International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
***Article Received:** 15 April 2023 **Revised:** 25 July 2023 **Accepted:** 18 August 2023*

Predicting Diabetes Risk Using an Improved Apriori Algorithm

Arpit Solanki
Department of Computer Science & Engineering
Dr. A.P.J. Abdul Kalam University
Indore, India
arpit.solanki29@gmail.com

Dr. Rajeev G. Vishwakarma
Department of Computer Science & Engineering
Dr. A.P.J. Abdul Kalam University
Indore, India
rajeev@mail.com

Abstract: The data mining is able to analyze data and recover the valuable insights from the data. These insights are used for serving in different applications. In this context different data mining algorithms has been developed among them frequent pattern mining has an essential role. In this paper, the frequent pattern mining technique has been implemented for analyzing the diabetic risk. In this context, the popular diabetic dataset has been obtained. Then, the preprocessing has been done on dataset for cleaning the dataset. Next, an encoding process has been developed to transform the dataset. This transform dataset is an effort to deal with the continuous values using the frequent pattern algorithm. Further a modified apriori algorithm has been employed to understand and establish the relationships between diabetic attributes. The experiments have been carried out and theexperimental performance of the improved apriori algorithms has been measured. Additionally a comparison has also been performed with three popular frequent pattern mining algorithms. According, to the performance, we found that the proposed apriori algorithm is efficient and accurate algorithm to predict the diabetic risk.

Keywords: Frequent pattern mining, Association rule mining, Data mining, Rule mining techniques, experimental comparison.

I. INTRODUCTION

How the medical emergencies are affecting entire world we learnt in recent years due to covid-19. But the sudden rising medical emergencies are less affecting as compared to the life style disease such as diabetes and blood pressure. These diseases are long term diseases and developed slowly in human body. These diseases are also the reason to develop more other diseases. Therefore, it is needed to cure and identify in early stages to prevent human life loss. In this context, diabetes risk analysis is an essential task. In order to analyze the data manually the human expertise is required. The human mistake and basic knowledge can be harmful for some one. Therefore, the Machine Learning (ML) based diabetic risk analysis has been proposed in this paper. The aim of the proposed model is to analyze the various risk attributes of the person and identify the possible risk of diabetes.

The ML based techniques are used for data analysis and insights recovery. The algorithm selection for data analysis has been done according to the nature of data and application requirements such as classification, clustering, prediction and rule mining. In this paper, a rule mining technique is used for the diabetic risk analysis. The association rule mining also abbreviated as frequent pattern mining is the process of identifying patterns within a dataset that occur frequently. These techniques have proven their importance in numberof applications such as **Recommender Systems**, **Network Intrusion Detection**, and **Medical data Analysis**.But there are some limitations in rule mining techniques such as:a large number of rules, data processing time and space complexity. In this paper, to address these issues and offer the solution an improved apriori algorithm has also discussed.

The aim is to employ the improved apriori algorithm to the diabetic risk factor dataset and performsearch in database to find frequent itemsets. These frequent itemsets are used to prepare the rules. The rules have been pruned to identify the high quality rules. The improved apriori algorithm (IAA) has allowed the user use high support values for minimizing the amount of rules generated. Additionally it also prevents the data loss that maximizes the accuracy of rules for identifying the diabetic risk. Thus the proposed data model is very promising for developing the Diabetic Risk Prediction (DRP) system. This section discusses the basics overview of the proposed DRP system. The next section includes a review, which demonstrates how the different types of ML algorithms are being used for DRP systems. Third section first highlights the architecture of the proposed system and their key components. Further, a performance analysis has been doneand comparative analysis has been performed. Finally, the conclusion and future plans to extend the proposed study has been discussed.

II. Related Study

This section offers the background study related to the frequent pattern mining and association rule mining.

