ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Disease Symptoms Analysis Using Data Mining Techniques to Predict Diabetes Risk

Jawad Kamal
Department of Computer Science
Jamia Hamdard
New Delhi, India
Jawadkamal10@gmail.com

Dr Safdar Tanveer
Department of Computer Science
Jamia Hamdard
New Delhi, India
safdartanweer@jamiahamdard.ac.in

Md. Tabrez Nafis
Department of Computer Science
Jamia Hamdard
New Delhi, India
nafis303@gmail.com

Abstract: Data mining field concentrates on theories, concepts, methodologies and mainly on extraction of useful knowledge from large amounts of data for decision making. During their day to day activities healthcare industry generates large amounts of healthcare information that has not been efficiently used to extract unknown information. Therefore the discovery of interesting and useful information usually remains hidden. Diabetes is a healthcare problem and is increasing at a greater rate with every passing year. If not recognized early, can lead to severe health problems, even in organ failures. Several data mining techniques like clustering, classification, association rule mining are used to identify early symptoms of the diseases and stopping them getting to a chronic level. In this paper, an efficient approach has been designed for prediction of risk of getting diabetes using diabetes database. The approach in this paper used more than one data mining techniques showing enhanced result in disease prediction. The data for diabetes is collected and processed to facilitate the mining process. Firstly, the preprocessed database is mined to extract frequent patterns related to diabetes using FP-Growth algorithm. After that ID3 algorithm approach has been used as the training algorithm to depict the risk of diabetes using a Decision Tree.

Keywords: Data Mining, Classification, Association Rule Mining, Symptoms, FP-Growth, Decision Tree.

I. Introduction

Data Mining is ahead more recognition because of its power to extract knowledge from voluminous data where it is beyond the reach of traditional techniques of knowledge discovery and human understanding [1]. Data mining is most beneficial in extracting hidden knowledge and exploratory analysis due to nontrivial information lying in huge volumes of data [2].

Data mining is used in many fields to find interesting patterns and sequences which helps in better analysis and enhanced decision making. The data mining techniques can also be inculcated in healthcare field, as they are useful in predicting various diseases found in the medical field by using huge amount of ever growing medical data[3].

Hospitals and clinics have been gathering huge amounts of medical and patient data over the years. This accumulated data if efficiently harnessed, can be used for medicinal analysis and also for the analysis of the risk factors for many diseases. For example, we can predict the risk of getting diabetes for a person by the patterns and knowledge that were mined using the diabetic patient's data. That could eventually help medical administrator to predict diabetic symptoms in early stages and suggest the respective measures accordingly. After the introduction of frequent itemset mining and association rules, the pattern extraction was recognized. At high frequency thresholds, previously known knowledge was extracted, while low frequency threshold revealed many unknown patterns. These hidden patterns can be used hospitals, clinics to provide better treatment and services. Now a days, a majority of fields related to medical services like medical tests, medication, discovery of relationship among clinical and diagnosis data and prediction of effectiveness of surgical process make use of Data Mining Methodologies [2]. Using data mining it can be analyzed that which course of action is effective, could be known by comparing symptoms, course of treatment followed and medication used for curing. In this paper we depicted prediction of the risk of getting diabetes by analyzing symptoms found in diabetic patients from diabetes database.

The diabetes database consists of mixed attributes consisting of numerical, binomial and categorical data. Records and attributes for diabetic patients are cleaned and filtered, so the data irrelevant to diabetes could be removed from the database before implementing the mining process. Then on the filtered data, frequent patterns related to diabetes diagnosis are mined using FP-Growth algorithm. And Finally, ID3 algorithm is used as the training algorithm to depict the symptoms resulting into diabetes with the help of decision tree.

II. Meaning Of Diabetes

Diabetes is commonly referred to a group of metabolic malfunctioning in which a person has high glucose levels in body either due to inadequate production of insulin in the body or the body cells do not respond to insulin properly or both. So, the uses of glucose levels increase in the body leading to symptoms such as heavy thirst, frequent Urination, weakness etc.

Insulin not only regulates the body glucose but it is also responsible for a person's lipid metabolism. This should also be known that insulin is given as medicine only when above criteria is broken.

Diabetes can be classified into three types [3]:

1. Type-1 Diabetes

Human body cells fail to reproduce insulin. Usually Type-1 diabetes affects people in early adulthood or before age 40. Almost 5-10% of the people around the world have been affected with type-1 diabetes.

2. Type-2 Diabetes

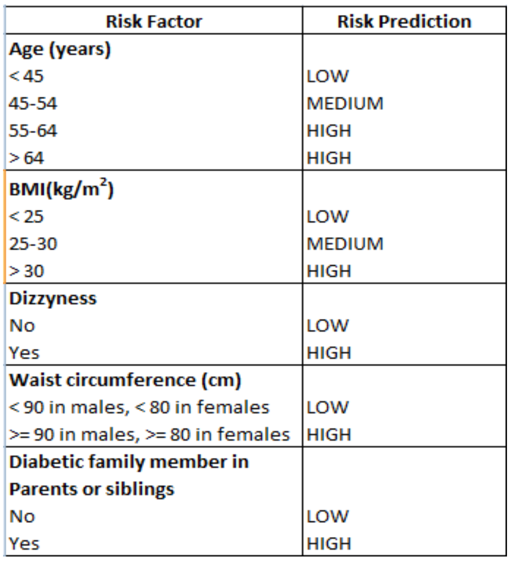
The human body cells stops to act in response with insulin or insulin confrontation. All over the world around 5-10% of all the diabetes cases are of this type.

3. Gestational Diabetes

This kind of diabetes usually affects females during pregnancy.

So, there is no complete cure for diabetes. The only way to protect diabetic people from its harmful consequences is to prevent or minimize blood glucose level, proper nutrition and regular physical exercise.

.



Risk Factor	Risk Prediction
Age (years)	
< 45	LOW
45-54	MEDIUM
55-64	HIGH
> 64	HIGH
BMI(kg/m^2)	
< 25	LOW
25-30	MEDIUM
> 30	HIGH
Dizzyness	
No	LOW
Yes	HIGH
Waist circumference (cm)	
< 90 in males, < 80 in females	LOW
>= 90 in males, >= 80 in females	HIGH
Diabetic family member in Parents or siblings	
No	LOW
Yes	HIGH

Figure 1: Shows the risk levels for diabetes risk factors.

III. RELATED WORK

Data mining is an analytical process which uses one or more techniques to analyze, search and extract the meaningful information from large amounts of available data. Use of data mining techniques with health care transaction has been becoming popular because of the ability of data mining techniques to extract unknown and precious knowledge for supporting decision making processes. There are many types of data mining algorithms such as association rule, forecasting, data clustering and classification. In this paper, the focus is on classification group, used to classify attribute of data objects.

Researchers use various data mining methods for addressing the healthcare issues. Several papers provides applications of different data mining techniques for prediction of various diseases. Sellappan Palaniappan, developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network[4] to predict the likelihood of patients getting a heart disease.

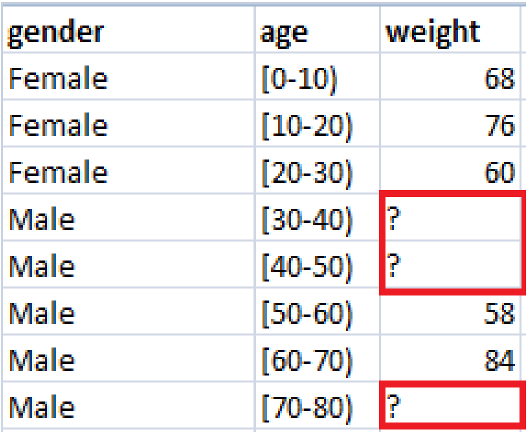
In his research work, S. Stilou applied the apriori algorithm to diabetic patient's database and attempted to extract association rules from the stored real parameters[5]. Dr. Bushra M. Hussan used K-means method for dealing with medical database for clustering [6]. Dr Harleen Kaur examined the impact of data mining techniques, using artificial neural networks on medical Diagnostics[7]. Ujma Ansar, Jyoti Soni experimented that Decision Tree outperforms Bayesian classification, KNN, Neural Networks in predictive data mining for Heart disease prediction [8].

IV. THEORITICAL BACKGROUND

A. *Data Preprocessing*

Data available to us for mining process should be cleaned and processed first to make mining process smooth and efficient. Preprocessed data used in data mining process will avoid the creation of inappropriate or meaningless rules and patterns. In preprocessing an attribute should be selected first. Then it should handle all missing values for it and investigate each possibility like using a global constant to fill the missing value or replacing the missing values with most popular etc. If an attribute has more than 5% missing values then the records should not be deleted instead impute values where data is missing by using a suitable method.

Figure 2 shows data available in diabetes database before preprocessing.



gender	age	weight
Female	[0-10)	68
Female	[10-20)	76
Female	[20-30)	60
Male	[30-40)	?
Male	[40-50)	?
Male	[50-60)	58
Male	[60-70)	84
Male	[70-80)	?

Figure 2: Missing values in the attribute of diabetes database.

B. *FP-Growth*

Frequent Itemset Mining is considered to be one of the basic data mining problem that focuses on groups of items or values or patterns that occur together frequently in a transaction. The discovery of significant patterns from the diabetes database is presented in this section. The diabetes database contains clinical data of diabetes patients. It is of great significance in different data mining tasks that targets to mine interesting patterns from databases, like correlation, association rules, sequences, clusters, classifiers. We used FP Growth algorithm to mine interesting frequent patterns applicable to diabetes from the data extracted.

FP-Growth is a two step approach to find frequent patterns which allows frequent itemset discovered without candidate itemset generation.

Steps involved in FP-Growth algorithm are as follows

Step 1: A compact data structure FP-Tree is build using 2 passes over the data-set.

Step 2: Frequent itemset are directly extracted from FP-Tree. FP-Tree is constructed using 2 passes over the dataset.

Pass-1: Compresses a large database into a compact, frequent pattern tree (FP-Tree) structure.

Pass-2: An efficient FP-Tree based frequent pattern mining is developed.

The major difference between Aprori and FP-Growth is that in FP-Growth candidate itemset is not generated.

C. *Decision Tree Representation*

Classification is an unsupervised learning which is used to predict the class of objects whose class label is not known. It is used for creating classification rules by means of decision trees from a given data set. Decision tree is usually suitable for experimental knowledge discovery as the tree structure depends on the data available to perform prediction. In decision tree model, leaves indicate the outcomes and branches depict attribute values of the dataset. In data mining, decision tree doesn't make any decision rather it visualizes the data for making decisions.

An advantage of decision tree model is that, decision trees can be converted easily into understandable rules. Decision trees have mostly been used to build diagnosis models for medical data. Most commonly used decision tree algorithms are ID3, CART, C4.5, CHAID and J48 etc. In this paper we use ID3 algorithm provided in RapidMiner software to build a decision tree for getting the risk of diabetes from the diabetic patient's data.

V. System Architecture

Figure 3 depicts the model of the system architecture of the approach used in this paper for risk prediction of getting diabetes.

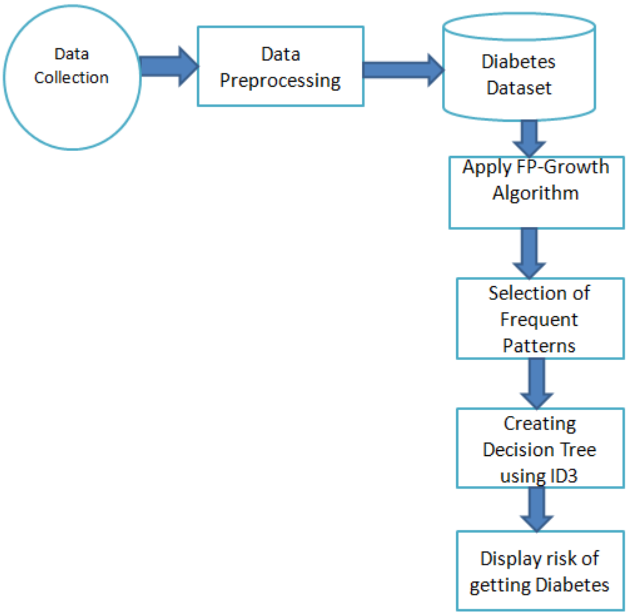
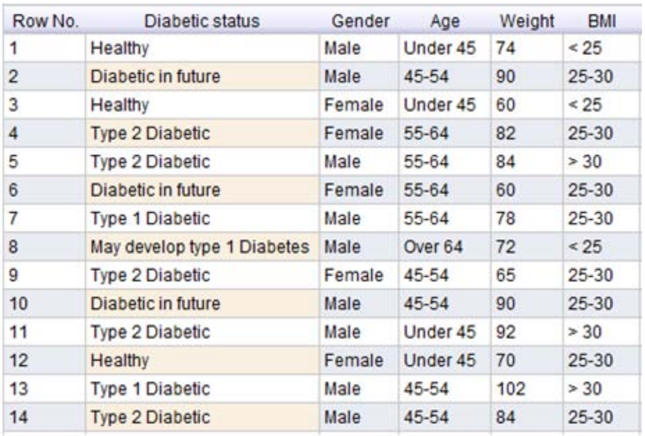

Data Collection
Data Preprocessing
Diabetes Dataset
Apply FP-Growth Algorithm
Selection of Frequent Patterns
Creating Decision Tree using ID3
Display risk of getting Diabetes

Figure 3: System architecture for the proposed model.

VI. Experimental Results

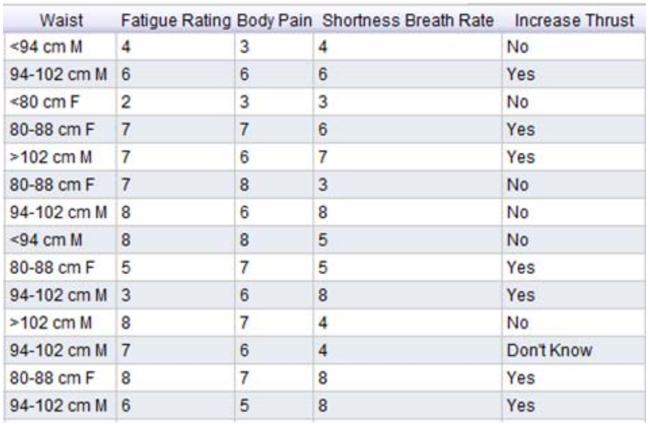
The experiment of finding significant frequent patterns for diabetes was conducted using the following approach. And the results found during the experimental are discussed in this section. With the help of the database, the patterns significant to the prediction of the risk of getting diabetes are discovered using the approach discussed.

The template for diabetes symptoms data was taken from [10][11] and [12].The attributes related to our research perspective were selected and data was gathered for those attributes. And then the collected data was stored in diabetes database. The diabetes database is preprocessed successfully by supplying the missing values as shown in Table 2 A&B. Then from refined diabetes database that we got after preprocessing, frequent patterns were mined efficiently using the FP-Growth algorithm. And then data is trained to predict the risk of getting diabetes with decision tree by using ID3 algorithm by information gain.



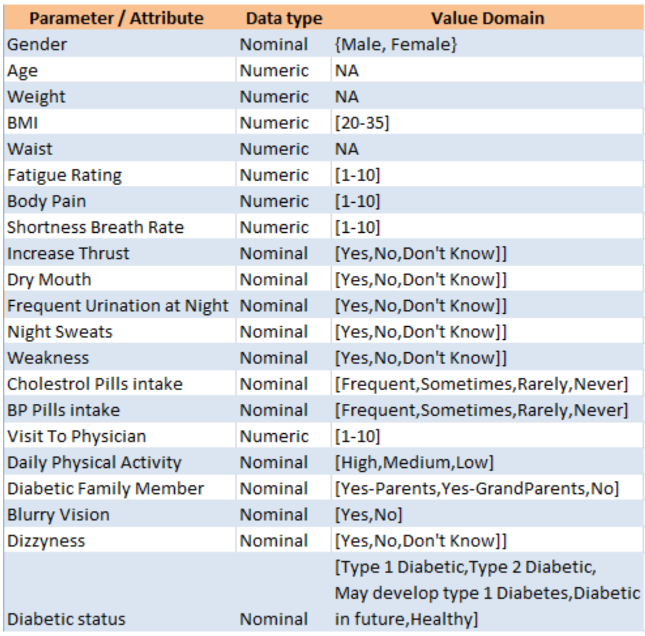
Row No.	Diabetic status	Gender	Age	Weight	BMI
1	Healthy	Male	Under 45	74	< 25
2	Diabetic in future	Male	45-54	90	25-30
3	Healthy	Female	Under 45	60	< 25
4	Type 2 Diabetic	Female	55-64	82	25-30
5	Type 2 Diabetic	Male	55-64	84	> 30
6	Diabetic in future	Female	55-64	60	25-30
7	Type 1 Diabetic	Male	55-64	78	25-30
8	May develop type 1 Diabetes	Male	Over 64	72	< 25
9	Type 2 Diabetic	Female	45-54	65	25-30
10	Diabetic in future	Male	45-54	90	25-30
11	Type 2 Diabetic	Male	Under 45	92	> 30
12	Healthy	Female	Under 45	70	25-30
13	Type 1 Diabetic	Male	45-54	102	> 30
14	Type 2 Diabetic	Male	45-54	84	25-30

Table 2 A



Waist	Fatigue Rating	Body Pain	Shortness Breath Rate	Increase Thrust
<94 cm M	4	3	4	No
94-102 cm M	6	6	6	Yes
<80 cm F	2	3	3	No
80-88 cm F	7	7	6	Yes
>102 cm M	7	6	7	Yes
80-88 cm F	7	8	3	No
94-102 cm M	8	6	8	No
<94 cm M	8	8	5	No
80-88 cm F	5	7	5	Yes
94-102 cm M	3	6	8	Yes
>102 cm M	8	7	4	No
94-102 cm M	7	6	4	Don't Know
80-88 cm F	8	7	8	Yes
94-102 cm M	6	5	8	Yes

Table 2 B: Shows a fragment of the diabetes dataset after applying the pre-processing techniques.



Parameter / Attribute	Data type	Value Domain
Gender	Nominal	{Male, Female}
Age	Numeric	NA
Weight	Numeric	NA
BMI	Numeric	[20-35]
Waist	Numeric	NA
Fatigue Rating	Numeric	[1-10]
Body Pain	Numeric	[1-10]
Shortness Breath Rate	Numeric	[1-10]
Increase Thrust	Nominal	[Yes,No,Don't Know]]
Dry Mouth	Nominal	[Yes,No,Don't Know]]
Frequent Urination at Night	Nominal	[Yes,No,Don't Know]]
Night Sweats	Nominal	[Yes,No,Don't Know]]
Weakness	Nominal	[Yes,No,Don't Know]]
Cholestrol Pills intake	Nominal	[Frequent,Sometimes,Rarely,Never]
BP Pills intake	Nominal	[Frequent,Sometimes,Rarely,Never]
Visit To Physician	Numeric	[1-10]
Daily Physical Activity	Nominal	[High,Medium,Low]
Diabetic Family Member	Nominal	[Yes-Parents,Yes-GrandParents,No]
Blurry Vision	Nominal	[Yes,No]
Dizzyness	Nominal	[Yes,No,Don't Know]]
Diabetic status	Nominal	[Type 1 Diabetic,Type 2 Diabetic, May develop type 1 Diabetes,Diabetic in future,Healthy]

Figure 4: A list of selected parameters for creating the model.

The proposed model is implemented on this dataset shown in Table 2 using RapidMiner [8] tool to predict risk of getting diabetes.

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, validation and optimization [16]. In our approach we used the operators for data retrieval, feature selection, classification and validation etc.

For mining Association rule from the diabetes dataset we first used data access operator (in this case we use Retrieve operator) for reading a data object from data repository. For feature selection we used Select Attribute operator. And then for mining association rule we used FP-Growth operator (using minimum support 20%) with Create Association Rule operator (using minimum confidence 20%).

An association rule is of the form X=>Y, where X ⊂ I, Y ⊂ I and X ∩ Y = ∅. The rule X=>Y holds in the transaction set D with confidence c, if c% of transactions in D that contain X also contain Y. The rule X=>Y has support s, in the transaction set D, if s% of transactions in D contains X ∪ Y [13].

Not all the patterns extracted would be interesting or beneficial for our work. But some of the patterns mined would be very useful and can bring up unknown facts and relations.

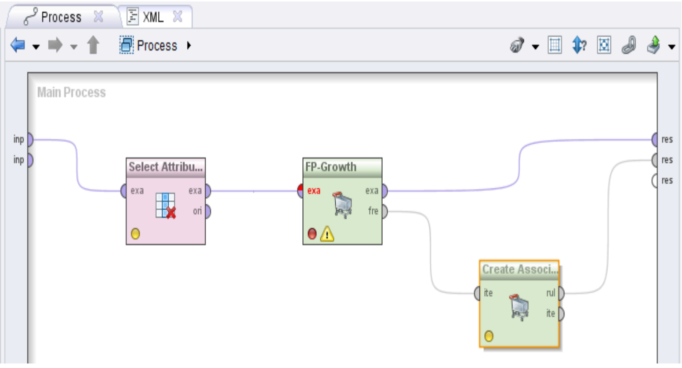
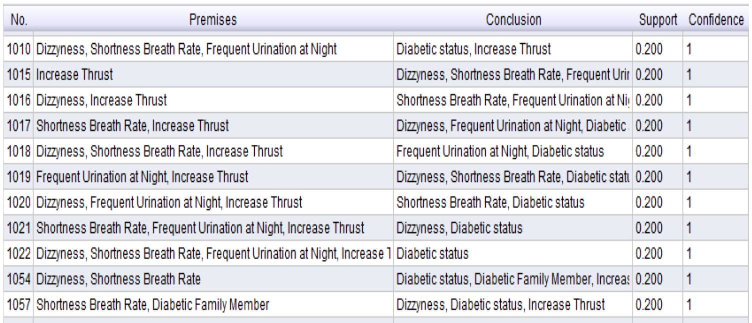


Figure 5: Shows operators setup in RapidMiner tool for mining frequent patterns.



No.	Premises	Conclusion	Support	Confidence
1010	Dizzyness, Shortness Breath Rate, Frequent Urination at Night	Diabetic status, Increase Thrust	0.200	1
1015	Increase Thrust	Dizzyness, Shortness Breath Rate, Frequent Urir	0.200	1
1016	Dizzyness, Increase Thrust	Shortness Breath Rate, Frequent Urination at Ni	0.200	1
1017	Shortness Breath Rate, Increase Thrust	Dizzyness, Frequent Urination at Night, Diabetic	0.200	1
1018	Dizzyness, Shortness Breath Rate, Increase Thrust	Frequent Urination at Night, Diabetic status	0.200	1
1019	Frequent Urination at Night, Increase Thrust	Dizzyness, Shortness Breath Rate, Diabetic stat	0.200	1
1020	Dizzyness, Frequent Urination at Night, Increase Thrust	Shortness Breath Rate, Diabetic status	0.200	1
1021	Shortness Breath Rate, Frequent Urination at Night, Increase Thrust	Dizzyness, Diabetic status	0.200	1
1022	Dizzyness, Shortness Breath Rate, Frequent Urination at Night, Increase T	Diabetic status	0.200	1
1054	Dizzyness, Shortness Breath Rate	Diabetic status, Diabetic Family Member, Increas	0.200	1
1057	Shortness Breath Rate, Diabetic Family Member	Dizzyness, Diabetic status, Increase Thrust	0.200	1

Figure 6: Rules mined using FP-Growth.

For making Decision tree using diabetes dataset in RapidMiner, in main process Data Retrieve operator is used to fetch dataset from the repository. And then values of the attributes in the dataset are normalized using Normalize operator and then Discretize operator is used with it to get better and clean values to create decision tree.

And as sub process xValidation operator is used encapsulating ID3 Decision Tree operator (which creates a decision tree and is implementation of ID3 algorithm in RapidMiner) learner and performance operator. ID3 operator was used in Information Gain mode to make decision tree.

The test option chosen was 10-fold cross-validation.

Cross-validation is a method of estimating the accuracy of a classification and it works as follows:

1. Separate data in to fixed number of partitions (or folds)
2. Select the first fold for testing, whilst the remaining folds are used for training.
3. Perform classification and obtain performance metrics.
4. Select the next partition as testing and use the rest as training data.
5. Repeat classification until each partition has been used as the test set.
6. Calculate an average performance from the individual experiments.

10-fold cross validation was chosen because experiments have shown that this is the best choice to get an accurate estimate [14].

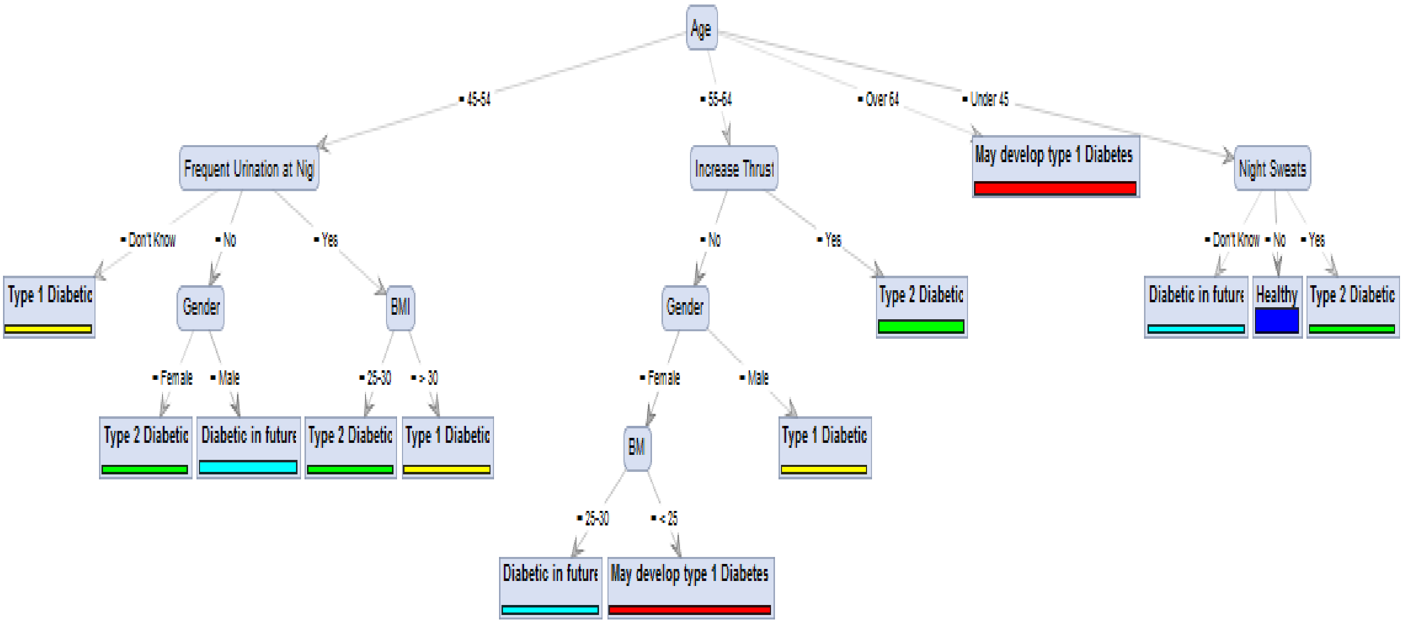
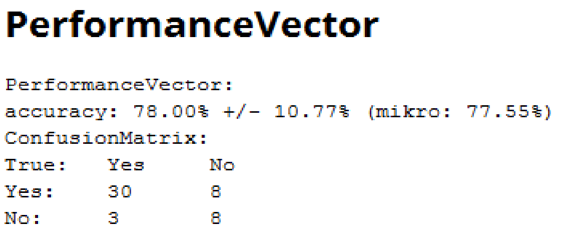


Figure 7: Generated Decision Tree using ID3 algorithm in RapidMiner.

Performing classification can be performed using the above decision tree as:

If Age 45-54 and Frequent Urination at Night = Yes and gender = Female
Then
She has a risk of getting Type 2 Diabetes.

The results of our experimental analysis for finding significant patterns for risk of getting diabetes have been presented.
The goal is to have high accuracy.



PerformanceVector

```
PerformanceVector:
accuracy: 78.00% +/- 10.77% (mikro: 77.55%)
ConfusionMatrix:
True:    Yes     No
Yes:     30      8
No:      3       8
```

Figure 8: Shows the confusion matrix for the Decision Tree.

Below equation is a formula of this confusion matrix in calculating the classifier's accuracy .

$$Accuracy = \left(\frac{\mathrm{TP + TN}}{\mathrm{TP\ +\ FP\ +\ TN\ +\ FN}}\right) \times 100$$

- True Positive (TP): The Total percentage of members objects classified as Class X actually belongs to Class X.
- False Positive (FP): The Total percentage of members objects of Class X but does not belong to Class X.
- False Negative (FN): The total percentage of members objects of Class X incorrectly classified as not belonging to Class X.
- True Negative (TN): The total percentage of members objects which does not belong to Class X and are classified not as a part of Class X.

VII. Conclusion

Data related to health care is very huge in nature and it keeps on increasing exponentially on daily basis. This data can arrive from different sources and all of them not completely suitable in quality and structure. Now a days, the efficient use of knowledge and experience of many specialists and health analysis data of patients collected in a database during the diagnosis procedure has been recognized on a large scale. In this paper we have represented an efficient approach segregating and extracting significant patterns from the diabetes data for the efficient prediction of the risk of getting diabetes.
So, if a person has some early symptoms of diabetes then by using Frequent Patterns other symptoms could be predicted. And by using the extracted sequence and applying the classification through decision tree, the risk of getting diabetes could be predicted. And the respective medical measures could be taken for that person at a very early stage.

In our future work, we have planned to design and develop an approach to predict risk of getting diabetes from the food and eating habits and physical work culture of people.

VIII. References

[1] Md. Tabrez Nafis, “Data Mining of Web Access Logs Using Classification Techniques” in Vol. 2 Issue VIII, (IJRAS ET), 2014, p-1.

[2] Hnin Wint Khaing, “Disease Forecasting System Using Data Mining Methods”. IEEE- International Conference on Intelligent Computing Applications, 2014, pp. 1–5.

[3] Gunasekar Thangarasu, Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques . PETRONAS 2011.

[4] Sellappan Palaniappan, “Intelligent Heart Disease Prediction System Using Data Mining Techniques,” Malaysia, 2010.

[5] S. Stilou, “Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare,” IOS Press, in MEDINFO 2001.

[6] Dr. Bushra M. Hussan, “Data Mining based Prediction of Medical data Using K-means algorithm”

[7] Harleen Kaur, “THE IMPACT OF DATA MINING TECHNIQUES ON MEDICAL DIAGNOSTICS”, Data Science Journal, October 2006.

[8] Ujma Ansar, Jyoti Soni “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” International Journal of Computer Applications, March 2011.

[9] http://patienteducation.stanford.edu, accessed on 10-03-2017, 11:20 pm

[10] self-management@stanford.edu , accessed on 10-03-2017, 11:47 pm

[11] http://www.idf.org/diabetes-prevention, , accessed on 11-03-2017, 8:27 pm.

[12] http://www.anderson.ucla.edu, accessed on 21-04-2017, 8:41 pm.

[13] Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare S. Stilou1 , P.D. Bamidis1,2, N. Maglaveras1 , C. Pappas1

[14] Asma A. AlJarullah, “Data miningDecision Tree Discovery for the Diagnosis of Type II Diabetes”, 2011.

[15] Patcharaporn Panwong and Natthakan lam-On, Predicting Transitional Interval of Kidney Disease Stages 3 to 5 Using Data Mining Method, 2016.

[16] https://en.wikipedia.org/wiki/RapidMiner, , accessed on 20-04-2017, 8:06 pm.