International Journal of Computer Applications Technology and Research
Volume 4– Issue 4, 291 - 295, 2015, ISSN:- 2319–8656

Android Application to Predict and Suggest Measures for Diabetes Using DM Techniques

V.Krishna Priya
Rajalakshmi Engineering College
8/14 F Kumar Brindavan Flats
43rd street nanganallur, ch-61
Chennai, India

A.Monika
Rajalakshmi Engineering College
2, Flowers road 4th lane,
Pursaiwakkam,ch-84
Chennai, India

P.Kavitha
Rajalakshmi Engineering College
819, Rajiv Gandhi St.,
Nazarathpet, Poonamalle.
Chennai, India

Abstract: Data Mining is an analytic process designed to explore data in search of consistent patterns and systematic relationships between variables, and then to validate the results by applying the patterns found to a new subset of data. Data mining is often described as the process of discovering patterns, correlations, trends or relationships by searching through a large amount of data stored in repositories, databases, and data warehouses. Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood [3] glucose (blood sugar), either because insulin production is insufficient, or because the body's cells do not respond properly to insulin, or both. This project helps in identifying whether a person has diabetes or not, if predicted diabetic[4] the project suggest measures for maintaining normal health and if not diabetic it predicts the risk of getting diabetic. In this project Classification algorithm was used to classify the Pima Indian diabetes dataset. Results have been obtained using Android Application.

Keywords: Android Application, Diabetes, Data Mining

1. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases[1,12], is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Diabetes has become a most common disease in today's world. So for every individual it is important to take a precautionary measure to check if the person has any chances of getting diabetes. For this purpose we use data mining techniques to predict if a person is diabetic or not. It is attractive as the results are obtained through an android application installed in mobile device. The main reason for accuracy of results is that only most significant attributes causing diabetes are considered for analysis

Data mining tools [8] predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [7]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Diabetes (diabetes mellitus)[24] is classed as a metabolism disorder. Metabolism refers to the way our bodies use digested food for energy and growth.. Most of what we eat is broken down into glucose. Glucose is a form of sugar in the blood - it is the principal source of fuel for our bodies.

A person with diabetes has a condition in which the quantity of glucose in the blood [19] is too elevated (hyperglycemia). This is because the body either does not produce enough insulin [5], produces no insulin, or has cells that do not respond properly to the insulin the pancreas produces. This results in too much glucose building up in the blood. This excess blood glucose eventually passes out of the body in urine [19]. So, even though the blood has plenty of glucose, the cells are not getting it for their essential energy and growth requirements

2. DATASET

Dataset is composed of 768 instances. Each patient is characterized in data set by 8 attributes. All attributes are numerical values. This attributes are: Diastolic blood pressure, plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function ,age (years), Class variable (0 or 1)[21].

3. CLASSIFICATION ALGORITHM

Classification [22] consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm [16] tries to discover relationships between the attributes that would make it possible to predict the outcome. Decision tree [3] builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes [6]. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees [3] can handle both categorical and numerical data.

3.1 C4.5 algorithm

The C4.5[18] algorithm constructs the decision tree with a divide and conquer strategy.

In C4.5algorithm, each node in the tree is associated with a set of cases. And also each cases are assigned with weights to take into account the unknown attributes values [17].

At the beginning, only the root is present, associated with the whole training set T S and with all case weights equal to 1:0. At each node the following divide and conquer method, the algorithm is executed, trying to find the locally best choice, with no backtracking allowed.

Let T be the set of cases associated at the node. The weighted frequency f req(Ci ; T) is computed (step (1)) of cases in T whose class is Ci , for i 2 [1; NC lass]. If all cases (step (2)) in T belong to a same class Cj (or the number of cases in T is less than a certain value) then the node is a leaf, with associated class Cj (resp., the most frequent class). The classification [22] error of the leaf is the weighted sum of the cases in T whose class is not Cj (resp., the most frequent class). If T contains cases belonging to two or more classes (step (3)), then the information gain of each attribute is calculated. For discrete attributes, the information gain is relative to the splitting of cases in T into sets with distinct attribute values. For continuous attributes, the information gain is relative to the splitting of T into two subsets, namely cases with attribute value not greater than and cases with attribute value greater than a certain local threshold, that is determined during information gain calculation [3].