A. Essential Keywords

Table 1 List of keywords

Abbreviation	**Full form**
AI	Artificial Intelligence
ARM	Association Rule Mining
ANN	Artificial neural network
BBC	Balanced Bagging Classifier
BMI	body mass index
DM	Diabetes Mellitus
DT	Decision Trees
EHR	Electronic Health Records
EMR	Electronic Medical Records
FP-Growth	Frequent Pattern Growth
FPMax	Frequent Pattern Maximal
FES	fuzzy expert systems
ICD-9-CM	International Classification ofDiseases, Ninth Revision, Clinical Modification
KNN	K Nearest Neighbor
LOF	Local Outlier Factor
MIMIC	Medical Information Mart for Intensive Care
MLP	Multi-Layer Perceptron
PCA	principal componentanalysis
RF	random forest
ROC	Receiver Operating Curve
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machines
T2DM	Type 2 DM

B. Related Study

Prediction of diabetes at an early stage can lead to improved treatment. ***T. M. Alam et al [1]***, predicted diabetes using attributes, and the relationship of the differing attributes. Various tools are used to attribute selection, and for clustering, prediction, and ARM for diabetes. Attributes selection was done via the PCA. The findings indicate a strong association of diabetes with BMI and glucose level, which was extracted via the Apriori. ANN, RF and K-means were implemented for the prediction. The ANN technique provided a best accuracy of 75.7%.

In India, the prevalence of diabetes mellitus is very high. In advanced stages it creates complications. Early detection is highly recommended. With the usage of EHR and EMR, it is possible to detect the disease. But, the problem is the storage and maintenance of records which is growing. ***Muni K. N. et al [2]*** predict DM and its risk factors using data mining techniques in the Hadoop. Authors report and describe MapReduce based Apriori algorithms and their performance.

Md. M. Hassan et al [3] have chosen methodologies that give the best performances for independent testing to confirm the universal applicability. They have focused on early detecting this disease. They have collected data from Khulna Diabetes Center where the instances are 289 and 13 features. They use the Logistic Regression with 88%, XGboost 86.36% and RF with 86.36% accuracy.

The healthcare datasets are undefined and influential to manage and operate. ***P. Tiwari et al [4]***, is estimated diabetes by characteristics and relation of contradictory attributes. Features selection was done via the recursive feature elimination with RF. The estimation specifies an alliance of diabetes with BMI and with glucose level was drawing out using the Apriori. The XGBoost gives better accuracy compared to the ANN.

S. Manikanta et al [5] is enhancing the accuracy of insulin quantity by comparing Ensembling Gradient Boost with the Apriori. Bootstrap ensembling and the apriori algorithm were used for predicting. Two sample groups are taken and tested, G-power is calculated, which contains two groups, alpha (0.05), power (90%) and environment ratio. It was observed that the Ensembling algorithm obtains accuracy as 88.42% and have better than the Apriori algorithm. The result proves that the Ensembling technique with varying seed value have significant improvement.

Diabetes patients have more risk of developing multiple conditions. This phenomenon is known as comorbidity. The extracting comorbidities patterns have a few limitations: (1) little is known about comorbidities; (2) identifying top comorbidities patterns. ***Bramesh S. M. et al [6]*** is identifying top-k diabetes-specific disease comorbidity patterns. They used the ICD-9-CM codes to recognize patients with diabetes in MIMIC datasets. They have extracted comorbidity patterns using ARM techniques, namely Apriori, FP-Growth and FPMax. They have proposed an approach to identify top-k association patterns. They also investigate to find differences in gender-specific patterns. Chronic kidney disease was the top most common comorbidity, followed by Atrial fibrillation. Also, diabetes-specific patterns are gender-specific and subtypes-specific.

Diabetes is known as the "Slow Poison" that deteriorates the condition of patient and causes multiple organs failure. ***R. Patil et al [7]*** is used noteworthy features and predict the likelihood of the disease, Decision Tree, RFand SVM have been applied to detect and predict diabetes. Pima Indians Diabetes Dataset is sourced from Kaggle. Precision, F1, Recall and Accuracy are measures on the basis of three Algorithms. RF outperforms and as a result it has the highest accuracy of 78.35%. These results are verified using ROC.

