First International Conference on Computer Engineering
International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
Volume 5 Issue 8, November-December-2020

# Survey Data Mining in a Diabetes Patient Database using WEKA Tool

**Shivani Patel[1], Sanjay Chaudhary[2], Prakashsingh Tanwar[2]**

[1]Research Scholar, Computer Science and Information Technology, Madhav University, Pindwara (Sirohi) Rajasthan, India

[2]Research Supervisor, Computer Science And Information Technology, Madhav University, Pindwara (Sirohi) Rajasthan, India

## ABSTRACT

Data mining tools play a significant role in the healthcare sector. As medical records systems become more standardized, data quantity increases with much of it going unanalyzed. Taking into account the prevalence of diabetes the study is aimed at finding out the characteristics that determine the presence of diabetes. In this research, WEKA an open-source data mining tool is used for the analysis of diabetes database. Classification techniques are applied to classify the data and the data is evaluated using 10-fold cross validation and the results are compared.

Keywords : Diabetes database, Machine Learning, Classification, Evaluation, Weka Toolkit

## I. INTRODUCTION

Computer-assisted care is becoming increasingly important in healthcare. There is no other domain with as many inventive developments that have such a large societal impact. There is already a long history of computer-based decision support in medicine, dealing with complicated challenges such as illness diagnosis, administrative choices, and aiding in the prescription of suitable therapy. Diabetes, an incurable chronic condition, is one of the useful applications in the field of medicine. It is a set of metabolic illnesses in which a person's blood sugar levels are elevated, either because the body does not make enough insulin or because cells do not respond to the insulin that is produced. Data mining is used to discover interesting patterns that can aid in the critical duties of medical diagnosis and therapy.

Data Mining seeks to extract knowledge from data and deliver it in a way that people can readily compress. It is the process of analysing data from several angles and synthesising it into meaningful knowledge. To tackle data mining difficulties, open data mining technologies are employed. WEKA, an open-source data mining tool, is utilised in this study to analyse a diabetic database. To categorise the data, classification algorithms are used, and the data is tested using 10-fold cross validation and the results are compared.

A supervised model that classifies a data item into one of many predetermined classifications is known as classification. The categorization of data is a two-step process:

1. A model is created to describe a predefined collection of data types or ideas.
2. The model is used to classify data.

The purpose of this study is to use the WEKA machine learning software platform to analyse the

Pima Indian Diabetes dataset. The purpose of this study is to find connections in Diabetes data in order to categorise it more effectively. The classification's purpose is to establish a class that will accurately find previously unseen records.

The data set for this study was gathered from the UCI repository and consists of 768 instances with 9 distinct characteristics [9]. The use of data mining methods will be investigated in order to determine the most efficient categorization of a diabetic dataset. The goal of this study is to use WEKA as a classification tool to learn about the practical elements of machine learning theory.

## II. RELATED WORK

In the subject of medical diagnostics, there is a lot of data mining study. It's a good idea to start by reviewing current developments in the field of the planned study.

Researchers have used Decision Tree and Incremental Learning to investigate the early stages of cardiac and diabetic symptoms [1]. The data set was collected from patient files kept in the medical records division of the BGS Hospital in Bangalore. A classification algorithm analyses a dataset, and the classifier or learner model is expressed as classification rules or decision trees.

Using data mining techniques, researchers have built a decision assistance system for sickness detection [2]. On the Diabetes dataset, three artificial intelligence classifier techniques were used: Multilayer Perceptron, Naive Bayes Classifier, and J.48. These classifiers are often employed in the disciplines of data mining, biomedical engineering, and medical diagnosis.

Researchers used data from a hospital repository to classify diabetes patients' data, which consisted of 249 cases with seven distinct variables [3]. The Dataset's examples are divided into two categories: blood tests and urine testing. WEKA tool is used to classify the data and the data is evaluated using 10-fold cross validation.

