Data Mining Models Comparison for Diabetes Prediction

Amina Azrar[1], Muhammad Awais[3]
Department of Software Engineering
Government College University,
Faisalabad, Pakistan

Yasir Ali[2]
Department of Computer Science and Engineering
University of Engineering and Technology,
Lahore, Pakistan

Khurram Zaheer[4]
Department of Software Engineering
Government College University,
Faisalabad, Pakistan

***Abstract*—From the past few years, data mining got a lot of attention for extracting information from large datasets to find patterns and to establish relationships to solve problems. Well known data mining algorithms include classification, association, Naïve Bayes, clustering and decision tree. In medical science field, these algorithms help to predict a disease at early stage for future diagnosis. Diabetes mellitus is the most growing disease that needs to be predicted at its early stage as it is lifelong disease and there is no cure for it. This research is intended to provide comparison for different data mining algorithms on PID dataset for early prediction of diabetes.**

Keyword—Diabetes; data mining; classification; decision tree; Naïve Bayes; KNN

I. Introduction

Knowledge discovery in databases (KDD) is the system of applying data mining algorithms. Knowledge Discovery in Databases (KDD) is common research area for researchers in machine learning, databases, high performance computing, data visualization and knowledge-based systems. The primary steps for data mining include data selection, data preprocessing, data transformation, data mining, and final evaluation (pattern evaluation and pattern recognition).

Data Mining is the process of getting meaningful outcomes from any given dataset. Some of the techniques used for data mining include association rules, classification, clustering, Naïve Bayes, Decision Tree and KNN. A variety of rules can be generated using data mining techniques. Data Mining is useful for Prediction or Description of a few records. Using prediction, we are expecting unknown values of various variables in dataset whilst description specializes in coming across designs that depict the information translated by means of People.

Data mining is useful for predicting diseases. Affected person's history, Hospitals, clinical devices and electronic facts offer a lot of records concerning a selected disease. Those datasets are used for extracting useful information by which we are able to take choices and generate rules. Multiple diseases can be diagnosed using data mining methodologies, for example, AIDS and diabetes. This paper is meant to predict diabetes for pregnant women depending on few given attributes. Some major factors that affect the diabetes or may cause its increase in severity include obesity, weight increase or hypertension.

A. Diabetes

Diabetes mellitus is a common disease where there is too much sugar (glucose) floating around in your blood. This occurs because either the pancreas can't produce enough insulin or the cells in your body have become resistant to insulin. Diabetes affects the capability of human body to utilize the energy present in food. Basic types of diabetes are:

Type1 – In this type of diabetes pancreas does not produce adequate amount of insulin and in consequence the level of glucose in blood exceeds from typical range. Individuals suffering from this type diabetes are usually dependent on external insulin injected in body after regular intervals. It is caused by a genetic predisposition. Medical risks associated with this type of diabetes include diabetic retinopathy (eyes disorder), diabetic neuropathy (nerves disorder) and diabetic nephropathy (kidneys disorder). It counts for 95% diabetes cases.

Type2 – In Type 2 diabetes body is unable to consume the insulin properly due to insulin resistance. It is usually caused due to obesity and overweight children. It is non-insulin dependent and milder than Type 1 diabetes. It causes major effects on heart diseases and heart strokes. It cannot be cured but controlled with proper nutrition, exercise and weight management.

Gestational Diabetes – This type of diabetes includes married women who are not affected with diabetes according to previous medical history but high glucose level is diagnosed during/after pregnancy. According to the National Institutes of Health, the reported rate of gestational diabetes is between 2% to 10% of pregnancies.

II. Literature Survey

- Tawfik Saeed Zeki et al. [1] in their research presented an expert system for diabetes diagnosis. Their proposed expert system was rule based that have the structure of IF THEN. Transforming experts' knowledge to stated rules, they defined 3 stages that are handled by Block Diagram, Mockler Charts and Decision Tables. Total 6 states of diagnosis had been described (by the block diagram of diagnosis) using 5 attributes with different combinations for various diagnosis. After inspecting multiple factors, expert system provides diagnosis for disease. It was coded in VP-Expert as it is specified for developing expert systems.

- Seyedeh Talayeh Tabibi et al. [2] proposed an expert system for checkup and treatment of different types of diabetes. Expert system was developed in 10 stages. They took 3 attributes that include patient's condition, patient's information and different tests. Multiple combinations are generated on the basis of given attributes as it was rule-based expert system. Questions are generated relating to test and background relating to diabetes and suitable advice is generated depending on situation. It was developed in VP-Shell to code while they also obtained experts advice from medical specialists and nurses of diabetics' department.

- Vishali Bhandari and Rajeev Kumar [3] compared Mamdani-type and Sugeno-type fuzzy expert systems with the help of multiple parameters for diabetes diagnosis. MATLAB fuzzy logic toolbox is used for comparative study for both types of expert systems. Different resulting parameters showed that Sugeno-type expert system is more useful as it less computational and optimized while Mamdani-type expert system is not computationally powerful. Mamdani-type fuzzy expert system generates outcome using defuzzification and has outcome membership functions while Sugeno-type uses weighted average for outcome and has no outcome membership function. 5 parameters are used and results are generated that are compared afterword.

- Neeru Lalka and Sushma Jain [4] presented an expert system for diagnosis and medication for Type-I diabetes. Multiple parameters are used that include body mass index (bmi), plasma glucose level, minimum blood pressure and serum insulin level. Specified dosage for insulin intake is recommended based on few attributes. Probability of diagnosis uses five fuzzy numbers to show results and also accurately calculates probability to avoid hypoglycemic (low level of blood sugar) condition. Three fuzzy numbers are used to show output results. This expert system is only for Type-I diabetes.

III. Dataset

A. Pima Indians Diabetes Data set

Dataset contains records of females, having age at-least 21 years and living in Phoenix, Arizona, USA. Dataset contains labeled data having class attribute in binary (0 or 1). 0 value of class attribute represents negative test and 1 value represents the diagnosis of diabetes. Dataset contains 768 records of different patients in which 268 (34.9%) records are positive test cases which means '1' value of class attribute and 500 (65.1%) cases in class '0' representing negative test. The attributes of dataset are given in Table I with their description, type and units.

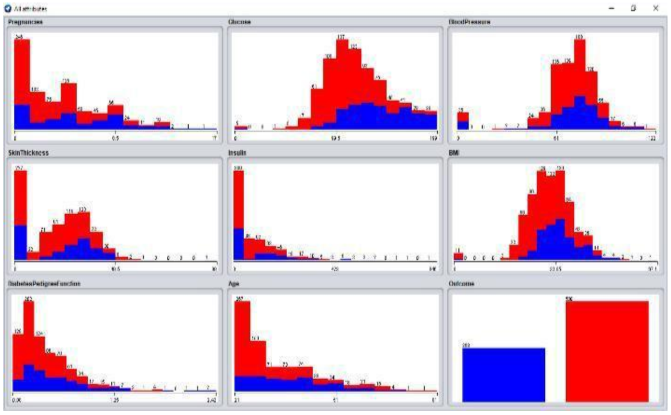


Fig. 1. Results of Class Label.

TABLE I. Dataset Description

Name	Description	Type	Unit
Pregnant	Number of pregnancies	Numeric	-
GTT	2-hour OGTT Plasma glucose	Numeric	mg/dl
Bp	Diastolic BP	Numeric	mmHg
Skin	Triceps Skin fold thickness	Numeric	Mm
Insulin	2-hour serum insulin	Numeric	Mm, U/ml
BMI	Body mass index (kg/m)	Numeric	Kg/m^2
DPF	Diabetes pedigree function	Numeric	-
Age	Age of Patient (years)	Numeric	-
Diabetes (Label)	Diabetes onset within 5 years (0,1)	Numeric	-

Visualization of dataset using WEKA presents the data distribution shown in figure. Fig. 1 shows 5 input variables and one outcome variable. Red color represents class label with negative results while blue color shows class label with values of labeled with negative result.

B. Preprocessing

Pre-processing consists of the steps of collection/cleaning, selection and transformation, data mining (integration and normalization) and last step is evaluation. Cleaning is used to fill the missing values in datasets using one-of-a-kind techniques like binning or replacing by mean or mode.

For pre-processing and applying few data mining algorithms, numerical data is converted to categorical data. Outcome is converted from integral data to categorical data with class labels as YES and NO while other categories are based on general items used and displayed in tables below (Tables II and III). Data for BP and Glucose is categorized on the basis of general categories used and ranges defined in below table.

TABLE II. BP LEVEL RANGES [6]

BP Level	Range
Low	Less than 80
High	80 to 100
Hypertension	Above 100

TABLE III. GLUCOSE LEVEL RANGES

Glucose Label	Range
Low	Less than 80
Normal	80 to 140
Early Diabetes	141 to 180
Diabetes	Above 181

IV. APPLIED ALGORITHMS

1) Decision Tree: Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It is tree like a graph used to display every possible outcome of a decision. It is most powerful classification algorithm used to predict possible outcome of a branch or tree. Classification is done by tree and leave nodes are generated on the basis of results on nodes in it.

Parameters on the dataset when applying DT set as criterion was gain ratio, maximal depth of the tree considered as 20. We also applied pruning as confidence=0.25 and pre-pruning techniques on DT as by setting minimal gain=0.1, minimal leaf size=2, minimal size of split=2 and number of pruning alternates considered as 3 in both datasets. We split data in DT as 70% training data and 30% as test data and apply model to show outcome and performance to check effectiveness and accuracy of both treatments.

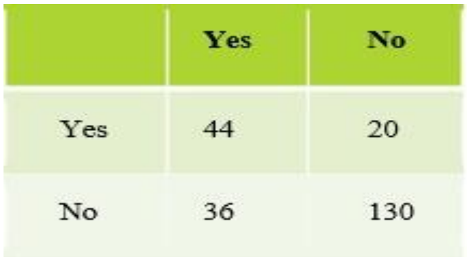
Result using information gain show the class precision of yes and no and the Accuracy of decision tree algorithm up to 75.65%.

$$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X)=P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

Fig. 2. Probability Calculation [5].

Results using Gain Ratio

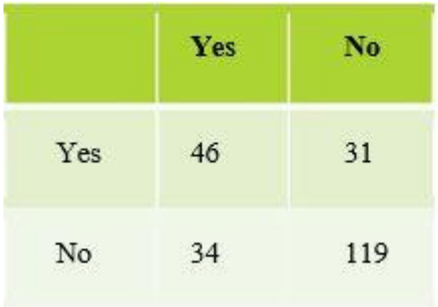


	Yes	No
Yes	44	20
No	36	130

2) Naïve Bayes: Naïve Bayesian is a well-known type of data mining classification technique. According to the definition it is a statistical technique which predicts the class of a new data record by estimation governed by the probabilities calculated from Bayesian rule formula. The Naïve Bayesian basic principle can be described as: Calculation of probability of Hypothesis that record belongs to class c given the new observed data record x.

Training process includes the calculation of marginal and conditional probabilities which are used in testing process for the calculation of probability of belonging of a new record to any class (Fig. 2).

Result using Naïve Bayes



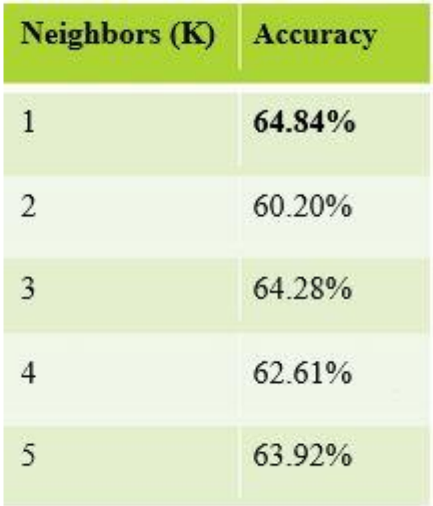
	Yes	No
Yes	46	31
No	34	119

Result using algorithm show the class precision of yes and no, the Accuracy of Naïve Bayes algorithm shows 71.74% and Distribution model for label attribute Outcome is:

- Class **Yes** (**0.349**) -> 8 distributions
- Class **No** (**0.651**) -> 8 distributions

3) K-Nearest Neighbor (KNN): K-Nearest Neighbor (KNN) is supervised learning algorithm used for classification of data. K means to select points from given dataset that how much data will be selected of nearest neighbor. This algorithm selects data on the basis of K value to nearest neighbor and decides that this point is similar to given sample. We apply KNN on dataset with K values ranging from 1 to 10. First, we make label to results of the treatment and split data into 70%, 30% as training and test records respectively, and then we make 10-fold of cross validation, also with 20 folds, by giving sampling as automatic to the split data value and apply KNN on the given data.

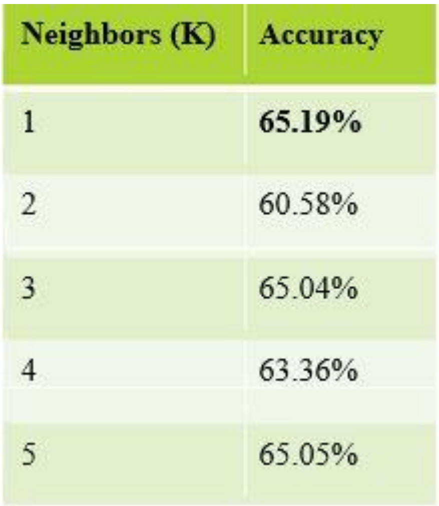
For 10 Folds (Fig. 3):



Neighbors (K)	Accuracy
1	**64.84%**
2	60.20%
3	64.28%
4	62.61%
5	63.92%

Fig. 3. Accuracy using 10-Fold.

