School of Computer Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

Final Review Report

Programme: B. Tech-Computer Science Engineering

Course: CSE2004 - Database Management Systems

Slot: D2

Faculty: Dr. M. PREMALATHA

Component: J

TITLE: PREDICTION OF DIABETES USING CLASSIFICATION ALGORITHMS

TEAM MEMBERS:

NAME: ATHUL SHANKER

REGNO: 20BCE1337

NAME: SAM MESHACH D

REGNO: 20BCE1356

NAME: KANNA LAKSHMI DATHATREYA

REGNO: 20BCE1687

ABSTRACT

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic Center andconsulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy.

Therefore, two machine learning **classification algorithms** namely **Decision Tree** and **Support Vector Machine (SVM)** are used in this experiment to detect diabetes at an early stage. Experiments are performed on **Pima Indians Diabetes Database (PIDD)** which is sourced from UCI machine learning repository. The performances of all the two algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

KEYWORDS:

Diabetes; SVM; Decision Tree; Accuracy; Machine Learning.

INTRODUCTION

1. Data Set Description:

This Dataset consists of the medical details for 768 instances which are pregnant female patients. The dataset comprises of 9 numeric value attributes which are:

1. Number of times pregnant
2. Plasma Glucose Concentration
3. Diastolic Blood Pressure (mm Hg)

4. Skin Fold Thickness (mm)
5. 2-Hour Serum Insulin (mu U/ml)
6. Body-Mass Index (BMI) (weight/(height)^2)
7. Diabetes Pedigree Function
8. Age
9. Class ‘0’ or ‘1’.

About this file

This dataset describes the medical records for Pima Indians

and whether each patient will have an onset of diabetes within veyears.

Fields description follow:

preg = Number of times pregnant

plas = Plasma glucose concentration 2 hours in an oral glucose tolerance test

pres = Diastolic blood pressure (mm Hg)

skin = Triceps skin fold thickness (mm)

test = 2-Hour serum insulin (mu U/ml)

mass = Body mass index (weight in kg/ (height in m) ^2)

pedi = Diabetes pedigree function

age = Age (years)

class = Class variable (1: tested positive for diabetes, 0: tested negative for diabetes)

2. Consider the schema alone and normalize it till BCNF using schema decomposition.

Step 1: Find merged minimal cover of FDs, which contains:

Insulin --> Glucose,BloodPressure,SkinThickness

DiabetesPedigreeFunction --> Pregnancies,Age,Outcome

Age --> BloodPressure

Initially rel[1] contains the original table, with the FDs above

Step 2: Checking whether table rel[1] is in BCNF

The FD [Insulin --> Glucose,BloodPressure,SkinThickness] violates BCNF as the LHS is not superkey. Table is split into the two below:

rel[2]= (Insulin,Glucose,BloodPressure,SkinThickness)

With FDs:

Insulin --> Glucose,BloodPressure,SkinThickness

rel[3] = (Pregnancies,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome)

With FDs:

DiabetesPedigreeFunction --> Pregnancies,Age,Outcome

Age --> BloodPressure

Step 3: Checking whether table rel[2] is in BCNF

Table rel[2] is in BCNF already.

Round3: Checking whether table rel[3] is in BCNF

The FD [DiabetesPedigreeFunction --> Age,Pregnancies,Outcome] violates

BCNF as the LHS is not superkey. Table is split into the two below:

rel[4]= (DiabetesPedigreeFunction,Age,Pregnancies,Outcome)

With FDs:

DiabetesPedigreeFunction --> Pregnancies,Age,Outcome

rel[5]= (Insulin,BMI,DiabetesPedigreeFunction)