Data mining approaches were described by the authors in order to analyse a dataset and determine the significance of categorization test data [4]. Their research demonstrates how WEKA analyses file conversions and selects attributes to mine, in comparison to Knowledge Extraction of Evolutionary Learning, which not only analyses data mining classifications but also genetic, evolutionary algorithms, is the most efficient learning method.

[5] Researchers have given an intelligent recommended paradigm for diabetic patient treatment. This web-based data mining application provides several benefits in well-equipped hospitals, including resource efficiency and patient illness prediction.

The authors conducted a thorough evaluation of data-mining methods' uses in diabetes research. For this study, the MEDLINE database from PubMed was employed [6]. Around 20 works in the linked topic were evaluated by the authors. The objective of the study, the group/topic of research, the kind of diabetes, the data set used, the data-mining methods employed, the data-mining software and technology used, and the outcome of the data-mining application are all collected from the publications and given.

The authors provided a comparative analysis of open-source data mining tools, focusing on WEKA's critical function in contrast to other tools and its use in a real-world scenario [7]. The frequently overlooked pre-processing and post-processing processes in knowledge discovery have been shown to be the most important factors in determining the success of a real-world data mining application [8]. A semi-automatic data cleaning system is used to solve the problem of noisy data.

## III. Research Framework

Data mining is the process of extracting knowledge from data using a computer-based technique [12]. Figure 1 depicts a diagrammatic depiction of the suggested study framework.

- Data collection and preparation are the first steps in the approach used to accomplish the research challenge. The Pima Indians Diabetes Database from the UCI Machine Learning Repository was utilised as the training dataset for data mining. The data preparation step included all tasks necessary to create the final dataset from the raw data.
- Various modelling approaches are chosen and used, and their parameters are calibrated to the best possible values. The same data mining problem is tackled using classification algorithm approaches. A portion of the dataset is translated into the needed format based on the requirements.
- To choose the most efficient algorithm, comparative research of algorithms for distinct models is conducted. The evaluation process determines how well the model matches the goals.

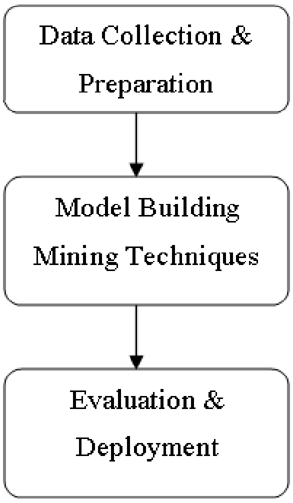

Data Collection & Preparation
Model Building Mining Techniques
Evaluation & Deployment

Figure 1. Research Framework

The classification's goal is to assign a class that will help you discover previously unseen records as precisely as possible. The goal is to create a classification model for class characteristics that uses a test set to assess the model's accuracy. Training and test sets are created from the provided data source. The model was built using the training data, and the test set was utilised to validate it.

## IV. IMPLEMENTATION USING WEKA

Data mining is a technique for making sense of enormous volumes of mainly unsupervised data in a certain topic. Data is mapped into predetermined groups by classification. Because the classifications are selected before the data is examined, it is commonly referred to as supervised learning. During the analysis of the diabetes dataset, two groups are formed based on the data attribute value: "tested positive" and "tested negative."

### *A. Data Collection*

The Pima Indians Diabetes Database from the UCI Machine Learning Repository [9] was utilised as the training dataset for data mining. There are 768 record samples in the collection, each with eight properties. The following are the attributes of this dataset:

1. Number of times pregnant
2. Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (kg/m)^2)
7. Diabetes pedigree function
8. Age (years)

### *B. Data Pre-processing*

Prior to mining, data processing techniques were used to significantly increase the overall quality of the patterns mined as well as the time necessary for the actual mining [10]. Because quality judgments must be founded on quality data, data pre-processing is an important phase in the knowledge discovery process. The dataset has 768 instances in total. Some of these examples include fictitious values, such as "age is 0, BMI is 0." The dataset has been cleansed of such fictitious records. Table 1 displays the number of records that were eliminated as part of the data cleansing process.