Diabetes is a metabolic disorder and one of the most appalling diseases. In diabetic disease, the body does not respond to “insulin”. ***M. Shuja et al [8]*** develops a model for diabetic prediction based on classifications techniques. Classification of imbalanced data is challenging and was the motivational factor for developing a classifier. A two-phase classification model is employed: first preprocessing is done usingSMOTE, and second feeding five classifiers (Bagging, SVM, MLP, Simple Logistic and DT) with the preprocessed data to select the best classifier to predict diabetes. They achieved an accuracy of 94.7013%, and 0.953 ROC curve with decision tree.

Early prediction for diseases positively impacts healthcare quality and can prevent serious complications. ***D. S. Khafaga et al [9]*** constructs a prediction system for predicting diabetes. The system starts with LOF-based outlier detection. A BBC technique is used to balance data. Finally, integration between association rules and classification algorithms is used to develop a prediction model. Four algorithms (ANN, DT, SVM, and KNN) were utilized to discovered relationships between various factors. Results revealed that KNN provided the highest accuracycompared to the others.

ML techniques are applied to predict the onset and reoccurrence of the disease. Continuously increasing diabetic patient data has necessitated for the applications of efficient ML algorithms, which learns from data and recognizes the critical conditions. ***A. Singh et al [10]***, an ensemble-based framework named eDiaPredict is proposed. It includes an ensemble of ML algorithms comprising XGBoost, RF, SVM,ANN, and DT. The performance of eDiaPredict has been evaluated using accuracy, sensitivity, specificity, Gini Index, precision, area under curve, area under convex hull, minimum error rate, and minimum weighted coefficient. Its application on the PIMA Indian diabetes datasetis achieved.

Diabetes has affected millions of people. Its diagnosis, prediction, cure, and management are crucial. ***F. A. Khan et al [11]*** present a review of the state-of-the-art in the area of diabetes diagnosis and prediction. The aim is:(1) explore and investigate the data mining based diagnosis and prediction solutions. (2)Provides a classification and comparison of the techniques that have been frequently used based on important metrics. They highlight the challenges and research directions that can be considered.

Most of the diabetes patients suffer from T2DM. ML techniques are tools that can improve the analysis and interpretation of knowledge. These techniques may enhance the prognosis and diagnosis. ***F. Kazerouni et al [12]*** applied classification models, including KNN, SVM, logistic regression, and ANN, and compared with each other. They performed the algorithms on six LncRNA variables and demographic data.They aimed to find the best approach for the prediction. According to the finding, the maximum AUC dedicated to SVM and logistic regression. KNN and ANN also had the high mean AUC, KNN had the highest mean sensitivity and the highest specificity.

DM is critical diseases and lots of people are suffering. Big Data Analytics plays a significant role in healthcare. In existing method, the prediction accuracy is not so high. ***A. Mujumdar et al [13]*** have proposed a diabetes prediction model, which includes few external factors responsible for diabetes along with regular factors. Classification accuracy is boosted with new dataset. A pipeline model used for diabetes prediction intended towards improving accuracy.

With increasing number of diabetic patients, its early detection becomes essential. ***H. Thakkar et al [14]***, focused on data mining and fuzzy logic in diabetes diagnosis. The FES analyzes the knowledge from data, which might be vague and suggests linguistic concept with approximation as its core to texts. The methodology delivers the pipeline of tasks such as selecting the dataset, preprocessing. Feature extraction technique is implemented for improving the accuracy and finally data mining and fuzzy logic applied. The accuracy through RF classifiers as high as 99.7% and in fuzzy logic approaches, high precision and low complexity was found.

The common diabetes complications are retinopathy, nephropathy and neuropathy. In order to prevent them, data mining technique needed to extract risk factor. ***C. Fiarni et al [15]*** is constructing a model for three major diabetes complications. The diabetes risk factors are Age, Gender, BMI, Family history, Blood pressure, duration of diabetes and glucose level. Naive Bayes Tree and C4.5 tree-based classification and k-means clustering were used. They evaluated the performance and found the feature and sub feature. The most influential risk factor for Retinopathy is a female patient. As for Nephropathy, the risk factor is the duration. But for Neuropathy, it dominated for females, with BMI more than 25. As for family history of diabetes, there is no distinct significant. The overall accuracy is 68%.