Step 4: Checking whether table rel[4] is in BCNF

Table rel[4] is in BCNF already

Step 5: Checking whether table rel[5] is in BCNF

Table rel[5] is in BCNF already

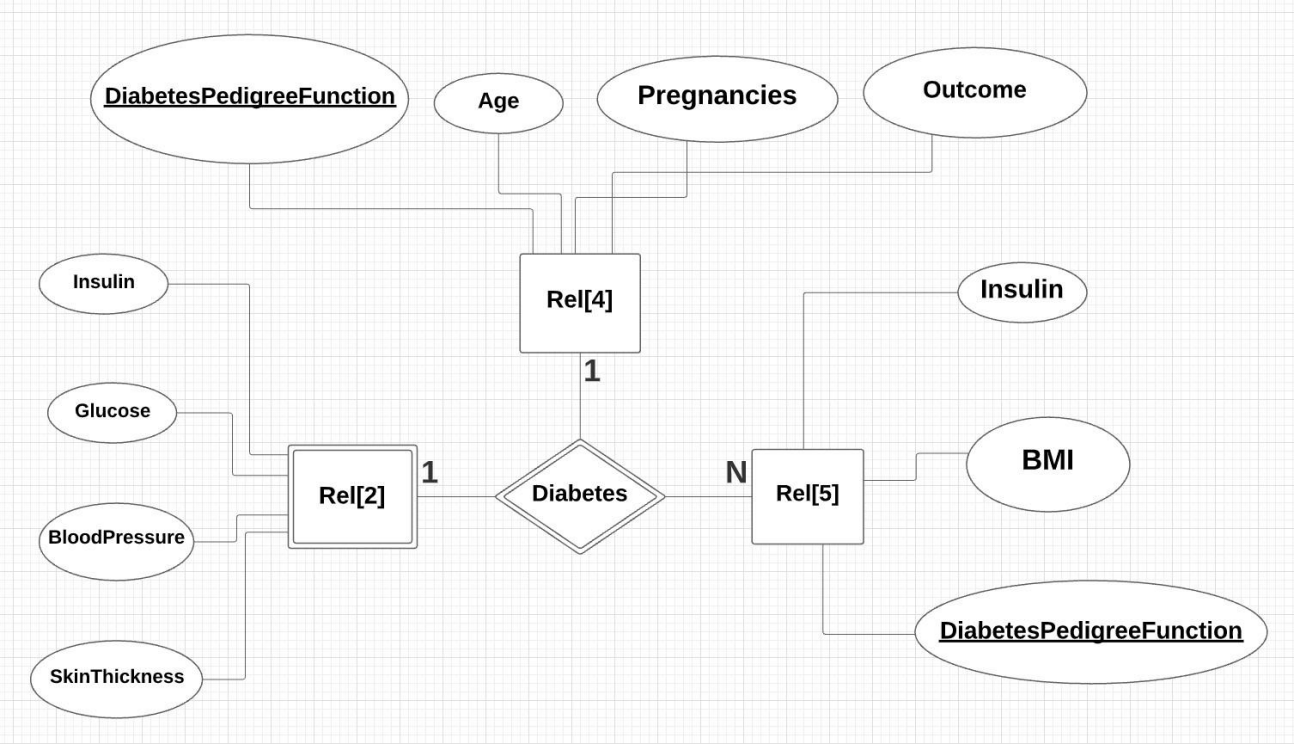
The Final relations are rel[2], rel[4], rel[5].

rel[2]= (Insulin,Glucose,BloodPressure,SkinThickness)

rel[4]= (DiabetesPedigreeFunction,Age,Pregnancies,Outcome)

rel[5]= (Insulin,BMI,DiabetesPedigreeFunction)

3. Draw the ER diagram for the final decomposed schema stating the key attributes, mapping cardinalities, participation constraint, and so on.



4. Methodology and Algorithm used:

The methodology used in the project is **classification**. Classification aims at identifying the category of a new observation among a set of categories based on a labelled training set. By using classification, we can

identify a diabetic person when there is a change in the data from the data that a non-diabetic person would have.

Algorithms used are **Support Vector Machine (SVM)** and **Decision Tree**.

Support Vector Machine (SVM):

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes. For better generalization hyperplane should not lie closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors.

Decision Tree:

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent.

classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes.

5. Implementation:

As per the project proposal we will be trying to predict diabetes in patients given the medical features of the patients using classification algorithms using Scikit_Learn and Python Programming.

1. Load the libraries, numpy, pandas and matlablib.pyplot
2. Import the dataset from local PC's location.
3. Check whether if there is any null data or not?
4. Summarizing each field of the dataset.
5. Divide the dependent variable (Outcome) and the independent variables.
6. Import the Standard scaler and performing the scaling of independent variable.
7. Grid Search cross Validation.
8. Apply Support Vector Classifier algorithm.
9. Grid Search Parameters for SVC.
10. Building the SVC with best parameters available.

11. Plotting the performance measures for SVC.

12. Plotting the Confusion matrix for SVC.

13. Apply Decision Tree algorithm.

14. Grid Search Parameters for Decision Tree.

15. Building the Decision Tree with best parameters available.

16. Plotting the performance measures for Decision Tree.

17. Plotting the Confusion matrix for Decision Tree.

18. Comparing the performance of both models.

19. Visualize the performance of SVC and Decision Tree using graphs.

20. Plotting the ROC Curve of Each Model.

Model Diagram

Proposed procedure is summed up in figure-1 underneath in the form of model diagram. The figure shows the pattern of the research conducted in constructing the model.

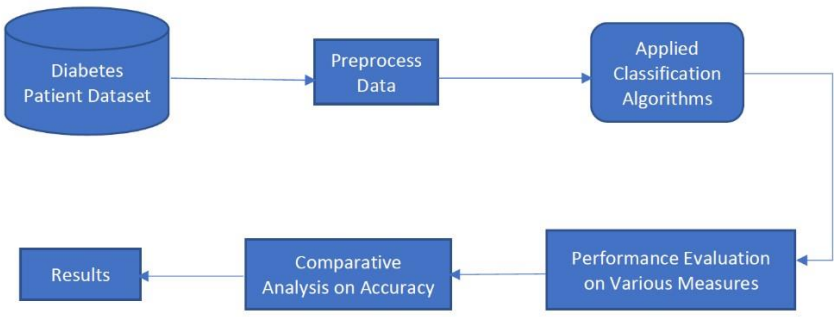


Fig-1: Proposed Model Diagram

6. Results and Discussion:

- We trained SVC, DecisionTree, using the GridSearchCV with internal 10 fold cross validation to find the best model using different parameters.
- For measuring the performance of these models we used various mesures such as Accuracy, Precision, Recall, F-Measures and ROC which one can see in the table above the plots.
- As it is clearly visible that all two models works similar for this dataset. So the best supervised Machine Learning algorithm is Support Vector Classifier (SVC) with the 82% accuracy and 0.76 Precision on test dataset in respective to other data model for this experiment.
- The results we have obtained are within the range expected in the project proposal (75-82%), which is a good indicator as we have got results in the upper bound of expected range.

7. Conclusion:

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, two machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 82 % using the Support Vector Classifier classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.