computing (ICMLC) IEEE* pp 330-34

[20] Sankaranarayanan S 2014 Diabetic Prognosis through Data Mining Methods and Techniques *International Conference on Intelligent Computing Applications (ICICA) IEEE* pp 162-66

[21] Butwall M and Kumar S 2015 A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier *International Journal of Computer Applications* **120** pp 36-39

[22] Vijayan V V and Anjali C 2015 Prediction and diagnosis of diabetes mellitus—A machine learning approach *IEEE Recent Advances in Intelligent Computational Systems (RAICS)* pp 122-27

[23] Shing L C and Wang M H 2011 A fuzzy expert system for diabetes decision support application *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **41** pp 139-153

[24] Kumari S and Singh A 2013 A data mining approach for the diagnosis of diabetes mellitus *7th International Conference on Intelligent Systems and Control (ISCO) IEEE* pp 373-75

[25] Fayssal B and Chikh M A 2013 Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm *Computer methods and programs in biomedicine* **112** pp 92-103

[26] Wenxin Z and Ping Zhong P 2014 A new one-class SVM based on hidden information *Knowledge-Based Systems* **60** pp 35-43

[27] Davar G et al. 2012 Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search *arXiv preprint* **1** (arXiv:1201.2173)

[28] Benhar H, Idri A, Alemán J F 2018 Data preprocessing for decision making in medical informatics: potential and analysis *World conference on information systems and technologies* pp 1208–18

[29] Abidin N Z, Ismail A R, Emran N A 2018 Performance analysis of machine learning algorithms for missing value imputation *Int. J. Adv Computer Science Appl.* **9** pp 442–7

[30] Liu H, Motoda H 2018 Feature selection for knowledge discovery and data mining *Springer Science & Business Media* **454** 214

[31] Malley B, Ramazzotti D, Wu J T-y. 2016 Data pre-processing *Secondary analysis of electronic health records Springer* pp 115–41

[32] Egi M, Bellomo R, Stachowski E, French C J, Hart G K, Hegarty C, et al. 2008 Blood glucose concentration and outcome of critical illness: the impact of diabetes *Crit Care Med* **36** pp 2249–55

[33] Menke A, Rust K F, Fradkin J, Cheng Y J, Cowie C C 2014 Associations between trends in race/ethnicity, aging, and body mass index with diabetes prevalence in the United States: a series of cross-sectional studies *Ann Intern Med* **161** pp 328–35

[34] Brunström M and Carlberg B. 2016 Effect of antihypertensive treatment at different blood pressure levels in patients with diabetes mellitus: systematic review and meta-analyses *BMJ* **352** 1-10

[35] Solanki S K and Patel J T 2015 A Survey on Association Rule Mining *Fifth International Conference on Advanced Computing & Communication Technologies* 15953566 pp 212-16

[36] Omana J, Monika S and Deepika B 2017 Survey on Efficiency of Association Rule Mining Techniques *International Journal of Computer Science and Mobile Computing* **6** pp 5-8

[37] Umadevi S and Marseline K S J 2017 A survey on data mining classification algorithms *International Conference on Signal Processing and Communication (ICSPC)* 3691956 pp 264-268

[38] Priyadharshini P 2017 Prediction of Diabetes Mellitus Using XGboost gradient Boosting *International Journal of Advances in Science Engineering and Technology* **5** pp 48-50

[39] Chen M, Liu Q, Chen S, Liu Y, Zhang C and Liu R 2019 XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System *IEEE Access* p 99

[40] Lichman M Pima Indians diabetes database *ed. Center for machine learning and intelligent systems: UCI Machine Learning repository*.