A Report on:
"Predictive Analysis on PIMA Diabetes Dataset Using a Supervised Learning Method: Decision Tree"

For the course
Data Warehousing and Data Mining
Section: 'A'

Submitted by
HOSSAIN, MD MUZAKKER
18-37801-2
Dept. of CSE, American International University-Bangladesh

Submitted to
DR. MD. MAHBUB CHOWDHURY MISHU
Assistant Professor and Head [Undergraduate Program], Computer Science
American International University-Bangladesh

Table of Contents

1. Introduction

In this age of information the amount of data is overwhelming. To cope up with the pace of astronomical growth of data and extract the valuable information from that is quite intimidating. The "Data Mining" term comes in this regard. In this report I will explain and perform a predictive analysis on a dataset using techniques and tools of data mining to find out certain information based on certain diagnostic measurements included in the dataset step by step.

2. About PIMA Diabetes Dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. There are 768 instances on this dataset and about 9 attributes. This dataset describes the medical records for Pima Indians and whether or not each patient will have an onset of diabetes within next years.

The Attribute description follows:

preg = Number of times pregnant

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

pres = Diastolic blood pressure (mm Hg)

skin = Triceps skin fold thickness (mm)

insu = 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)

3. Explanation of Attributes and Features

There are 9 distinct attributes in this PIMA-Diabetes dataset. The following graph shows the Data Point of these attributes that were collected from at least 21 years old female of Pima Indian heritage:

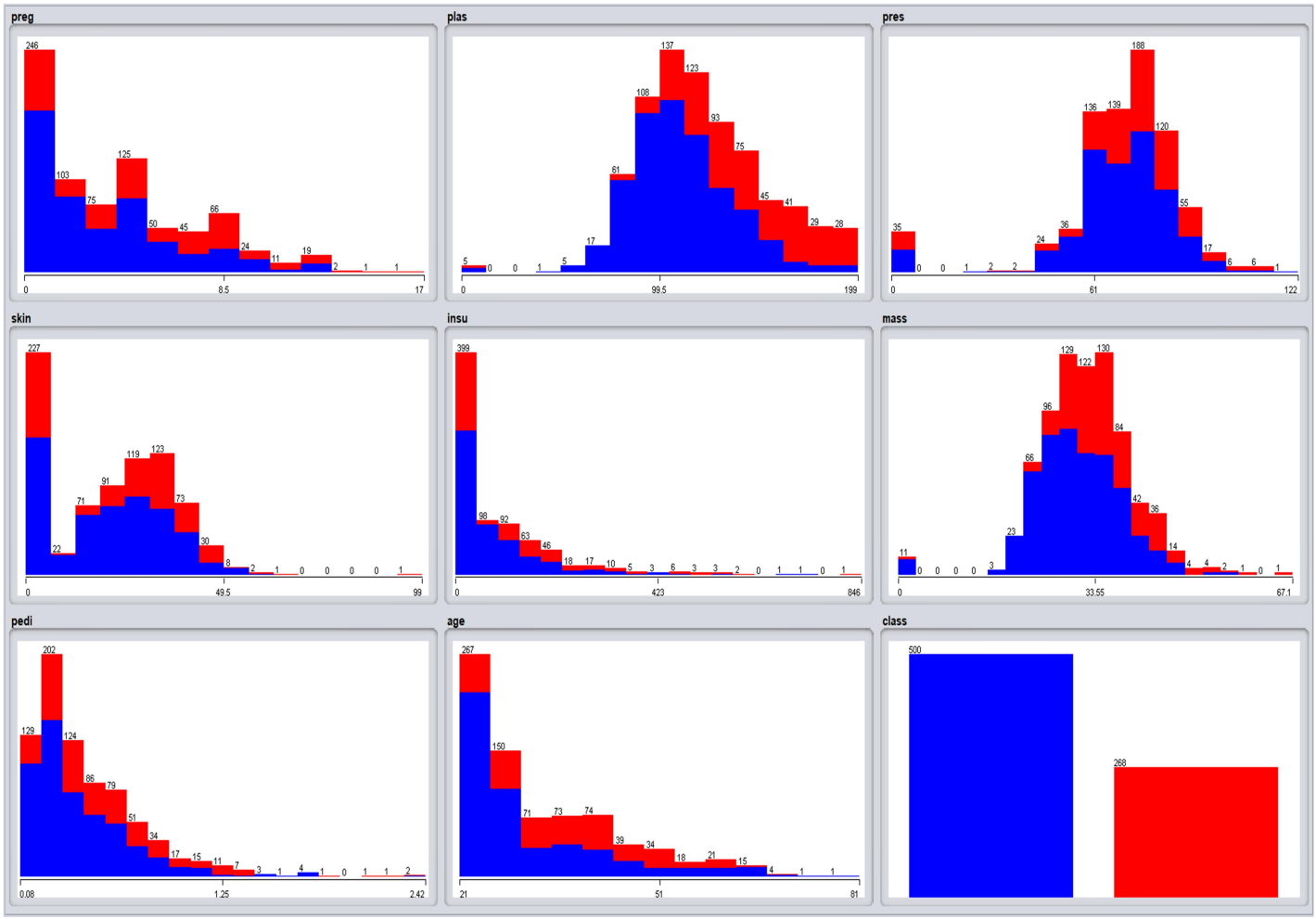

Figure 1: Graphical Representation of Attribute Data Point

preg: The preg bar-chart represents the number of times that the females of at least 21 years of age have been pregnant. It is a numeric type of attribute with a minimum value of 0 and a maximum of 17. It shows that about 246 instances have been pregnant from 0 to 1.308 times and the average number of pregnant time is

around 3.845. The red color represents the instances with positive result of diabetes and the blue one represents the negative result.

plas: The plas chart shows Plasma glucose concentration a 2 hours in an oral glucose tolerance test with an average of 120.895. It is a numeric value with minimum value of 0 and a maximum of 199. The highest number of instance has been encountered between 99.5 and 111.938 of glucose concentration with 137 instances. The risk of getting diabetes increases with the increase of Plasma glucose concentration.

pres: This chart shows Diastolic blood pressure (mm Hg) with a minimum value of 0 to a maximum of 122. The average Diastolic blood pressure among the instances is about 69.105 mm Hg and the largest instances have between 71.765 to 78.941 mm Hg.

skin: Skin chart shows Triceps skin fold thickness in mm in numeric value. This chart shows a large gap among the instances in this criterion and around 227 instances of 768 have been found between 0 to 6.188 mm. Although, the average Triceps skin fold thickness remains with 20.536 mm.

insu: It refers to serum insulin or more precisely the test of 2-Hour serum insulin (mu U/ml) shows numeric value with a minimum of 0 and a maximum of 846 mu U/ml serum insulin. An even larger gap is found in this chart among the instances. The largest instances of 399 have this serum insulin between 0 to 44.526 mu U/ml which tend to the minimum values and the average is quite higher than this which is 79.799 mu U/ml.

mass: This chart show the BMI or Body mass index (weight in kg/(height in m)^2) of instances from a minimum numeric value of 0 to 67.1 of maximum. The average BMI of the instances is found 31.993 whereas majority of instances reside between 27.45 to 36.6 BMI. This index shows the direct relation between obesity and diabetes. With the increased BMI the number of positive diabetes instances increases.

pedi: The Diabetes pedigree function that indicates the risk getting lower with the increase of this function. The average value of pedigree function is found 0.472 where 2.42 is the maximum value.

age: This chart shows the relation of age and diabetes where the minimum age is 21 and the maximum is 81. The chart shows that the age of 21 to 25.615 have the higher risk of being diabetes positive. Surprisingly with the increase of ages the diabetes positive number drops steeply.

class: This is a nominal attribute, Class variable (1:tested positive for diabetes, 0: tested negative for diabetes) representing the number of diabetes positive and negative. Among the instances around 500 have been tested negative and the rest are diabetes positive. Based on this variable I will be performing prediction whereas an instances can be diabetes positive or not.

4. Chosen Classification Method

The main objective of this report is to predict whether a patient has diabetes or not. In order to do so I'll need a class variable that can categorize the patient based on their diagnostic measures. If the measures of a patient is found to be potentially diabetes positive this can be categorize in positive class and negative otherwise. We've learnt about Naive-Bayes technique, K-Nearest-Neighbors and Decision Trees technique to predict a possible outcome. Among these three I choose the third option as my technique to do the prediction. Followings are the reasons that I choose this technique –

Naive-Bayes: Naive-Bayes is a good technique that solely depends on categorical or nominal attributes but it doesn't perform well for the combination of categorical and continuous attributes or even only continuous attributes. But in the dataset there are mostly numerical attributes are present. Naive Bayesian prediction requires each conditional prob. be non-zero, otherwise the predicted prob. will be zero which is another major drawback of this technique. Hence, this technique has not been chosen for this particular dataset.

K-Nearest-Neighbor: KNN classification is mainly applied when the dimension (i. e., the number of attributes) is small. But in this dataset there are 768 instances which can be hard when we will need to analyze the confident intervals.

Decision-Tree: A is a supervised learning method used for classification and regression. It is a tree which helps us by assisting us in decision-making. It breaks down a data set into smaller and smaller subsets and simultaneously decision tree is incrementally developed. The final tree is a tree with decision nodes and leaf nodes. A decision node has two or more branches. For the dataset it is a better choice as we can visualize the attributes in a tree-from and make decision quicker.

5. Result Analysis

To apply decision tree classification I used WEKA tool to apply the algorithm. In WEKA there is an option called preprocess where one can choose how can attribute he want to work with from the dataset and this can help one with specific requirement of attribute. In the classify option there is a choose option from where open can select a desired algorithmic technique to perform prediction.

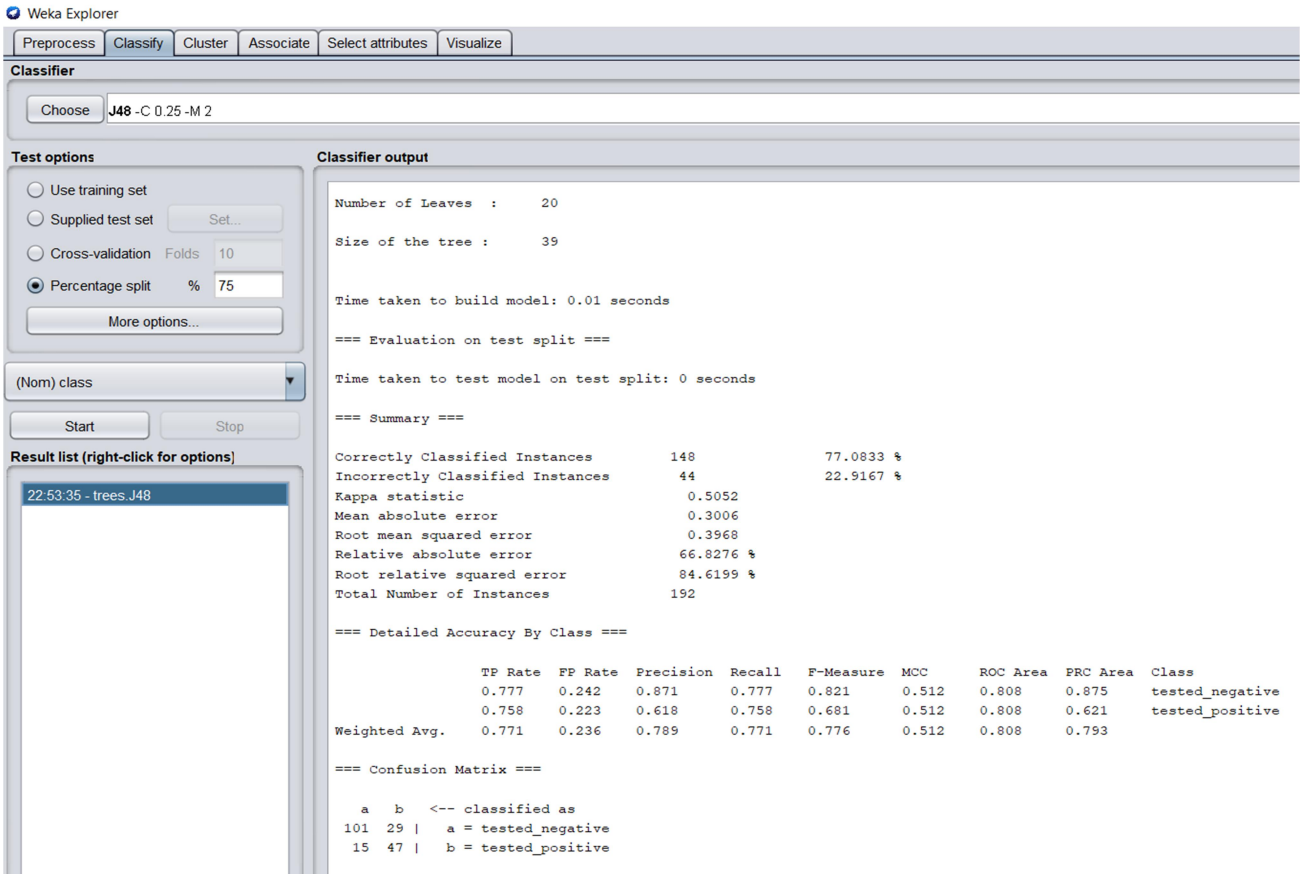

Figure 2: The Procedure of Performing Algorithmic Technique in WEKA

In my case, I've chosen J48 technique which is an extension of ID3 algorithm of decision tree. I've chosen a percentage split of 75% meaning I've taken 75% data from the dataset as training data and rest of 25% as test data. After performing the algorithm on nominal class attribute I've found the following result –

```
=== Summary ===

Correctly Classified Instances         148               77.0833 %
Incorrectly Classified Instances        44               22.9167 %
Kappa statistic                          0.5052
Mean absolute error                      0.3006
Root mean squared error                  0.3968
Relative absolute error                 66.8276 %
Root relative squared error             84.6199 %
Total Number of Instances              192

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.777    0.242    0.871      0.777    0.821      0.512    0.808     0.875     tested_negative
                 0.758    0.223    0.618      0.758    0.681      0.512    0.808     0.621     tested_positive
Weighted Avg.    0.771    0.236    0.789      0.771    0.776      0.512    0.808     0.793

=== Confusion Matrix ===

   a   b   <-- classified as
 101  29 |   a = tested_negative
  15  47 |   b = tested_positive
```

Figure 3: The Summary of the Classifier Output

Here, using the J48 classifier I've found about 77.0833% of accuracy and about 44 instances have been identified incorrectly. The mean absolute error is found 0.3006 and it measures how similar the forecasts are to the final outcomes. The root mean squared error 0.3968 indicates the standard deviation of the sample from the variations between the expected values and the observed values. The confusion matrix here represents a visualization method. It shows the particular diagnostic measures and how they are contributing to the final outcome class.

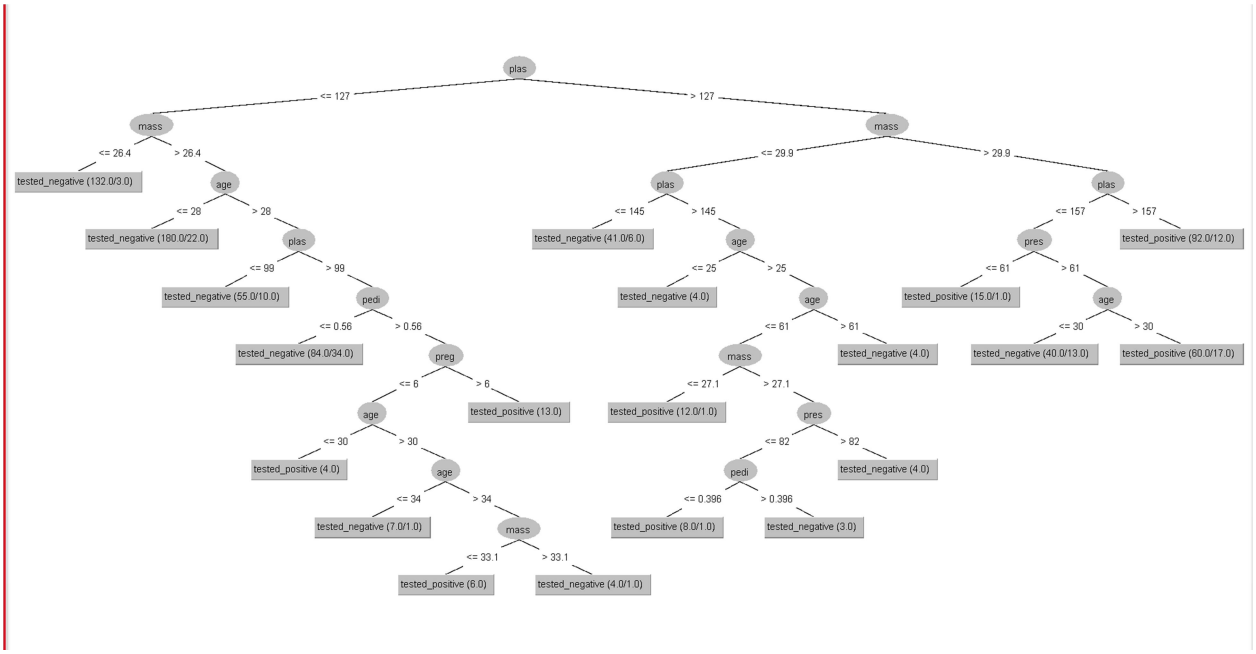

Figure 4: Tree Representation of Class

From the decision tree one can come to the final outcome of either of two tested result and also see how an attribute is affecting in decision-making.

6. Conclusion

In order to perform a predictive analysis on PIMA-Diabetes dataset I've used a supervised learning method, Decision Tree (J48 classifier – WEKA Tool). From the dataset the overall performance of the decision tree is found around 77% accuracy where other models didn't score this accuracy. So it can be said that it is a sufficient model to provide one useful predictions about a patient whether he/she is diabetes positive or not.

7. Appendix

1. The PIMA-Diabetes Dataset – UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/support/diabetes)