Application of Two Machine Learning Techniques for Predicting and Classifying the Type II Diabetes mellitus

Kasem. A. Farag1*, Fathia. H. Mohammed2 and Marfoua. S. Ali3, Abobaker M. Jaber 4

1Department of Mathmatics, Faculty of Science, University of Omar Al-Mukhtar
2Department of Physiology, Faculty of Medicine, University of Omar Al-Mukhtar
3Department of Zoology, Faculty of Science, University of Omar Al-Mukhtar
4Department of Statistics Faculty Of Science, Benghazi University , Benghazi
Application of Two Machine Learning Techniques for Predicting and Classifying the Type II Diabetes mellitus

Abstract
The main objective of the study is to evaluate the prediction and the classification accuracy of two Supervised Machine Learning Techniques which are linear discriminant analysis (LDA) and logistic regression analysis (LRA) using real data of Type II diabetes. The classification accuracy for both models was determined by the classification accuracy rate. LRA and LDA correctly classified 78.70% and 80.00% of the Type II diabetes mellitus (diabetics and non-diabetics) respectively. The LRA has sensitivity and specificity was 64.38% and 85.35% respectively and the LDA had a sensitivity and specificity of 70.88% and 84.77% respectively. Both algorithms had a good overall classification rate. In terms of proper classification rate, the LDA model slightly outperformed the LRA approach. In general, the findings of this study revealed that the LRA model appears to be appropriate for prediction accuracy while the LDA model appears to be appropriate for classification procedures.

Keywords: predictive accuracy, classification accuracy Logistic Regression Analysis, linear discriminant analysis.
1. Introduction

Diabetes mellitus (DM) is a severe and chronic metabolic disease that has been steadily increased over the last few decades. It appears either when the pancreas produce insufficient amount of insulin (Type I diabetes) or when the body cannot properly benefit from the insulin it produces (Type II diabetes). Diabetes may cause many healthy problems, if it is untreated [3][11].

Type II diabetes mellitus (T2DM) is the most common and prevalent type of diabetes. It is caused by problems with a chemical in the body (Hormone) called insulin. In type 2 diabetes, your body is not able to effectively use insulin to bring glucose into your cells. Increased thirst, unexplained weight loss, and frequent urination are all indications of high blood sugar. Uncontrolled type 2 diabetes can lead to chronically high blood glucose levels which can cause several and serious complications include heart disease, strokes and diabetic retinopathy which may result in blindness, kidney failure [9].

Machine learning (ML) is a kind of an artificial intelligence (AI) technique which allows computers to learn without having to be explicitly programmed [1]. Machine learning techniques are impressively employed in the medical field. These algorithms have been shown to be more effective and powerful in detecting, diagnosing and predicting, classifying and identifying various illnesses because they can handle enormous amounts of data, aggregate data from several sources, and integrate background information into the research. There are many (ML) techniques like linear discriminant analysis (LDA), logistic regression analysis (LRA), and Support vector machine (SVM), Decision Tree... etc. [2][5][8][12]. This paper compares performance of two Supervised Machine Learning Techniques classifiers for Predicting and Classifying the Factors Affecting Type II diabetes mellitus (T2DM)

2. Data and Methodology

The aim of the study is to predict and classify the effective of the factors affecting Type II diabetes mellitus (T2DM). The study data were collected from diabetes clinic (known as clinic number 1) in Al-Baydha city. The data included 230 instances with 2 classes (The outcome is either 0 for cases that does not have diabetes or 1 for cases that has diabetes). The selected related independent variables after detecting multicollinearity and correlated
variables included age, sex, Marital status, occupation status, body mass index (BMI), hypertension, Total cholesterol (mmol/l), Triglyceride (mmol/l), Family History, physical activities. Two supervised machine learning classifiers, The Logistic regression analysis (LRA) and linear Discriminant Analysis (LDA), techniques were implemented in the main study. The statistical analysis and data management has been carried out using R.