ML is a subset of AI, plays a promising role in prediction. ***R. Birjais et al [16]*** have tried to focus on diagnosis of Diabetes. They have also tried to show different techniques like Gradient Boosting, Logistic Regression and Naive Bayes, which can be used for the diagnosis of diabetes with accuracy as 86% for the Gradient Boosting, 79% for Logistic Regression and 77% for Naive Bayes.

Diabetes Retinopathy is a disease which results from a prolongedcase of DM and cause loss of vision. ***T. O. Oladele et al [17]*** was aimed atdetermining the effect of feature selection in predicting DiabetesRetinopathy. The dataset used for this study was gotten from diabetes retinopathyDebrecen dataset from the University of California.Feature selection was executed then the Implementation of k-NN, C4.5 decision tree, MLP and SVM was conducted. It is observed that, use offeature selection increases the accuracy and the sensitivity and mostly reflected in the SVM. Feature selection also increases thetime taken for the prediction of diabetes retinopathy.

III. PROPOSED WORK

The main aim of the proposed work is to design an accurate diabetic risk prediction model using ML algorithm. In this context, the previously proposed modified apriori algorithm has been considered for system development. The required diabetes risk prediction system is demonstrated in figure 1.

Article Received: *15 April 2023* ***Revised:*** *25 July 2023* ***Accepted:*** *18 August 2023*

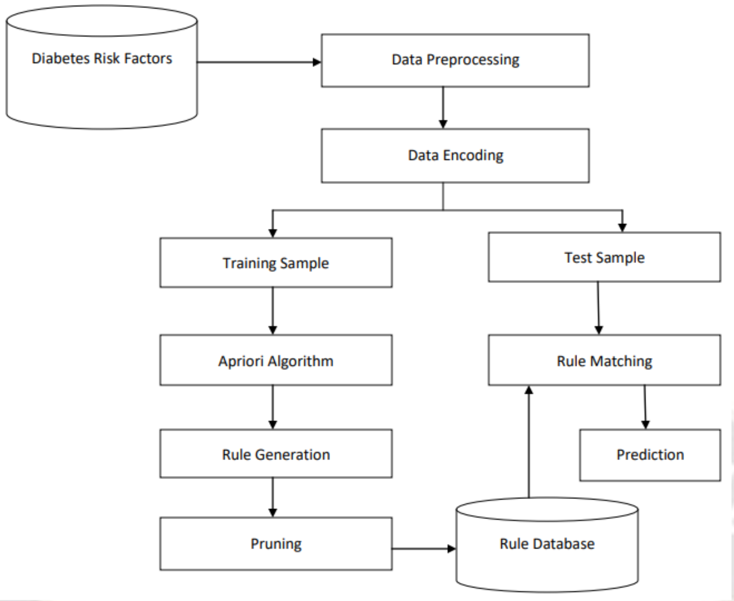

Figure 1: Proposed Diabetes Risk Prediction Model

The initial component of the proposed model is diabetic dataset, which is taken from the Kaggle [18]. The dataset consist of the following set of attributes:

Table 1 dataset attributes and it's type

Attribute	**Type**
Gender	Male, Female
Age	Continuous values
Hypertension	0 or 1
Heart disease	0 or 1
Smoking history	No info, never, other
Body Mass Index (BMI)	Continuous values
Hemoglobin A1c (HbA1c)	Continuous values
Blood Glucose Level	Continuous values
Diabetes (Class label)	0 or 1

Now we perform a preprocessing operation to deal with the dataset. The aim of the preprocessing is to transform and encode the dataset in transactional patterns. In this context first we create a set of symbols, which is used to map the dataset. In this context, the attributes Age,Smoking history, BMI, HbA1c and blood glucose level has been first transformed into a set of attributes. The new set of attributes is given in table 2.In order to create the new variables from the continuous values we have created two thresholds. And then map the values into a new variable. In order to understand the calculation of the proposed concept of threshold calculation considers the figure 2. Let an attribute X, which is needed to be transforming into three new symbols(X_1, X_2, X_3).