Table 1. Dataset Pre-processing

| Sr. No | Pre-processed Attributes (value zero) | No. of records deleted |
|---|---|---|
| 1 | Blood pressure in the diastole | 35 |
| 2 | BMI glucose tolerance test | 5 |
| 3 | The thickness of the triceps skin folds | 4 |
| 4 | 2-Hour serum insulin | 192 |
| 5 | Blood pressure in the diastole | 140 |

Following the deletion of these instances, 392 cases remained with no missing values. The statistical features of qualities are explained in Table 2.

Table 2: Properties of attributes after data pre-processing

| Attribute No. | Minimum & Maximum | Mean Deviation | Standard Deviation |
|---|---|---|---|
| 1 | 0-17 | 3.301 | 3.211 |
| 2 | 56-192 | 122.628 | 30.861 |
| 3 | 24-110 | 70.663 | 12.496 |
| 4 | 7-63 | 29.145 | 10.516 |
| 5 | 14-846 | 156.056 | 118.842 |
| 6 | 18.2-67.1 | 33.086 | 7.028 |
| 7 | 0.085-2.42 | 0.523 | 0.345 |
| 8 | 21-81 | 30.865 | 10.201 |

### *C. Model Building - Classification Algorithm Execution*

On the basis of a training set of data comprising observations whose category membership is known, classification is the issue of determining which of a set of categories a new observation belongs to [11]. The training examples are used to develop a model that can categorise data samples into well-defined categories.

Different categorization approaches are used on the Pima Indians Diabetes Database in this study, and the data is reviewed using 10-fold cross validation, with the findings compared. This means that each classifier training phase contains 100 steps. The class label is predicted by the classification method. The ultimate product will be patterns that may be utilised to determine whether or not a person has diabetes. This is explained in fully in Table 3.

### *D. Model Evaluation*

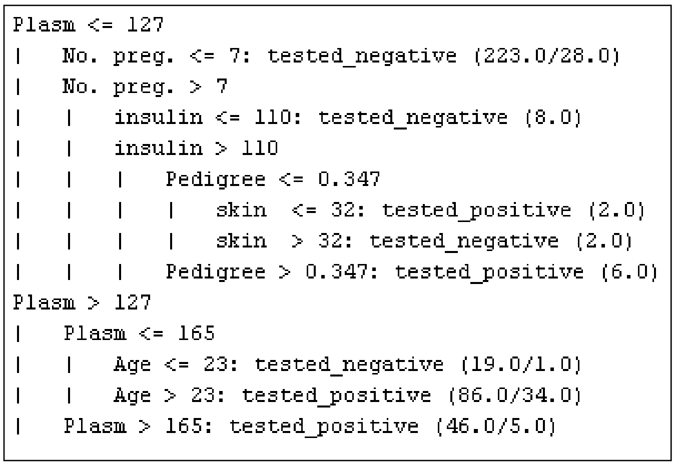
The Pima Indians Diabetes Data Set was subjected to twelve different classifiers in this study. The percentage of test set tuples properly categorised by a classifier on a given test set is known as its accuracy. Table 3 lists the performance indicators.

Ten-cross validation training was used to test each of the classifiers. J48 was shown to be the strongest predictor in this simulation.

J48 constructs a decision tree from a data set utilising information gain and analyses the effects of selecting an attribute for data splitting. The method then repeats on smaller selections. When all instances in a subset belong to the same class, the splitting method comes to an end. The leaf node of a decision tree is then built, instructing the user to select that class. Figure 2 shows a trimmed J48 tree.

Table 3. Classification Algorithm Execution

| Sr. No | Classification technique | Accuracy | Time |
|---|---|---|---|
| 1 | NAÏVE BAYES | 77.80 | 0 |
| 2 | K-Star | 71.17 | 0 |
| 3 | Random Tree | 75 | 0.02 |
| 4 | OneR | 75.76 | 0 |
| 5 | ZeroR | 66.83 | 0 |
| 6 | Decision Table | 77.8 | 0.06 |
| 7 | Multilayer Perceptron | 74.23 | 1.09 |
| 8 | Simple Logistic | 77.8 | 0.31 |
| 9 | SimpleCart | 76.02 | 0.09 |
| 10 | Hyper Pipes | 66.32 | 0 |
| 11 | JRip | 77.2 | 0.05 |
| 12 | J48 | 79.33 | 0 |



```
Plasm <= 127
|   No. preg. <= 7: tested_negative (223.0/28.0)
|   No. preg. > 7
|   |   insulin <= 110: tested_negative (8.0)
|   |   insulin > 110
|   |   |   Pedigree <= 0.347
|   |   |   |   skin  <= 32: tested_positive (2.0)
|   |   |   |   skin  > 32: tested_negative (2.0)
|   |   |   Pedigree > 0.347: tested_positive (6.0)
Plasm > 127
|   Plasm <= 165
|   |   Age <= 23: tested_negative (19.0/1.0)
|   |   Age > 23: tested_positive (86.0/34.0)
|   Plasm > 165: tested_positive (46.0/5.0)
```

Figure 2. J48-Pruned tree

## V. CONCLUSION AND FURTHER WORK

Data mining is a branch of traditional data analysis and statistical approaches that incorporates analytical techniques from a variety of disciplines, including but not limited to numerical analysis, pattern matching, and artificial intelligence areas like machine learning, neural networks, and genetic algorithms. In this study, an open-source data mining programme called WEKA was employed to analyse a diabetic database.

In this study, a data mining approach was used to classify Diabetes Clinical data and predict whether or not a patient will develop Diabetes. The Pima Indians Diabetes Database is subjected to several categorization methods, with the results summarised in a table. Using association mining, this research may be furthered. This portion of the dataset has been converted to the appropriate format. This project also includes the installation of many datasets.

## VI. REFERENCES

1) Ashwinkumar.U.M and Dr Anandakumar.K.R, "Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques", 2012 2nd International Conference on Computer Design and Engineering (ICCDE 2012), IPCSIT vol. 49 (2012) © (2012) IACSIT Press, Singapore
2) Murat Koklu, Yavuz Unal, "Analysis of a Population of Diabetic Patients Databases with Classifiers", International Journal of Medical Science and Engineering World Academy of Science, Engineering and Technology, vol. 7, no. 8, 2013, pp. 772-774
3) P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in Weka Tool", International Journal of Scientific & Engineering Research, vol. 2, no. 5, May 2011
4) Trilok Chand Sharma1, Manoj Jain, "WEKA Approach for Comparative Study Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 4, April 2013, pp. 1925-1931
5) MD. Ezaz Ahmed, Dr. Y.K. Mathur and Dr Varun Kumar, "Knowledge Discovery in Health Care Datasets Using Data Mining Tools", (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 3, no.4, 2012
6) Miroslav Marinov, Abu Saleh Mohammad Mosa, Illhoi Yoo, and Suzanne Austin Boren, "Data-Mining Technologies for Diabetes: A Systematic Review", Journal of Diabetes Science and Technology, Diabetes Technology Society, vol. 5, no. 6, Nov. 2011
7) M.Vijayakamal, Mulugu Narendhar, "A Novel Approach for WEKA & Study On Data Mining Tools", International Journal of Engineering and Innovative Technology (IJEIT), vol. 2, no. 2, Aug. 2012
8) Wynne Hsu, Mong Li Lee, Bing Liu and Tok Wang Ling, "Exploration Mining in Diabetic Patients Databases: Findings and Conclusions"
9) http://archive.ics.uci.edu/ml/datasets/Diabetes
10) Asha Gowda Karegowda , M.A. Jayaram, A.S. Manjunath, "Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients", International Journal of Engineering and Advanced Technology (IJEAT), vol. 1, no. 3, Feb, 2012, pp. 147-151
11) Jie Gao1, J¨org Denzinger, and Robert C. James, "A Cooperative Multi-agent Data Mining Model and Its Application to Medical Data on Diabetes"
12) Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Students Performance" International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, 2011