For 20 Folds (Fig. 4):



Neighbors (K)	Accuracy
1	**65.19%**
2	60.58%
3	65.04%
4	63.36%
5	65.05%

Fig. 4. Accuracy using 20-Fold.

By applying K-nearest Neighbor algorithm, we find out several accuracies using 10 and 20-fold in KNN algorithm using 1 to 5 nearest neighbors on dataset. Using 1 nearest neighbor in 20-fold, highest accuracy can be seen.

V. Results

The results obtained from these 3 applied algorithms are different that as each algorithm worked on different technique. Results obtained from this dataset can be enhanced by applying more pre-processing techniques and data filtration. Accuracy obtained from Decision Tree is highest yet the graph is more dispersed that can be enhanced too. Lowest accuracy is from KNN. KNN is tested with wide range of K values from 1 to 10 and with changing folds from 10 to 20 but still accuracy is not that much. Pictorial representation of results is shown below in the form of graph.

Comparisons

Fig. 5 shows the accuracy rate of different data mining (DM) models.

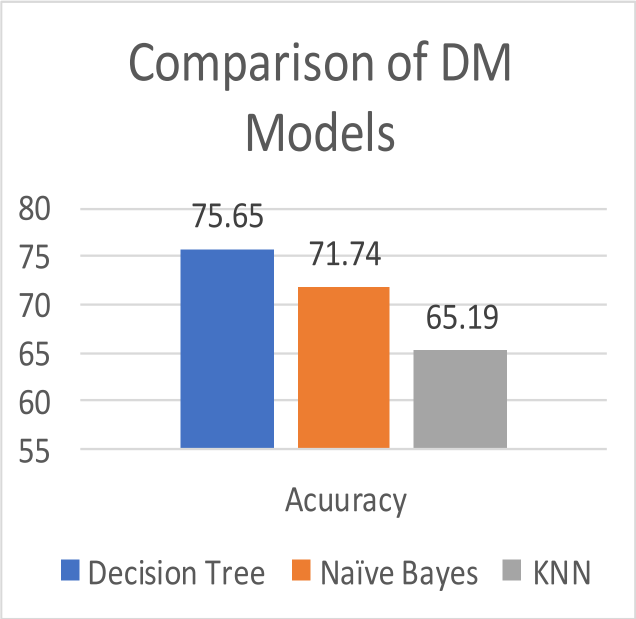


Fig. 5. Comparison of Different DM Models.

VI. Conclusions

The prevalence of diabetes is increasing among young adults and old age people. This paper focuses that the use of data mining algorithms can be very helpful in early prediction and in consequence early precautions before the diagnosis of disease. The main goal of this paper is to provide a comparison and suggest best algorithm which can be used for the pattern recognition or prediction in healthcare fields. These algorithms are of much importance for medical datasets because these algorithms can be used for automatic classification tools which can help doctors or experts for taking necessary steps for any disease before diagnosis. Each of these algorithms can give high accuracy and efficiency depending upon the type of data and attributes. After the implementations of these algorithms it can be said that for PID dataset Decision Tree gives best accuracy 75.65%. The tool used for testing and validation is Rapid Miner while all algorithms worked with 70:30 ratio for training and testing.

References

[1] Tawfik Saeed Zekia, Mohammad V. Malakootib, Yousef Ataeipoorc, S. Talayeh Tabibid. An Expert System for Diabetes Diagnosis. American Academic & Scholarly Research Journal Special Issue Vol. 4, No. 5, Sept 2012.

[2] Seyedeh Talayeh Tabibi, Tawfik Saeed Zaki, Yousef Ataeepoor. Developing an Expert System for Diabetics Treatment Advices. International Journal of Hospital Research 2013, 2(3):155-162.

[3] Vishali Bhandari and Rajeev Kumar. Comparative Analysis of Fuzzy Expert Systems for Diabetic Diagnosis. International Journal of Computer Applications (0975 – 8887) Volume 132 – No.6, December 2015.

[4] Neeru Lalka and Sushma Jain. Fuzzy Based Expert System for Diabetes Diagnosis and Insulin Dosage Control. International Conference on Computing, Communication and Automation (ICCCA2015).

[5] (Data Mining Map, an Introduction to Data Science, 2010-2018)

[6] (Blood Pressure UK, 2008).