The attribute [17] with the highest information gain (step (4)) is selected for the test at the node. Moreover, in case a continuous attribute is selected, the threshold is computed (step (5)) as the greatest value of the whole training set that is below the local threshold.

A decision node has s children if T1;Ts are the sets of the splitting produced by the test on the selected attribute (step (6)). Obviously, s = 2 when the selected attribute is continuous, and s = h for discrete attributes with h known values.

For i = [1; s], if Ti is empty, (step (7)) the child node is directly set to be a leaf, with associated class the most frequent class at the parent node and classification[22] error 0.

If Ti is not empty, the divide and conquer approach consists of recursively applying the same operations (step (8)) on the set consisting of Ti plus those cases in T with unknown value of the selected attribute. Note that cases with unknown value of the selected attribute are replicated in each child with their weights proportional to the proportion of cases in Ti over cases in T with known value of the selected attribute.

Finally, the classification error (step (9)) of the node is calculated as the sum of the errors of the child nodes. If the result is greater than the error of classifying all cases in T as belonging to the most frequent class in T , then the node is set to be a leaf, and all sub-trees are removed[3].

4. USAGE OF SIGNIFICANT ATTRIBUTES

The attributes used for calculation is the significant attributes done using Attribute Selection algorithm of WEKA[9] tool. The most significant attributes are plasma, body mass index, diabetes pedigree function, insulin level. These are the most significant attributes for the prediction of diabetes status of a person. The class attribute of the dataset specifies class 0 i.e not diabetic and class 1 i.e diabetic.

- **Not All Attributes Are Equal**

Whether you select and gather sample data yourself or whether it is provided to you by domain experts, the selection of attributes is critically important. It is important because it can mean the difference between successfully and meaningfully modeling the problem and not.

- **Misleading**

Including redundant attributes can be misleading to modeling algorithms. Instance-based methods such as k-nearest neighbor use small neighborhoods in the attribute space to determine classification and regression predictions. These predictions can be greatly skewed by redundant attributes.

- **Overfitting**

Keeping irrelevant attributes in your dataset can result in overfitting. Decision tree algorithms like C4.5 seek to make optimal spits in attribute values. Those attributes that are more correlated with the prediction are split on first. Deeper in the tree less relevant and irrelevant attributes are used to make prediction decisions that may only be beneficial by chance in the training dataset. This overfitting of the training data can negatively affect the modeling power of the method and cripple the predictive accuracy.

It is important to remove redundant and irrelevant attributes from your dataset before evaluating algorithms. This task should be tackled in the Prepare Data step of the applied machine learning process.

- **Feature Selection**

Feature Selection or attribute selection is a process by which you automatically search for the best subset of attributes in your dataset. The notion of "best" is relative to the problem you are trying to solve, but typically means highest accuracy.

A useful way to think about the problem of selecting attributes is a state-space search. The search space is discrete and consists of all possible combinations of attributes you could choose from the dataset. The objective is to navigate through the search space and locate the best or a good enough combination that improves performance over selecting all attributes.

Three key benefits of performing feature selection on your data are:

Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.

Improves Accuracy: Less misleading data means modeling accuracy improves.

Reduces Training Time: Less data means that algorithms train faster

5. SYSTEM ARCHITECTURE

The system architecture (see Figure 1) describes the flow of the project work. The first step in the process is the collection of data needed for the work. Here the dataset used is Pima Indian diabetes dataset[21], which is collected in the first step. The next step in the process is preprocessing of the data[4]. Here we covert the raw data into understandable format. Now the preprocessed data is classified into a decision tree[3] to predict the status of a person whether diabetic or not using the algorithm(C4.5)[25]. The user enters the details to know his results for the test into an android app[22] installed in his mobile device. The attributes entered by the user is compared with the decision tree and the results are generated.