Table 2 New symbol table

Old symbol	**New symbol**
Age	A_1, A_2, A_3
Smoking history	S_1, S_2, S_3
BMI	B_1, B_2, B_3
HbA1c	H_1, H_2, H_3
blood glucose level	b_1, b_2, b_3

Then In a two-dimensional search space which is lies between two values V_{max} and V_{min}. Now the mean of the search space is defined using black dotted line. Additionally calculated using:

$$Avg = \frac{V_{max} + V_{min}}{2}$$

Additionally, the threshold is defined by the following equation:

$$T_1 = \frac{V_{max} + avg}{2}$$

$$T_2 = \frac{avg + V_{min}}{2}$$

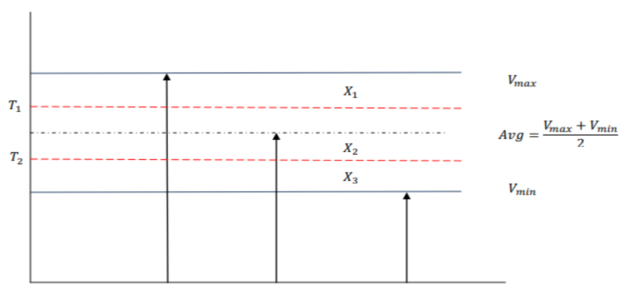

Figure 2: Shows the concept of threshold calculation

In figure 2 the threshold is demonstrated using red dotted line. Using these two thresholds the attributes are divided into three different symbols:

$$f(X) = \begin{cases} if\ V_{min} \le X \le T_2\ then\ X_3 \\ if\ T_2 < X \le T_1\ Then\ X_2 \\ if\ T_1 < X \le V_{max}\ then\ X_1 \end{cases}$$

The above given function has work as the mapping function to convert the values into a new symbol. Using this process we convert the continuous values into three variables. Next for performing the encoding an algorithm is proposed which is given below in table 3. This process has initiated when the complete mapping of the dataset has been done using the symbols given in table 2. Now to generate the new transactional sets we apply the process given in table 3. The algorithm accepts the attribute one by one from the dataset, and then finds the number of unique variables in that attribute. Then, the attribute has mapped differently for both two symbol attribute and three symbol attribute.

Before performing the encoding of the entire dataset into the transactional set, a new dataset has been developed. In this dataset the table contains all the Boolean variables of the dataset (Gender, Hypertension, Heart disease and smoking history). These variables are directly incorporated into the new dataset table then the remaining attributes (Age,Smoking history, BMI, HbA1c and blood glucose) are transformed into three variables.Therefore, for each attribute three new attributes has been added to the new dataset.The final preprocessed dataset has a total of 4 Boolean attributes and 15 new attributes. Thus the new dataset contains a total of 19 attributes and 1 class label. Now the table demonstrates how the new dataset values are prepared by using the previous dataset.

Table 3 encoding algorithm

Input: The Dataset D,

Output: New Dataset ND

Process:

```
1. $A_n = readDatasetAttributes(D)$
2. $for(i = 1;\ i \le n\ ; i++)$
    a. $temp = GetUniqueCount(A_i)$
    b. $if(temp == 2)$
        i. $if(V_A{}^i == 1)$
            1. $A_i = 1$
        ii. Else
            1. $A_i = 0$
        iii. End if
    c. $Else\ if(temp == 3)$
        i. $[A_1, A_2, A_3] = CreateMapping(A_i)$
        ii. $ND.Add([A_1, A_2, A_3])$
        iii. $Temp = f(V_A{}^i)$
        iv. $if(Temp == X_1)$
            1. $A_1 = 1$
        v. Else
            1. $A_1 = 0$
        vi. End if
        vii. $if(Temp == X_2)$
            1. $A_2 = 1$
        viii. Else
            1. $A_2 = 0$
        ix. End if
        x. $if(Temp == X_3)$
            1. $A_3 = 1$
        xi. Else
            1. $A_3 = 0$
        xii. End if
    d. End if
3. End for
4. Return ND
```

