DESIGN AND IMPLEMENTATION OF PREDICTIVE MODEL FOR PROGNOSIS OF DIABETES USING DATA MINING TECHNIQUES

Shuja Mirza
Scholar School of computer and system sciences
Jaipur National University
Rajasthan India

Dr. Sonu mittal
Associate Professor School of computer and system sciences
Jaipur National University
Rajasthan India

Dr. Majid zaman
Scientist D Directorate of IT&SS
University of Kashmir, Srinagar

Abstract: Purpose: The aim of this research is to design a predictive model using data mining tools and techniques that could be employed in prediction if diabetes, with the intention of enhancing the capability and efficiency of decision making.
Methods: The research was carried out on primary data that was collected from one of leading diagnostic centers in Srinagar (J&K). Data was preprocessed so as to remove inconsistencies'. Data mining techniques’ Naïve Bayes and support vector machine, K-nn were used for decision making. Data mining is a strong novel innovation for the extraction of concealed clairvoyant and significant facts from vast databases that can be utilized to increase thoughtful and novel insights. Utilizing superior data mining techniques to exhume vital knowledge is considered as a dissident way to enhance the quality and exactness of health services in order to improve healthcare service while bringing down the services cost and time.
Findings: The study demonstrated that model created using SVM outperformed Naïve Bayes and K-nn in predicting the disease.

Keywords: Data mining, Decision support system, SVM, Naïve Bayes, K-nn, Diabetes, weka.

1. INTRODUCTION

Data mining is one of the young and promising fields that have gotten consideration of the scientists in both industry and scholastics. “Knowledge discovery in databases” a synonym used by many for term data mining is a technique aimed at revealing significant and interesting patterns from data. Data mining is a well-defined characterized procedure of inimitable distinctive steps. It was Frawley et.al in 1996[1] who gave formal definition of knowledge discovery in databases as” Data Mining is the non-trivial extraction of implicit previously and potentially useful information about data”. Different sorts of data mining tools and procedures are available which allow specialists to foresee behavior and patterns in data for taking dynamic decisions. Utilization of data mining has been effective in various fields of life, for example, in marketing, Banking, Business etc. however its trend in healthcare application is conglomering with each passing day. Diabetes is the lifestyle disorder, the predominance of which is increasing globally. Diabetes is a complex and convoluted disease characterized by either insufficiency of insulin or immunity to insulin, a hormone which is basic for metabolism of blood sugar. This happens when pancreas don't create enough insulin or when the body can't utilize insulin it produces. Raised glucose level in blood can lead to serious harm to body eventually. People experiencing diabetes ordinarily have indications like frequent urination, excessive thirst, weight reduction or extraordinary appetite. Many a times patient may experience nausea, vomiting or stomach pain. As indicated by WHO report of 2016, 422 million individuals were experiencing diabetes in 2014 which is fourfold increment since 1980. Diabetes was the direct cause of 1.5 million deaths worldwide while as high glucose level was cause of another 2.2 million [2]. According to IDF 1 in 2 adults with diabetes is undiagnosed. It has anticipated around 642 million individuals will have diabetes by year 2040 i.e., 1 adult in 10. IDF estimates there are about 78 million individuals with diabetes living in south and East Asia which will ascend to around140 million by 2040[3]. Diabetes has also grown like epidemic in India which is facing twin burden of under nutrition and over nutrition. India has the unique distinction of being on the hunger map of world as well as being the diabetic capital. In India there is high prevalence of diabetes in urban area rather than in rural areas.

2. RELATED WORK:

In year 2005 [4] came up with an approach ‘CoLe’ to detect diabetes in initial stage. It was a multi-agent system having multiple data miners other than being a combination agent. Its fundamental goal was to accomplish a blend of learning which depicts information in alternate points of view. [5] Applied Association rule on data to classify diabetes. They used equal interval binning approach for discretinizing continuous valued attributes and then applied Apriori algorithm to make diabetic classification and lastly generated Association rules to understand relationship among measured fields, which was used to predict the disease. 4 rules were developed for condition ‘yes’ and 10 rules for condition ‘no’. [6] Used decision tree method for prediction of type II diabetes. Data discretization was used
to pre-processing data. The precision procured was about 78.1768%. [7] Used CART as a technique to build up a data mining tool to monitor diabetes. The algorithm was used to make peculiarity between high risk and low risk patients. The system had an achieved accuracy of 96.39%. [8] Proposed a Hybrid classification system (HCS) to predict diabetes. Fuzzy ID3 was used in designing the model. Estimation-Maximization (EM) algorithm was used for clustering data. The system worked in 2 phases. In phase I cleaned data is fed to EM algorithm. In phase II adaptation rules were obtained, which were essential for predicting diabetes. Accuracy of model was about 91.32%. [9] Developed an expert system using improved version of extended classifier system which is a most successful learning agent in artificial intelligence and has greater accuracy than other data mining techniques. A set of simple rules with “if-than” format were composed. The system achieved an accuracy of 91.32%. [10] proposed a novel artificial bee algorithm with an additional mutation operator for improving performance; they achieved an accuracy of 84.21% using this method. [11] Applied F-P growth and Apriori algorithm to generate association rules on diabetic dataset. Data discretization was used in preprocessing in order to boost precision. [12] Applied fuzzy c means and support vector machine for prognosis of diabetes. FCM gave an accuracy value 94.3% as compared to SVM’S accuracy of 59.5% which is below par. [13] Presented an approach that was jointly implemented using SVM and Naïve Bayes. The system had an accuracy of 97.6%. [14] Used random forest classifier to diagnose diabetes with different test parameters. They achieved an accuracy of about 99.7%. [15] Developed a tool for diabetes prediction and monitoring by using ID3 (decision tree) algorithm. The achieved accuracy was about 94%.

3. MATERIALS AND METHODS

3.1 Dataset used for experiments:
The dataset used in this research is a clinical dataset and is collected from one of reputed diagnostic labs. The dataset at first contained 593 records of patients with age group in range of 25-70. The dataset has 10 attributes namely age, fasting, postprandial, waist, BMI, systolic, diastolic, Hba1c, gender, history and a class label. The dataset initially had some inconsistencies’ so in order to overcome those we pre-processed the data. The dataset had some missing values and replaced it with an appropriate value by applying ReplaceMissingValue filter from weka tool. The filter scans all the values and replaces the missing value by mean mode method. Those attributes that had zero values were removed. After preprocessing we were left with only 532 records. Table 1 Shows all the attributes with description that were used in research.

3.2 WEKA software for data analysis:
WEKA (Waikato Environment for Knowledge Analysis) is a popular suite for data analysis which was developed at university of Waikato New Zealand. The software is available free of cost under GNU public licenses’ that is written in java. WEKA contains a collection of visualization tools and state of art machine learning algorithms for data analysis and modeling. Some of data mining tasks supported by weka are data preprocessing, clustering, classification, regression and visualization. WEKA supports ARFF data format which has two sections namely Header information section and Data information section [16].

Table 1. Attributes used in Dataset.

<table>
<thead>
<tr>
<th>Serial</th>
<th>Attribute name</th>
<th>Description</th>
<th>Values</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Patients age in years</td>
<td>25-70 years</td>
<td>numeric</td>
</tr>
<tr>
<td>2</td>
<td>Fasting</td>
<td>Fasting blood sugar</td>
<td>75-115mg/dl</td>
<td>numeric</td>
</tr>
<tr>
<td>3</td>
<td>Post_pran</td>
<td>Post Prandial blood sugar</td>
<td>75-140mg/dl</td>
<td>numeric</td>
</tr>
<tr>
<td>4</td>
<td>Waist</td>
<td>Waist measurement</td>
<td>30-40 inches</td>
<td>numeric</td>
</tr>
<tr>
<td>5</td>
<td>BMI</td>
<td>Weight in kg’s/height in m²</td>
<td>20-40kg/m²</td>
<td>numeric</td>
</tr>
<tr>
<td>6</td>
<td>Systolic</td>
<td>Systolic blood pressure</td>
<td>90-170</td>
<td>numeric</td>
</tr>
<tr>
<td>7</td>
<td>Diastolic</td>
<td>Diastolic blood pressure</td>
<td>60-100</td>
<td>Numeric</td>
</tr>
<tr>
<td>8</td>
<td>Fha1c</td>
<td>3 month plasma glucose concentration</td>
<td>3.5-10</td>
<td>Numeric</td>
</tr>
<tr>
<td>9</td>
<td>Gender</td>
<td>Gender of patient</td>
<td>M-male</td>
<td>Nominal</td>
</tr>
<tr>
<td>10</td>
<td>History</td>
<td>Family history</td>
<td>Father</td>
<td>Nominal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mother</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Both</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Class</td>
<td>Diagnosis of disease</td>
<td>Yes</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

3.3 Algorithms used in Experimenting:
3.3.1 Naïve Bayes: The Naïve Bayes classification is based on probabilistic Bayes theorem and is one of the most proficient and extremely scalable learning algorithms in data mining [17]. Bayes methods are called eager learners, when giving a training set, eager learners immediately analyze the data and build a model. When it wants to classify an instance it uses this internal model. Theorem is stated as:

$$P(Cj | d) = \frac{P(d | Cj) * P(Cj)}{P(d)}$$
Where \( P(C|d) \) is probability of instance ‘d’ being in class \( C_j \).
\( P(d|C_j) \) is probability of generating instance ‘d’ given a class \( C_j \).
\( P(d) \) is probability of instance ‘d’ occurring.

According to [18] Naïve Bayes is the most used method in case of prediction problems, implementation of which is based on linear function
Some of the advantages of naïve Bayes are:
A). It is fast to train and classify.
B). it is not sensitive to irrelevant features.
C). it handles real and discrete data
D). it handles streaming data well.

3.3.2 Support Vector Machines: Support Vector Machines
were invented by Vladimir Vapnik in 1979[19]. Explaining in its simple and linear form SVM can be defined as a hyperplane that separates’ a set of positive examples from a set of negative examples with maximum margin. SVS’s are useful techniques for data classification, here instances are represented as points in space and are mapped in such a way that separate classes are divide by a clear gap. SVM aims at finding the maximum hyperplane i.e. the one that gives maximum separation between the classes. The instances that are close to maximum-margin hyperplane are called support vectors.
Training data points can be represented as

\[
\{(R_1, S_1), (R_2, S_2), \ldots, (R_n, S_n)\}
\]
Where \( R_i \) is a k-dimensional vector and \( S_i \) is +1 or -1 denoting the class to which a given vector belongs to. Training data is then divided by a hyperplane of general form

\[
W \cdot R + B = 0 \quad (I)
\]
Here ‘W’ is k-dimensional vector, perpendicular to hyperplane and ‘B’ is scalar. Two hyperplanes that belong to two different classes can be described by equations

\[
W \cdot R + B = 1 \quad (II)
\]
\[
W \cdot R + B = -1 \quad (III)
\]
The distance between hyperplanes is \( 2/|W| \), aiming to minimize |w|.
The main advantage of SVM is that it avoids the overfit of data and increase the accuracy or prediction [20],

3.3.3 K-Nearest Neighbor: The technique also known as
Lazy Learner method is a simple technique which stores all the cases and classified new ones’ based on the similarity measure. It is an instance based method which discovers the unidentified data points using previously known data points and classified data points according to voting system [21]. It classifies data the data on close instances, which is decided in terms of Euclidean distance, Euclidean distance between two points say \( P \) and \( Q \) is given as

\[
D(P, Q) = \sqrt{\sum_{i=1}^{n}(P_i - Q_i)^2}
\]

4. EXPERIMENTATION

Experiment was carried out on diabetic data set that was collected from some of the leading diagnostic labs, the description of which is given in table 1. The visualization of dataset is given in figure 1.

In this research we used K-nn, Naïve Bayes and SVM algorithms for experimentation. For sampling the training and testing dataset we used 10-fold cross validation method. This method is a general way of calculating the error rate of a learning scheme on particular dataset. In 10-fold cross validation method, there are 10 equal sized partitions of dataset. Figure 2; figure 3and figure 4 exhibits the experimental results of SVM, Naïve Bayes and K-nn respectively.
5. RESULTS AND DISCUSSION

The main focus of this study was on implementation of data mining in healthcare for predicting diabetes and for that we performed some of the experiments on our diabetic data set by applying data mining algorithms in order to find out which of the algorithm is best in predicting diabetes. We calculated the performance and accuracy of each experiment by using standard metrics which include TP rate, TN rate, Precision, recall, and F-measure which are calculated by confusion matrix and is known as predictive classification. Precision (P) is calculated using True Positive (TP) rate and False Positive (FP) rate and is define as:

\[ P = \frac{TP}{TP + FP} \]

True positive rate (TPR) gives the proportion of positive cases that were correctly identified and is calculated as:

\[ \text{Recall} = \frac{TP}{TP + FN} \]

accuracy \( A \) is calculated as:

\[ A = \frac{TP + TN}{TP + TN + FP + FN} \]

The F-measure is computed as an average of information retrieval precision and recall and is defined as:

\[ F = \frac{2 \times P \times R}{P + R} \]

In general, results of all the implemented algorithms were compared Table 2 shows the various error values of the algorithms applied. The performance accuracy of all algorithms is given in Table 3. Interpreting the values contained in table we found that SVM has highest accuracy of 91.3534% and lowest error rate of 8.6466%.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Error Rate</th>
<th>Correctly Classified</th>
<th>Incorrectly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>91.3534</td>
<td>8.6466</td>
<td>486</td>
<td>46</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>88.1579</td>
<td>11.8421</td>
<td>469</td>
<td>63</td>
</tr>
<tr>
<td>KNN</td>
<td>88.9098</td>
<td>11.0902</td>
<td>473</td>
<td>59</td>
</tr>
</tbody>
</table>

Other performance measures viz. TP rate, FP rate, Precision, Recall, F-measure, ROC were also compared and is given in Table 4. The TP rate and FP rate of SVM is 0.914 and 0.088 respectively. Based on the above results we concluded that SVM has highest accuracy rate, highest TP rate, Precision and F-measure. So we selected SVM classifier model to predict diabetes for our data.
### Table 4. Detailed Accuracy

<table>
<thead>
<tr>
<th>Classifier</th>
<th>TP rate</th>
<th>FP rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.882</td>
<td>0.112</td>
<td>0.886</td>
<td>0.882</td>
<td>0.882</td>
<td>0.963</td>
</tr>
<tr>
<td>SVM</td>
<td>0.914</td>
<td>0.088</td>
<td>0.914</td>
<td>0.914</td>
<td>0.914</td>
<td>0.913</td>
</tr>
<tr>
<td>KNN</td>
<td>0.889</td>
<td>0.119</td>
<td>0.889</td>
<td>0.889</td>
<td>0.889</td>
<td>0.881</td>
</tr>
</tbody>
</table>

### 6. CONCLUSION

The research presented in this paper concentrated on creating and assessing a prescient predictive model for diabetic prognosis. We analyzed frequently used classification techniques Naïve Bayes and Support Vector Machine, k-nearest neighbor on clinical dataset in order to find an optimum solution for diabetic prognosis. The research established the application data mining techniques to predict diabetes effectively and efficiently.

### REFERENCES