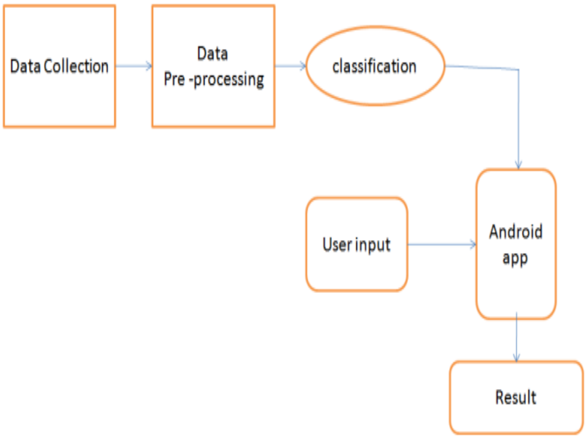


Figure 1. System Architecture

6. SCREENSHOTS

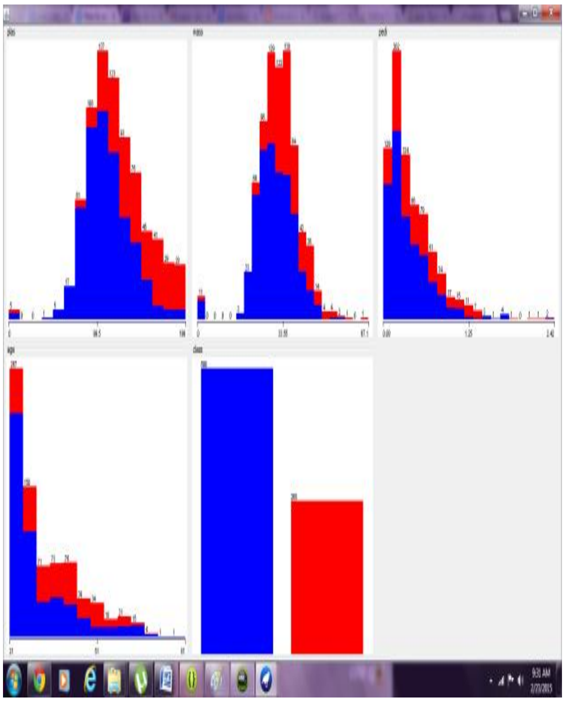
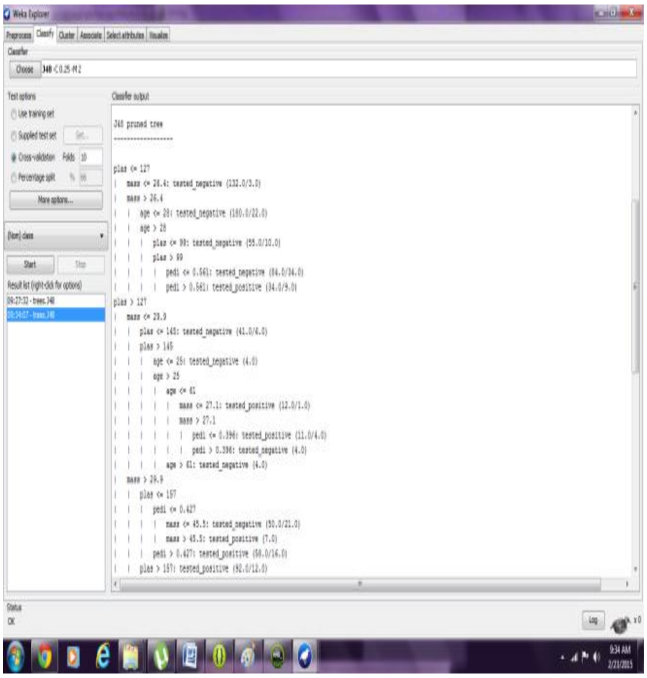


Figure 2. Pre-Processing

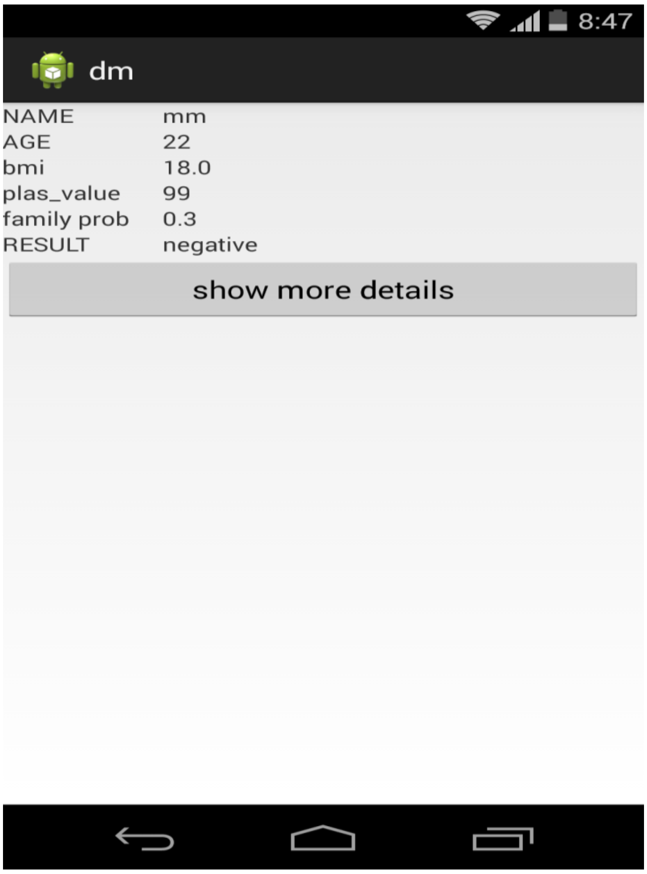


```
J48 -C 0.25 -M 2
```

Cross-validation Folds 10

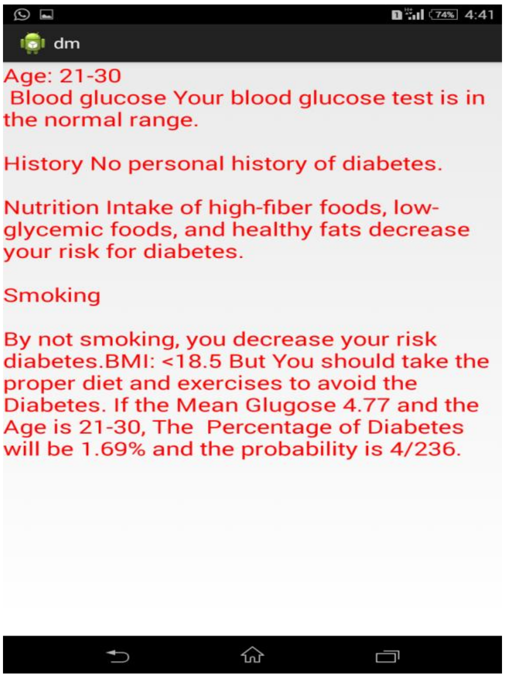
```
J48 pruned tree
------------------

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561: tested_positive (34.0/9.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pedi <= 0.396: tested_positive (11.0/4.0)
|   |   |   |   |   |   pedi > 0.396: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pedi <= 0.427
|   |   |   |   mass <= 45.5: tested_negative (50.0/21.0)
|   |   |   |   mass > 45.5: tested_positive (7.0)
|   |   |   pedi > 0.427: tested_positive (58.0/16.0)
|   |   plas > 157: tested_positive (92.0/12.0)
```

Figure 3. Decision tree



NAME	mm
AGE	22
bmi	18.0
plas_value	99
family prob	0.3
RESULT	negative

Figure 4. Result after entering details in Android application



Age: 21-30
Blood glucose Your blood glucose test is in the normal range.

History No personal history of diabetes.

Nutrition Intake of high-fiber foods, low-glycemic foods, and healthy fats decrease your risk for diabetes.

Smoking

By not smoking, you decrease your risk diabetes.BMI: <18.5 But You should take the proper diet and exercises to avoid the Diabetes. If the Mean Glugose 4.77 and the Age is 21-30, The Percentage of Diabetes will be 1.69% and the probability is 4/236.

Figure 5. Suggestive measures for user

7. CONCLUSION

The discovery of knowledge[9] from datasets is important in order to make effective diagnosis. The aim of data mining is to extract information stored in dataset and generate clear and understandable patterns. This study aims at the discovery of a decision tree model for the prediction of diabetes. Pre-processing is used to improve the quality of data. While preprocessing[6], the significant attributes of the dataset are considered for prediction of diabetes. This is an important factor for consideration. The decision tree algorithm[14] used for classification also produces maximum accuracy when compared to other algorithms of classification. Finally the results of the system are obtained in an android application which is very useful for the present generation.

8. FUTURE SCOPE

In future this system can designed for any prediction of any other disease such as cancer, thyroid, lung diseases etc., if these an android application of such disease prediction would be of great use in the near future. Another future enhancement would be to reduce the no of attributes considered for the prediction purpose. Considering less no of attributes and produce more accurate results is needed as an enhancement for the existing system.

9. ACKNOWLEDGMENT

Sincere thanks to our guide Mrs.P.Revathy who contributed to the development of this paper

10. REFERENCES

[1] P. Yasodha, M. Kannan, “Analysis of a Population of Diabetic Patient Databases in Weka Tool”, International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011

[2] WEKA, by university of Waikato, http://www.cs.waikato.ac.nz/ml/weka/

[3] T. Mitchell, "Decision Tree Learning", in T.Mitchell, Machine Learning (1997) the McGraw- Hill Companies, Inc., pp.

[4] Han, J., Kamber, M.: Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers (2000).

[5] Gloria L.A. Beckles and Patricia E. Thompson-Reidy the authors of“ Diabetes and Women’s Health Across the Life Stages”.

[6] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining Concepts and Techniques” Third edition .

[7] Folorunso O and Ogunde A. O (2004), “Data Mining as a Technique for Knowledge

[8] Management in Business Process Redesign” The Electronic Journal of Knowledge Management Volume 2 Issue 1, pp 33-44

[9] P.Yashoda, M.Kanan, Analysis of a population of diabetic patients databases in WEKA tool, IJSER, vol2, issue5, may 2011.

[10] Mukesh kumari, Dr. Rajan Vohra ,Anshul arora Prediction of Diabetes Using Bayesian Network (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178

[11] M. Khajehei, F. Etemady, "Data Mining and Medical Research Studies," cimsim, pp.119-122, 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, 2010

[12] Kaur H, Wasan SK,” Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science,2(2):194-200,2006

[13] Analysis of a Population of Diabetic Patients Databases with Classifiers using c4.5 Algorithm” World Academy of Science, Engineering and Technology International Journal of Medical, Pharmaceutical Science and Engineering Vol: 7 No: 8, 2013.

[14] Margaret H. Dunham,-“Data Mining Techniques and tice hall publishers

[15] P. Radha , Dr. B. Srinivasan Predicting Diabetes by cosequencing the various Data Mining Classification Techniques IJISET - InternationalJournal of Innovative Science

[16] E.Knorr.E and R.Ng, “Algorithms forming distance -based outliers in large datasets”, in proceedings of 1998 International Conference on Very Large Data Bases (VldB’98), pp. 392-403 New York, 1998.

[17] E.Jiawei Hen and Micheline Kamber “DataMining Concepts and Techniques”, *CA:Elsevier Inc,SanFrancisco*, 2006

[18] U.M.Piatetsky-Shapiro and G.Smyth “From DataMining to Knowledge Discovery : An Overview”,1996, pp.1 -36

[19] S.C.Liao & M.Embrenchts, “Data Mining techniques applied to medical information”, *Med.Inform*, 2000, pp.81 102.

[20] L.Breiman, J.Friedman, J.Olsen C.Stone, “Classification and Re-gression Trees”, *Chapman & Hal, 1984, 122-134. Engineering & Technology, Vol. 1 Issue6, August 2014.*

[21] *https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes UCI MACHINE LEARNING REPOSITORY*

[22] Szakacs-Simon, P. Dept. of Autom., “Transilvania” Univ., Brasov, Romania Moraru, S.A. ; Perniu, L.Android application developed to extend health monitoring device range and real-time patient tracking International Journal of Advanced Research in Computer Science and Software Engineering

[23] Rohanizadeh.s “A proposed data mining methodogy application to industrial procedures”

[24] en.wikipedia.org/wiki/Diabetes_mellitus