After encoding of the dataset into a vector of 20 attributes the rule mining is required. In this context, the previously introduced apriori algorithm has been implemented. This algorithm includes the goodness of frequent patterns obtained from apriori algorithm and rule generation. The generated rules are pruned and created two lists of rules according to the rules quality and to prevent the data loss. The steps of the proposed modified apriori algorithm are given below:

Table 7 Proposed Apriori Algorithm

Input: Dataset D, Support Threshold T, confidence threshold conf

Output:$SList, FList$

Process:

```
1. Find total symbols (S) in dataset (D)
2. Apply support threshold to filter symbols
3.
    a. $if\ condidate = true$
        a. Generate candidate set C
        b. Filter C using support threshold T
        c. F = F + L
    b. Else
        a. Flag = false
    c. End if
    d. While next
4. End while
5. Use F to generate association rules $R_n$
6. $For(i = 1, i \le n; I++)$
    a. $conf = CalculateConf(R_i)$
    b. $if\ conf < confidence\ threshold$
        a. $SList.Add(R_i)$
    c. $else$
        a. $FList.Add(R_i)$
    d. $end\ if$
```

Article Received: *15 April 2023* ***Revised:*** *25 July 2023* ***Accepted:*** *18 August 2023*

```
7.  $end\ for$
8.  Return $SList, FList$
```

The proposed algorithm returns two lists as compared to one list of rule. The list $SList$ contains the rules those have confidence less then support confidence, additionally the high quality rules are kept on the $FList$. During the utilization of these rules first the $FList$ rules are being used and if the data has not satisfied than the $SList$ rules are used. That process enhances the reorganization process and improves the accuracy of rule based classification. Next the validation set has been used to test the proposed algorithm. Thus the 30% of encoded data has been used here as validation samples.These samples are use with the generated rules and classify them according to their class labels. Finally the experiments have been carried out. The next section includes the results obtained by the experiments.

IV. RESULTS ANALYSIS

In this section, performance of the proposed diabetes risk prediction system using association rule mining algorithm has been evaluated. Additionally, a comparison with the traditional association rule mining algorithms has also been presented. In this section the experimental results has been discussed.

A. Accuracy

First we have measured accuracy of the proposed model and compared with the three popular traditional association rule mining algorithms. The accuracy of the model has been measured using the following equation.

$$\Delta = \frac{T_C}{T_T}$$

Where T_C is correctly classified samples and T_T total samples.

The measured accuracy of the algorithms for diabetes risk prediction is given in Figure 3.The X axis shows the dataset and Y axis shows the accuracy of prediction. According to the performance the proposed method provides higher accuracy of the prediction. Therefore, the proposed algorithm is reliable algorithm and demonstrates less data loss. Therefore the proposed method is accurate and can predict the risk of diabetes. The proposed algorithm demonstrates the accuracy above 80% and also helps to identify the precise value range of attributes to identify the risk level.

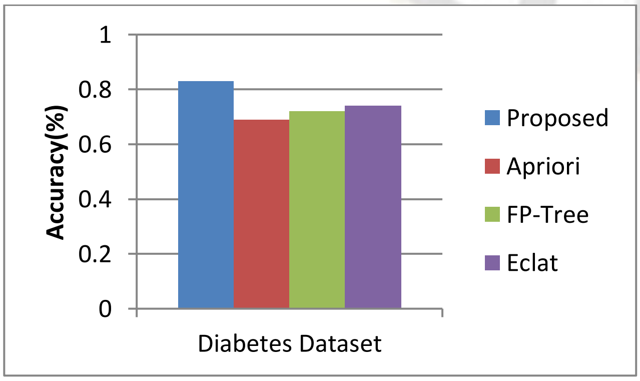

Figure 3: Rule based classification accuracy

B. Training Time

The amount of time consumed in calculating the decision rules is termed here as training time. The training time is calculated by the following equation:

$$T = T_E - T_S$$

Where T_E the last time when the rule generation algorithm ends and T_S is start time of rule generation.