2.1 The Logistic Regression Analysis (LRA): It's a machine learning classification model that has been used in many medical researches. It is used to examine the association between independent variables, also known as exposure or predictor variables, and a binary or dichotomous dependent variable, also known as the outcome or response variable, and to determine the likelihood of a disease as a function of a risk factor. Furthermore, it is a prediction model that determines if a patient has a disease based on diagnostic measures contained in the dataset and investigates various approaches to improve performance and accuracy.[4][7]

2.2 Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) or Fischer Discriminants is a robust classification and prediction algorithm as well as a dimension reduction model. By establishing a decision region between the distinct classes, LDA is considered as a powerful technique for class separability. The main objective of LDA is to optimize the ratio of between-class and within-class variances [6][7].

2.3 Confusion Matrix: It is used for evaluating the performance of machine learning classification algorithms. It gives us a comprehensive picture of how well our classification model is performing. It also compares the actual class with those predicted by the machine learning model. This matrix is also used to show the correct and incorrect instances.

Table 1. Confusion Matrix

<table>
<thead>
<tr>
<th>Predicated class</th>
<th>Actual class</th>
<th>Positive – P</th>
<th>Negative – N</th>
</tr>
</thead>
<tbody>
<tr>
<td>True – T</td>
<td>TP</td>
<td></td>
<td>TN</td>
</tr>
<tr>
<td>False – F</td>
<td>FP</td>
<td></td>
<td>FN</td>
</tr>
</tbody>
</table>

- True Positive – [TP]: This instance indicates patient samples that were classify as patient.
• True Negative – [TN]: This instance indicates not – patient samples that were classify as not – patient.

• False Positive – [FP]: This instance indicates not – patient samples that were classify as patient.

• False Negative – [FN]: It indicates patient samples that were classify as Not – patient.

2.4 The Evaluation Measures of algorithm performance:

Sensitivity (SE), Specificity (SP), positive predicted value (PPV), Negative predicted value (NPV), Accuracy (ACC), Error Rate (ER) and Cohen’s kappa coefficient were used to evaluate the algorithms performance[10][13].

Results

The confusion matrix and the classifiers performance results for LRA and LDA models over Evaluation Measures like Sensitivity (SE), Specificity (SP), positive predicted value (PPV), Negative predicted value (NPV), Accuracy (ACC), Error Rate (ER), and Cohen’s kappa coefficient values are listed in Table-2. The results of confusion matrix showed that how well the models have been performed. The LRA model correctly classified 181 out of 230 instances while 49 erroneously classified. 26 instances with diabetes were classified as non – diabetics and 23 who do not have diabetes have been classified as diabetic. For the LDA model, 184 out of 230 instances were classified correctly and 23 instances were classified incorrectly for each label.

Table 2. Classifiers Performance Analysis

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Evaluation Measure</th>
<th>LRA</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>47</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>TN</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>26</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>FN</td>
<td>134</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>64.38%</td>
<td>70.88%</td>
</tr>
<tr>
<td></td>
<td>SP</td>
<td>85.35%</td>
<td>84.77%</td>
</tr>
</tbody>
</table>
By comparing the evaluation measures for both models, it can be observed that for LDA performs somewhat lower than LRA at classifying negative instances, which is when the patient does not have diabetes with 85.35% and 84.77% respectively. On the other hand, the LDA algorithm performs better in the case of classifying positive instances, which is when the patient has diabetes. For LRA is 64.38% and for LDA it is 70.88%. Also, the accuracy for both algorithms is relatively similar with LRA’s accuracy at 78.70% and LDA’s accuracy at 80.00% which means that LDA performs somewhat better in general. The kappa statistics for LRA was 50.29% and this statistic for the LDA was 55.65%. The ER for the LDA is lower.

**Conclusion:**

The objective of this paper was to use LRA and LDA models to predict and classify of the Factors Affecting Type II diabetes mellitus. Both models have given slightly similar results performance. The overall classification rate for both was good. Although the LDA slightly exceeds LRA in the correct classification rate, either can be helpful in predicting and classifying. In order to decide which algorithm should be applied, the assumptions for each application must be considered.
References:


