AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH (AIUB)
Faculty of Engineering
Department of Electrical and Electronic Engineering

Data Warehousing and Data Mining (Sec: C)

Submitted By:

Name: Hameem, Md. Asir Id : 17-33856-1

Submitted to:
Dr. Md. Mahbub Chowdhury Mishu
Faculty of Computer Science & Engineering
American International University, Bangladesh

Diabetes Prediction (UCI Machine Learning Repository)

Introduction:

Data is just a gathered file of information about various people and there behaviours. Data Information and Knowledge these are three different things. But each of them is inter-related, and each term has its own meaning. According to a common view, data is collected and analyzed; data only becomes information suitable for making decisions once it has been analyzed in some fashion.

Data Set Characteristics: Multivariate

Number of Instances: 768

Area: Disease, Predictive models, Medical Data Mining

Attribute Characteristics: Integer/Categorical

Number of Attributes: 9

Associated Tasks: Classification

Missing Values? No

File formats: diabetes.csv

Source:

Pima Indians Diabetes Database, Predicts the onset of diabetes based on diagnostic measures.

https://www.kaggle.com/uciml/pima-indians-diabetes-database?select=diabetes.csv

Dataset Information:

This is an medical data set which is collected UCI Machine Learning Repository, which has been designed to facilitate machine learning through the supply of various kinds of datasets. Such system provides users with a synchronous access to educational resources from any device with Internet connection.

The dataset consists of 768 person records and 7 attributes. The features are classified into one major category that is Demographic features such as Age, Blood Pressure.The dataset consists of 500 patients that are tested negative and 268 are tested positive. We've age between 21 to 81. Where we can find most of them are 21 and there are 63 person aged 21. We have pregnancy data and Insulin reaction of the person to ensure which type of diabetes is present. Blood glucose level and Blood Pressure level , Skin thickness etc. Which will help us to predict diabetes patients.

Attributes

1 Pregnancies- How many times the patient got pregnant (Numeric values).
2 Glucose- Level of glucose in the patient body. (Numeric)

3 Blood Pressure- Level of blood pressure of the patient. (Numeric)

4 Skin Thickness- How thick is the skin of patient body.(Numeric)

5 Insulin- How many mg Insulin is produced in the patient body. (Numeric)

6 Age- Patient age (Numeric)

7 BMI - Body mass index weight and height. (Numeric)

Relevant Papers:

Prediction of Type 2 Diabetes using Machine Learning Classification Methods Author Neha Prerna Tiggaa ShrutiGarg

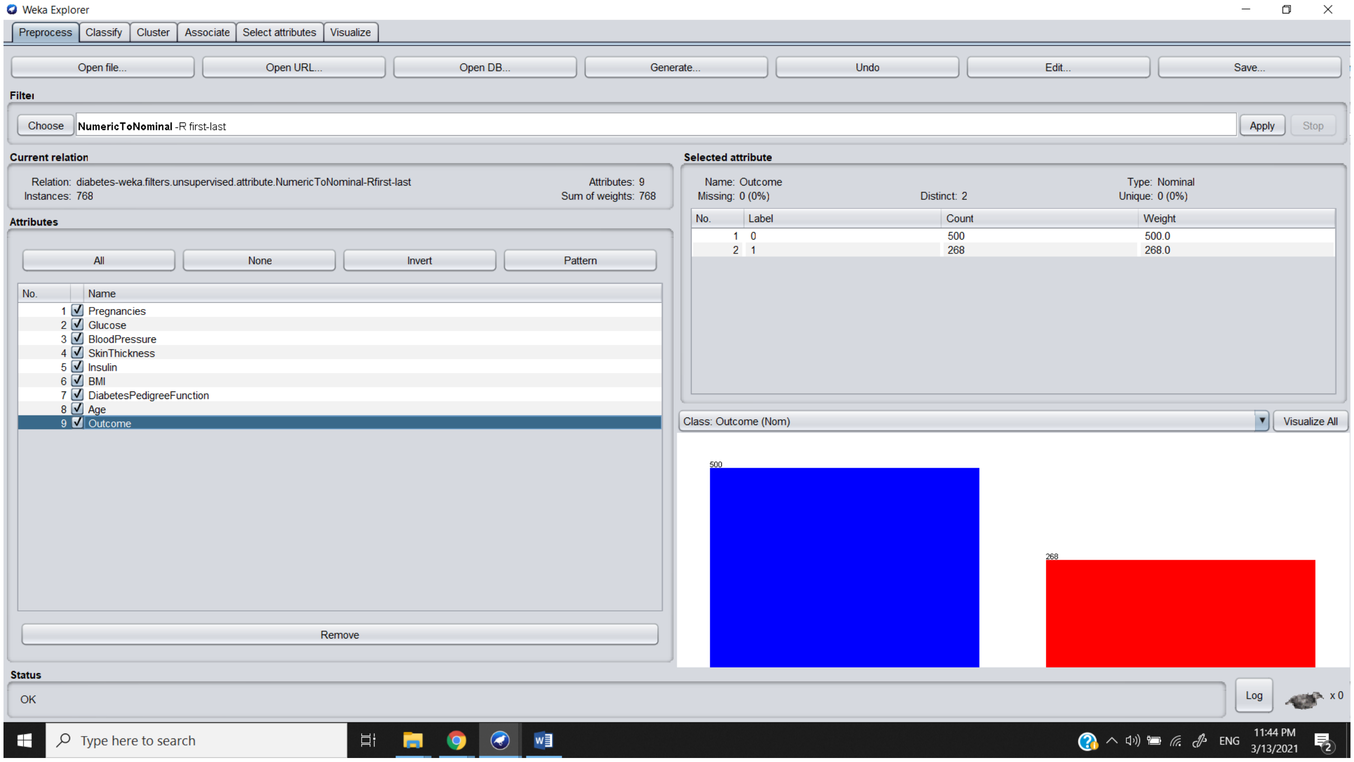
https://www.sciencedirect.com/science/article/pii/S1877050920308024

Diabetes Prediction using Machine Learning Algorithms

Aishwarya Mujumdar, V Vaidehi Dr.

https://www.sciencedirect.com/science/article/pii/S1877050920300557

At first, Weka software was lunched and from the Weka GUI chooser Explorer was chosen. Then the Weka Explorer window was opened, from that window open file option was pressed. After that, "diabetes.csv" was selected which data type was .csv type to load in Weka Explorer for applying different classification algorithms.



Filter: NumericToNominal -R first-last

Relation: diabetes-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 768
Attributes: 9
Sum of weights: 768

No.		Name
1	☑	Pregnancies
2	☑	Glucose
3	☑	BloodPressure
4	☑	SkinThickness
5	☑	Insulin
6	☑	BMI
7	☑	DiabetesPedigreeFunction
8	☑	Age
9	☑	Outcome

Name: Outcome
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

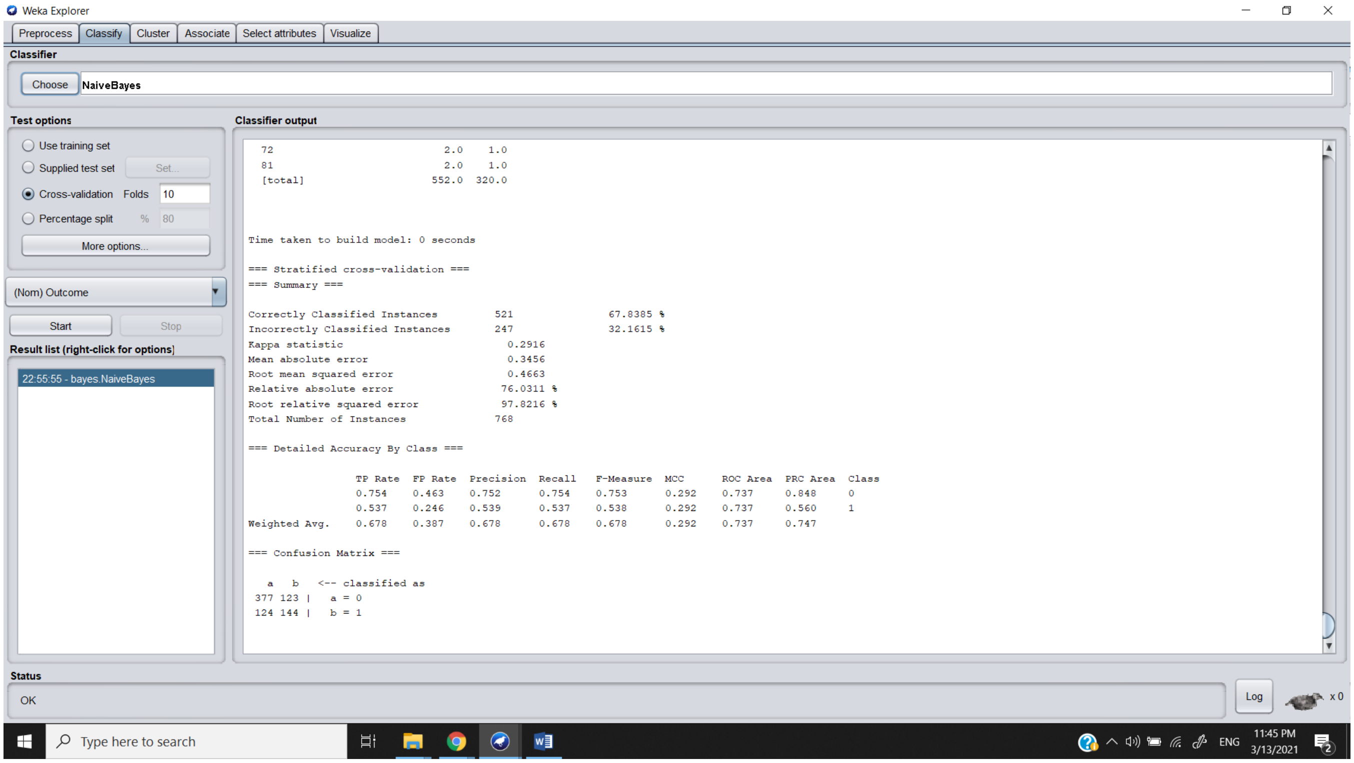
No.	Label	Count	Weight
1	0	500	500.0
2	1	268	268.0

Class: Outcome (Nom)

Status: OK

Figure: Loaded dataset into Weka

After opening the desired dataset into Weka, Classify tab was chosen. In the classify tab Crossvalidation folds 8 was chosen as test options for all three classifier. Which meant the data was folded in 8 and the following process was repeated 8 times (because it was 8 folds): 7 folds was used for learning and I folds was left for testing. Every time a different fold was left for testing. At first Naïve Bayes classifier algorithm was chosen from the classifier menu. Then the start button was pressed to start classifying the data. After classification the summary of the output was shown in in the classifier output area.



Classifier: NaiveBayes

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 80

(Nom) Outcome

Result list (right-click for options)

22:55:55 - bayes.NaiveBayes

Status: OK

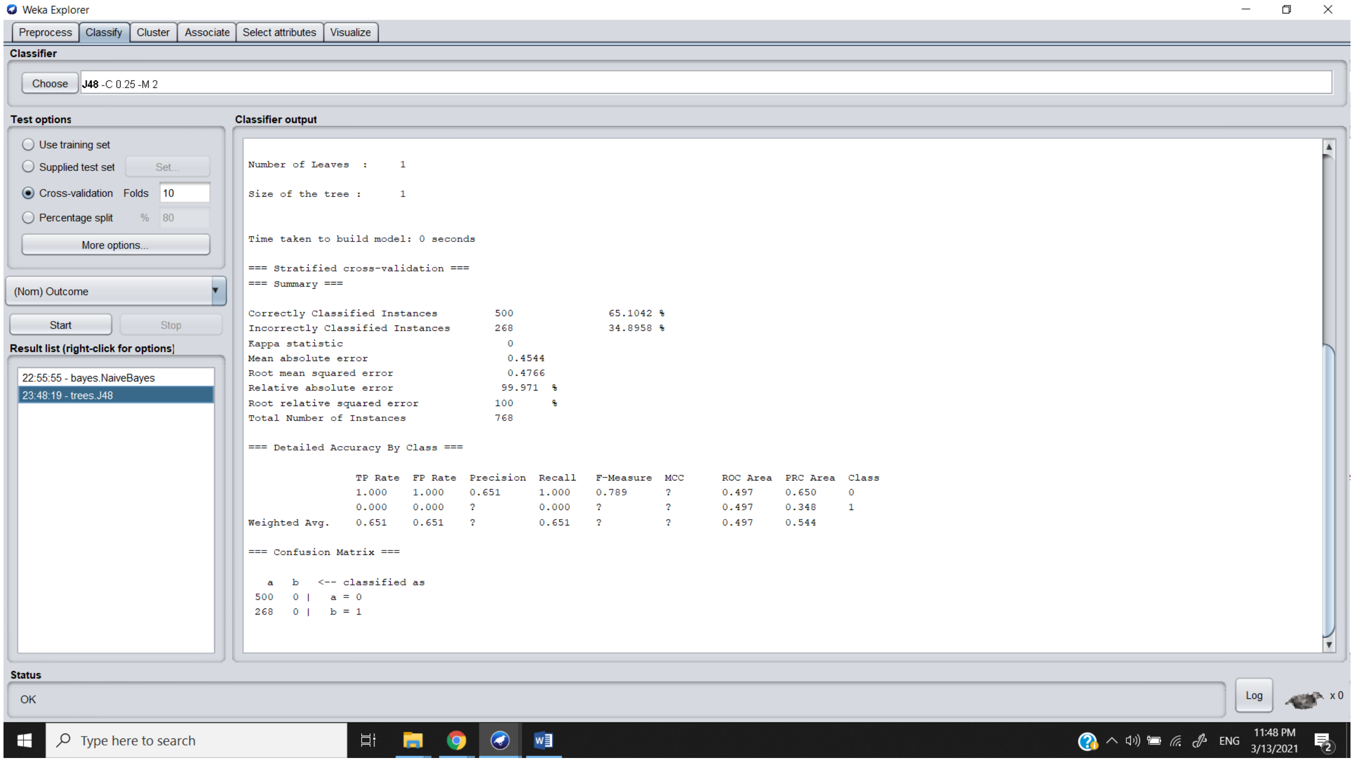
Classifier output

```
  72                          2.0    1.0
  81                          2.0    1.0
  [total]                   552.0  320.0



Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         521               67.8385 %
Incorrectly Classified Instances       247               32.1615 %
Kappa statistic                          0.2916
Mean absolute error                      0.3456
Root mean squared error                  0.4663
Relative absolute error                 76.0311 %
Root relative squared error             97.8216 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.754    0.463    0.752      0.754    0.753      0.292    0.737     0.848     0
                 0.537    0.246    0.539      0.537    0.538      0.292    0.737     0.560     1
Weighted Avg.    0.678    0.387    0.678      0.678    0.678      0.292    0.737     0.747

=== Confusion Matrix ===

   a   b   <-- classified as
 377 123 |   a = 0
 124 144 |   b = 1
```

Figure: Output of Naïve Bayes Classifier Algorithm

From the output we can see that, correctly classified instances is 768, incorrectly classified instances is 247 which is 32.1615% of total instances. The accuracy of correctly classified instances 67.8385%. Mean absolute error: 0.3456%

After that J48 classifier algorithm was chosen from the classifier menu. Then the start button was pressed to start classifying the data.



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 80

(Nom) Outcome

22:55:55 - bayes.NaiveBayes
23:48:19 - trees.J48

```
Number of Leaves  :     1

Size of the tree :      1


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         500               65.1042 %
Incorrectly Classified Instances       268               34.8958 %
Kappa statistic                          0
Mean absolute error                      0.4544
Root mean squared error                  0.4766
Relative absolute error                 99.971  %
Root relative squared error            100      %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.651      1.000    0.789      ?        0.497     0.650     0
                 0.000    0.000    ?          0.000    ?          ?        0.497     0.348     1
Weighted Avg.    0.651    0.651    ?          0.651    ?          ?        0.497     0.544

=== Confusion Matrix ===

   a   b   <-- classified as
 500   0 |   a = 0
 268   0 |   b = 1
```

Status: OK

Figure: Output of J48 Classifier Algorithm

From the output we can see that, correctly classified instances is 768, incorrectly classified instances is 268 which is 34.8958% of total instances. The accuracy of correctly classified instances 65.1042%. Mean absolute error: 0.4766%

From the summary of the output we can see that,
Correctly classified instances: 500
Incorrectly classified instances: 268
Accuracy of correctly classified instances: 65.1042%
Accuracy of incorrectly classified instances: 34.8958%
Mean absolute error: 0.4766
Comparing the outputs of Naïve Bayes, J48 and Input Mapped Classifier we can see that classifier correctly classified 116 instances out of 364 instances. J48 classifier algorithm also gives least amount of mean absolute error among all three classifier algorithms. But Naive Bayes classifier gives the accuracy of correctly classified instances 67.8385% which is best accuracy among all three classifier algorithms for the selected dataset.
So, we can conclude that Naive Bayes classifier algorithm was best for classification of the selected dataset