Analytical study on diabetes prediction-using random forest classifier

Dr. Kabirdoss Devi
Assistant Professor, Department of Commerce, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai

Mr. J. Sathish Kumar
Assistant Professor, Department of Commerce, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai

Dr. S. Poovizhi
Assistant Professor, Department of Management Studies, St. Joseph’s College of Engineering, Chennai

Mr. M. Krishnan
Associate Business Intelligence, 4i Apps Solutions Private Ltd.

Abstract---Diabetes predictions have gained major attention due to its consequences on the healthy well-being of an individual. When glucose levels go high due to non-availability of the hormone called insulin which digest glucose, together with other side effects like frequent urination, excessive thirst, and hunger with sudden weight reduction, one can be confirmed of suffering from diabetes. This requires a consistent treatment and monitoring of its complications which are considered fatal in some cases. There are various ways to keep a tract of the glucose level in blood to adjust the diet and dosage of insulin. However, predicting it as early as possible is a challenging task due to its inter-dependency factor that causes trouble to human organs like viscera, peripherals, nervous system, cardiovascular, eyes and excretory system. This research paper aims to provide five different machine learning methods for the prediction of diabetes such as SVM, Logistics regression, KNN Classifier, Random Forest and Logistic algorithm. These proposed methods are effective techniques for earlier detection of the diabetes.

Keywords---SVM, KNN, classifier, diabetes, random forest, algorithm.
Introduction

Diabetes is one the most troubling disorder which mandates regular check-ups for the awareness of insulin level and to keep up the intake. It sometimes considered to be fatal when ignored as it damages the important vital organs of cardiovascular and respiratory organs. Diabetes affects all the important organs including eye causing cataract and severe cases blindness. It also affects the peripheral nervous system, skin and feet. Gangrene is the another major issue resulting in tissue necrosis and mandatory surgical intervention. There are two ways diabetes in patients can be calculated. Firstly, data pre-processing where all the attributes are identified and secondly, the predictive models which are constructed based on the decision tree methods. There are many available tools for the purpose like WEKA mining tools, Decision tool algorithm and Artificial Neural Network using kappa statistics, mean absolute error and relative squared error.

Objectives of the study

The objective of the study is:

1. To predict the presence of diabetes in an individual based on insulin level, age and BMI.
2. To analyse the preset data obtained from Kaggle for the presence of diabetes.
3. To test all the five algorithm for the best fit algorithm identification.

Review of Literature

Diabetes is made as a chronic disease as there is no permanent cure leading to many consequences and demanding life long conservative management for keeping the sugar level at low. (Roshi Saxena et.al 2022) conducted a analysis on models for diabetic predictions and reviewed many researches to find the best one. They brought out that the analysis of these diabetic datasets were quite challenging because these were non linear and complex in nature and they proved that these shortages could be overcome by the machine learning based stratification of risk system for the classification of patients into diabetic and control. (Quan Zou et.al 2018) used decision tree, random forest and neural network to predict DM. Their devised models were used five-fold cross validation for the purpose of examination. They conducted independent experiments to check the methods that had the probability for the better performance. They did a random selection of 64994 healthy people and diabetic datasets for the training set. The outcome projected high accuracy with the random forest when all the other attributes used. (More, Rana 2017) they summarized all the literature survey between the years 2000 and 2016 with respect to the RFC and resolving class imbalance. The research paper (Gerald Biau 2012) on random forest discussed about the set of decision trees that was growing between the spaces of the datasets. He also discussed about the statistical and mathematical properties that was driving the algorithm of the random forest. This paper they discussed the in-depth analysis of this model proposed by Breiman in 2004. This was very close to the original algorithm. Not only diabetic predication and some of the health care contributions but also environmental issues like ground water yield
was predicted with the help of self-learning Random Forest (Maher Ibrahim, Biswajith, Saro Lee 2018). They conducted this study in South Africa having inventory data segmented into two groups where the first group was made 70% of training and 30% of testing. They used Bayesian Optimization Method and compared their model with other ML models such as Support Vector Machine, Artificial Networks and Decision trees.

**Research Methodology**

This methodology uses Random Forest which is Machine Learning and part of Artificial Intelligence enabling computer ability to learn. Reinforcement leaning is one of the category of machine learning which learns through past experiences. However, Random forest uses supervised learning algorithm that uses both regression and classification. Random forest has many decision trees similar to real forest that has many trees. These various subsets improves the accuracy of the outcome. This research paper aims to predict presence of diabetes through five different supervised machine learning methods like SVM, Logistics Regression, KNN Classifier, Random Forest and Logistic algorithm.

The architectural design of the research methodology includes -

![Architectural Design Diagram]

**Data Design**

Datasets for this study has been taken from the secondary source where the details of the past records of diabetes predictions stored. These datasets were obtained from the third party source known as Kaggle and it consisted of seven features and two thousand plus rows. The variables such as glucose, insulin, BMI, BP included in the model. The information related to the study alone were fitted to the database model. This database model was designed such a way that it must avoid redundant to prevent wasting of space which would further increase the faults and discrepancies within the database. There are many features being used in the machine learning model which used to predict the output accurately. These datasets has a major drawback; where certain features in the datasets might be high value and some might be low value to make it more feasible.

**Data Collection**

The multiple diabetic datasets were obtained from Kaggle and Merged together. The variables for the study includes glucose level, BMI, prediction function and insulin etc.
Data pre – processing

There are four steps involved in the data pre-processing mainly to avoid errors, inconsistency and incomplete status of the data. Errors do occur when the raw data being transferred to an understandable format for the analysis. To qualify for the data analysis, the datasets should be split into training and validation sets, should be rectified for all the missing values, should be divided based on the categorical features and the datasets should be normalized.

Tools used for analysis

The whole computation process was made on the web-based application for the ease of capturing entire calculation. These calculation process includes developing, documenting and coding and communicating outcomes. Another component used was browser based where the documents could be combined for explanatory texts with rich media output computations such as HTML, LaTeX, PNG, SVG etc.

Data Structure

Data Analysis

Below is the datasets that has been included for the study highlighting all the variables. The values are arranged in an array for the easy calculation and for the inclusion in the study.
Data Set

Detecting correlation between input features

High and low correlation between the variables are plotted in the correlation matrix. This brings out the fact that the changes in direction of the one variable with respect the another variable in specific manner. Revealing the relationship between the variable is useful because value of one variable helps to predict the value of another variable.

Pair plot of the attributes

Pairwise relationship is being made through the help of attributes of pair plot methodology. Pair plot is made with variables across the columns and rows, pairplot() function is used to plot multiple pairwise bivariate distributions. This
pair plot shows the relationship for \((n,2)\) in a DataFrame as matrix with the combination of variables as well as diagonal plot of univariate. This pair plot is very useful in identifying the relationship between both single and two variables. Follow-up analysis are better made with the pair plots and which can also easily be implemented with Python.

**Pair plot**

Confusion Matrix

The first step after data has been cleaned, pre-processed and wrangled is to feed it to a model and get the probability output. For the better effectiveness and performance the datasets requires confusion matrix.
Confusion Matrix

Performance measurements for machine learning comes with the four different values of predicted and actual values.

True Positive: Interpretation: You predicted positive and it’s true.
True Negative: Interpretation: You predicted negative and it’s true.
False Positive: (Type 1 Error) Interpretation: You predicted positive and it’s false.
False Negative: (Type 2 Error) Interpretation: You predicted negative and it’s false.

Results and Discussion

Chronic diseases like diabetes requires early detection to prevent irrecoverable damages to the vital organs. This random forest technique not only helps researchers to create veracious and effective tool but also clinician to make the best decision about the disease condition. Availability of surplus datasets of epidemiological and genetic risk and the advanced computational methods makes the machine learning with an enhanced way of predictions. There are different data mining techniques and its applied algorithm on different medical data sets and these have strong power of machine learning on datasets.

Accuracy Score

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>ACCURACY SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>77.29%</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>78.73%</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>75.28%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>75.84%</td>
</tr>
<tr>
<td>Random Forest Classifier</td>
<td>96.87%</td>
</tr>
</tbody>
</table>

Accuracy score works well when there are equal number of samples in each class. If there are two classes with 98% in one class and 2% in another class, then the model considers the 98% training accuracy and omits the other class. Similarly, when the same model is being utilized for 60% and 40%, then it would drop to 60%. Hence, classification accuracy gives the false sense while achieving high accuracy.

F_Score

The highest possible value of an F-core is 1.0. This indicates the perfect precision and the lowest value is 0 and if either then the precision or the recall is zero.

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>F - SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>61.5%</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>67.48%</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>56.5%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>61.26%</td>
</tr>
<tr>
<td>Random Forest Classifier</td>
<td>99.74%</td>
</tr>
</tbody>
</table>
Graphical User Interface

Many GUIs come with touchscreen and voice-command and allow the users to interact with different electronic devices. GUI are found to be great when compared to the command line interface. GUI which was commercially deployed into Apple Macintosh and Windows Android designed to respond to the problem of inefficient usability. The text-based command-line interfaces for the average user.

There are different kinds of algorithms of machine learning being used to discover patterns in machine language like python which directs to the actionable visual precepts. Based on the learning pattern this algorithm can be classified into two groups such as supervised and unsupervised learning. Predicting the class for the given data point is known as classification. This predictive model does a mapping function on the X input variables to Y discrete output variables. Under the supervised learning, it is possible to predict because computer learns the data input given. This data has the ability to identify bi-class like male or female or multi-class where it includes many variables.
Feature Interpretation

Stream Lit GUI
Stream lit is an open source python library. It makes it easy to create and share beautiful web apps for data analysis. This particular app does not require any knowledge on web development. It automatically saves and updates all the codes and the codes runs from top to bottom and does not require a callback.

Suggestions

Clinicians with the help of analysts should explore the possibilities of predicting the glucose level in the blood using these models. HTML and FLASK can also be used to predict well compare to steam lit web application. HTML would provide better web application compare to GUI. HTML and FLASK can also be easily uploaded into cloud storage compare to SLA.

Conclusion

Patients stage and the grade has to be found and the parameters like accuracy, classification report, confusion matric by SML algorithm method for the analytical process after the data cleaning, pre-processing and wrangling done. Finding out the level of glucose through machine learning and predicting at a minimum effort is the need of an hour. When the advanced technology grows exponentially, the problem should be addressed and explicitly used. LDA and SVM balances with 92% accuracy. RNN balances with 94%.
References