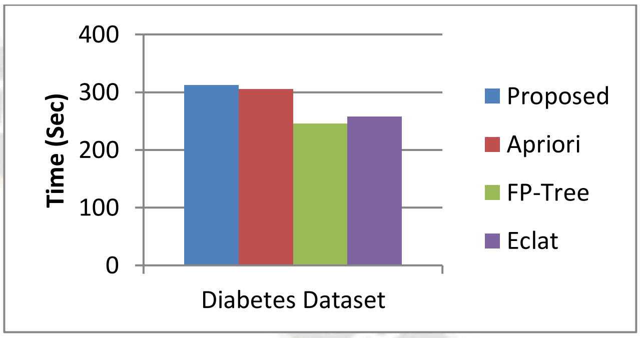

Figure 4: Training time

Figure 4 shows the training time of the algorithms implemented for diabetes risk prediction. The 70% of total dataset has been used for performing training or rule generation. The time has measured here in terms of seconds (sec). The X axis shows the dataset used and Y axis shows the training time of the algorithm. According to the obtained results it is found that the proposed algorithm has similar time complexity as apriori algorithm. Therefore, it is little expensive in terms of time consumption. According to the recorded training time the éclat and FP-Tree algorithm are time efficient algorithms. But the accuracy of the proposed algorithm makes it acceptable for the utilization in real world applications.

C. Decision Time

The classification of samples for decision making is also requires an amount of time. This amount of time is termed as the decision time. The decision time of the algorithms for entire 30% of samples has been measured and reported. The similar formula has been used as the training time to measure the decision time. The decision time of the algorithms are demonstrated in figure 5. The X axis contains the dataset used and Y axis shows the time consumption of processing the validation set.

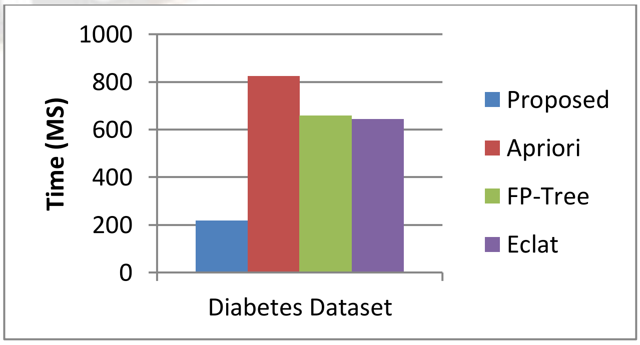

Figure 5: decision time

According to the decision time of the proposed algorithm is one third of the other algorithms. the proposed algorithm generates less amount of rules as compared to the other algorithms thus it provide faster results as compared to other algorithms. The time has measured here in terms of milliseconds (MS).

V. CONCLUSION & FUTURE WORK

The frequent pattern mining or association rule mining technique has a wide area of applications.Additionally the rule mining techniques has proven their importance in medical data analysis.In this presented worka diabetes risk prediction systemhas been proposed. The DRP system utilizes the risk factors and predicts the possible risk of diabetes. The DRP system first preprocesses the dataset and then transforms the dataset into a new mapping vector. This mapping vector is used for rule generation. The generated rules have further pruned to improve the rules quality. Finally the validation of the model has been done using a test sample set. By classifying the test samples the performance of the proposed DRP model has been measured. The performance of the system has been measured in three parameters and compared with the traditional rule mining algorithms.

The experimental results demonstrate the proposed model can provide accurate prediction of the diabetic risk. In addition, the method provides faster decisions due to less number of rules and high quality of rule selection. The proposed DRP technique not only predict the risk of diabetes it can also make precise decision due to extension of dataset. Therefore proposed data encoding technique has also reflected their importance. Based on the advantage of the proposed model the system is accurate, efficient and reliable for performing the prediction. The proposed technique can also be used for different disease prediction systems. Thus recommended to use in other kind of prediction.

REFERENCES

[1] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, Z. Abbas, “A model for early prediction of diabetes”, Informatics in Medicine Unlocked, 16, 100204, 2019

[2] Muni K. N., Manjula R, “Survey on map reduce based apriori algorithms in medical field for the prediction of diabetes mellitus”, Research Journal of Fisheries and Hydrobiology, 11(4): 13-18, 2016.

[3] Md. M. Hassan, Z. J. Peya, S. Mollick, Md. A. M. Billah, Md. M. H. Shakil, A. U. Dulla, “Diabetes Prediction in Healthcare at Early Stage Using Machine Learning Approach”, 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021

[4] P. Tiwari, V. Singh, “Diabetes disease prediction using significant attribute selection and classification approach”, Journal of Physics: Conference Series, 1714, 012013, 2021

[5] S. Manikanta, A. Rama, “Comparing Gradient Boost and Apriori Technique for Increasing Diabetes Insulin Quantity Accuracy”, Journal of Survey in Fisheries Sciences 10(1S) 2448-2456, 2023

[6] Bramesh S. M., Anil K. K. M., “An Effective Rule Based Approach For Identification Of Comorbidity Patterns In Diabetic Patients”, Indian Journal of Computer Science and Engineering (IJCSE), Vol. 13 No. 4 Jul-Aug 2022

[7] R. Patil, L. Majumder, M. Jain, V. Patil, “Diabetes Disease Prediction Using Machine Learning”, International Journal of Research in Engineering, Science and Management, Volume-3, Issue-6, June-2020

[8] M. Shuja, S. Mittal, M. Zaman, “Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE”, Diabetes Nutrition & Metabolism, 15(4), 215–221

[9] D. S. Khafaga, A. H. Alharbi, I. Mohamed, K. M. Hosny, “An Integrated Classification and Association Rule Technique for Early-Stage Diabetes Risk Prediction”, Healthcare 2022, 10, 2070

[10] A. Singh, A. Dhillon, N. Kumar, M. S. Hossain, G. Muhammad, M. Kumar, “eDiaPredict: An Ensemble-based Framework for Diabetes Prediction”, ACM Trans. Multimedia Comput. Commun. Appl., Vol. 17, No. 2s, Article 66. Publication date: June 2021.

[11] F. A. Khan, K. Zeb, M. A. Rakhami, A. Derhab, S. A. C. Bukhari, “Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review”, IEEE Access, Vol. 9, 2021

[12] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, Z. Mansoori, “Type2 diabetes mellitus prediction using data mining algorithms based on the longnoncoding RNAs expression: a comparison of four data mining approaches”, BMC Bioinformatics (2020) 21:372

[13] A. Mujumdar, Dr. Vaidehi V, “Diabetes Prediction using Machine Learning Algorithms”, Procedia Computer Science 165 (2019) 292–299

[14] H. Thakkar, V. Shah, H. Yagnik, M. Shah, “Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis”, Clinical eHealth 4 (2021) 12–23

[15] C. Fiarni, E. M. Sipayung, S. Maemunah, “Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm”, Procedia Computer Science 161 (2019) 449–457

[16] R. Birjais, A. K. Mourya, R. Chauhan, H. Kaur, “Prediction and diagnosis of future diabetes risk: a machine learning approach”, SN Applied Sciences (2019) 1:1112

[17] T. O. Oladele, R. O. Ogundokun, A. A. Kayode, A. A. Adegun, M. O. Adebiyi, “Application of Data Mining Algorithms for Feature Selection and Prediction of Diabetic Retinopathy”, ICCSA, LNCS 11623, pp. 716–730, 2019

[18] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset