Abstract

Knowledge Discovery in Databases (KDD) process is also known as data mining. It is a most powerful tool for medical diagnosis. Due to hormonal changes, diabetes may occur during pregnancy is referred as Gestational diabetes mellitus (GDM). Pregnant Women with GDM are at highest risk of future diabetes, especially type-2 diabetes. This paper focuses on designing an automated system for diagnosing gestational diabetes using hybrid classifiers as well as predicting the highest risk factors of getting Type 2 diabetes after delivery. One of the common predictive data mining tasks is classification. It classifies the data and builds a model based on the test data values and attributes to produce the new classified data. For detecting GDM and also its risk factors, two classifier models namely modified SVM and modified J48 classifier models are proposed. The data set were collected from various hospitals and clinical labs and preprocessed with discretize filter using weka tool. Missing values are replaced by the suitable values. The final preprocessed data applied in the proposed classifier Model. The output of the proposed model is compared with all the other existing methodologies. Since the proposed model modified J48 classifier model produces more accuracy and low error rate against other existing classifier models.

Keywords: GDM, risk factors, classification, SVM, J48, hybrid model.

1. Introduction

Data mining

From the huge amount of data, extracting information or mining knowledge is called Data mining. Descriptive Data mining and Predictive Data Mining are two major classification of data mining. The descriptive describes the interesting properties of data set in concise manner whereas the predictive constructs a set of data models and to predict the behaviour of new data sets. Class description, frequent patterns, Associations, correlations and clusters are some of the descriptive data mining functions. Some of the predictive models of data mining are classification rules, decision trees, mathematical formulae and neural networks. Data mining plays a major role in the areas of market analysis management, fraud detection, sports, medical, education, astrology and corporate analysis. This research work uses data mining to detect the gestational diabetes.

Types of Diabetes

Type 1 DM, Type-2 DM, Gestational diabetes are the three major types of diabetes mellitus.

- Type 1 DM also referred as ‘Insulin Dependent Diabetes Mellitus’ (IDDM) or ‘juvenile diabetes’. In this case, pancreas is failure to produce insulin. Root cause of this type is unknown. Approximately 10% of diabetes people affected by Type 1 DM. Children are mostly affected by this type. It must be managed with only insulin injections.

- Type 2 DM is a category of ‘Non-Insulin Dependent Diabetes Mellitus’ (NIDDM) or ‘adult-onset diabetes’. In this case, pancreas could not produce enough insulin or the body cells are failed to respond to insulin produced. Excessive weight and not enough exercise are the common factors. It may treat by proper exercise, diet and by medication with or without insulin. Approximately 90% diabetes people are of this type.
• Gestational diabetes affects the female only in the time of pregnancy. Gestational Diabetes is diabetes first recognition during pregnancy in which occurs in about 5% of all pregnancies. If it is untreated, it may cause many health problems for both mother and fetus like Miscarriage, preterm birth, fetal death and congenital malformation. Age, previous unknown stillbirth, family history and excess weight are some of the reasons for GDM. Various screening tests are done at the different weeks of pregnancy to diagnose the diabetes. At the first prenatal routine fasting glucose measurement screening test will be done to assess GDM. The fasting glucose level is <5.1mmol/l, then it is treated as normal. Suppose the glucose level is >7.0mmol/l, it is suspected pre gestational diabetes. For suspected GDM the glucose level should be in the range of 5.1-7.0 mmol/l. Oral Glucose Tolerance Test (OGTT) should be done at the 24 weeks of gestation. If the GDM symptoms are present the OGTT should be done again.

Diagnosing the diabetes

Tests for type 1 and type 2 diabetes and pre diabetes

2. Proposed methodology

In this research work, a systematic approach has been developed to detect the gestational diabetes using data mining classifier algorithms.

Proposed algorithms are,

• Modified Support Vector Machine
• Modified J48

Steps Involved

1. Data set undergoes the Preprocessing technique to identify the missing or zero values.
2. Using the discretize filter, the missing or zero values replaced by the proper value.
3. The preprocessed final data set
4. Uploaded in weka as a csv format.
5. The data in the preprocessed data set can be separated by the risk factors of GDM.
6. The identified data set screening values are considered.
7. Proposed modified SVM and modified J48 applied on data set to classify the GDM and Non-GDM.
8. Decision tree is generated to find out the highest risk factor of Gestational Diabetes Mellitus.
9. Conclude the highest risk factor of Gestational Diabetes Mellitus
10. Compare the results with the three proposed methodologies and find the best classifier model with the evaluation measures.
11. Compare the results of proposed methodology with existing methodologies with the parameters like time taken to build a model, accuracy and error rate.

Data set

Data collected from the various hospitals and labs from pregnant women. Age, Blood pressure, height, weight, number of times pregnant, hemoglobin, skin fold thickness, insulin level like various measures are taken into consideration. Totally 6629 instances are taken. All these data are preprocessed using discretize filter before implemented to weka. The following table 2 gives the brief description about the data.

<table>
<thead>
<tr>
<th>S.No</th>
<th>ATTRIBUTE</th>
<th>DESCRIPTION</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Age (years)</td>
<td>Numeric</td>
</tr>
<tr>
<td>2</td>
<td>Ht</td>
<td>Height (in cms)</td>
<td>Numeric</td>
</tr>
<tr>
<td>3</td>
<td>Wt</td>
<td>Weight (in kgs)</td>
<td>Numeric</td>
</tr>
<tr>
<td>4</td>
<td>Bp</td>
<td>Blood Pressure</td>
<td>Numeric</td>
</tr>
<tr>
<td>5</td>
<td>Hmg</td>
<td>Heamoglobin</td>
<td>Numeric</td>
</tr>
<tr>
<td>6</td>
<td>Preg</td>
<td>Number of times pregnant</td>
<td>Numeric</td>
</tr>
<tr>
<td>7</td>
<td>Plas</td>
<td>Plasma glucose concentration (glucose tolerance test)</td>
<td>Numeric</td>
</tr>
<tr>
<td>8</td>
<td>Pres</td>
<td>Diastolic blood pressure (mm Hg)</td>
<td>Numeric</td>
</tr>
<tr>
<td>9</td>
<td>Skin</td>
<td>Triceps skin fold thickness (mm)</td>
<td>Numeric</td>
</tr>
<tr>
<td>10</td>
<td>Insu</td>
<td>2-Hour serum insulin (mu U/ml)</td>
<td>Numeric</td>
</tr>
<tr>
<td>11</td>
<td>Mass</td>
<td>Body mass index (weight in kg/(height in m)^2)</td>
<td>Numeric</td>
</tr>
<tr>
<td>12</td>
<td>Pedi</td>
<td>Diabetes pedigree function(Family history)</td>
<td>Numeric</td>
</tr>
<tr>
<td>13</td>
<td>Hba1c</td>
<td>HbA1c test value</td>
<td>Numeric</td>
</tr>
<tr>
<td>14</td>
<td>Class</td>
<td>Class variable (1-tested positive for diabetes, 0-tested negative for diabetes)</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

Schematic diagram of the proposed
3. Results and discussions

Output of support vector machine (SVM)

- Time taken to build a model: 0.06 seconds
- Correctly classified instances: 5735
- Incorrectly classified instances: 894
- Accuracy: 86.51%
- Error rate: 0.13
- Precision: 0.74
- Recall: 0.60

Output of Modified J48 classifier

- Time taken to build a model: 0.12 seconds
- Correctly classified instances: 6420
- Incorrectly classified instances: 209
- Accuracy: 96.84%
- Error rate: 0.03
- Precision: 0.92
- Recall: 0.93
According to the results, clearly see that decision tree the high accuracy 96.84%. J48 gives the next highest accuracy among all the classifiers in case of diabetes disease. The proposed classifier model used to identify the highest risk factors of the gestational diabetes. All the patients’ data are analysed by using different classifiers such as SVM and J48 Decision Tree. From the results of proposed modified J48 classifier model, conclude that Age factor and strong family history of diabetes, especially first degree of relatives are the two main factors for Gestational Diabetes Mellitus. Fitting predictive models like those utilized as a part of this study might be utilized to create individual/clinical decision support systems, to keep up wellness or enhance the administration of chronic diseases, for example, heart disease and diabetes. Recognizable proof of real hazard factors and creating decision support system, and powerful control measures and wellness programs will decrease in the chronic disease mortality.

4. Conclusion

This paper developed a framework for automatic detection of gestational diabetes using hybrid classifier algorithms of data mining. The data collected from various hospitals and labs are preprocessed and applied on two proposed models namely modified SVM classifier and modified J48 classifier model. The result of the proposed model is compared with all the other existing methodologies and also with each other. Accuracy, error rate, precision and recall are taken as evaluation measures. The decision tree J48 classifier model shows that the proposed model produce more accuracy and low error rate against other existing classifier models.

References