DIABETES AND DATA MINING

HAP 780
SEMESTER PROJECT
Data Mining in Healthcare

By Group 8
Nihar Kaluvala
Swetha Meruva
Tejaswi Pulusu
Vineeth Reddy Gunreddy

Under the Guidance of Professor Hong Xue

Master's in Health Informatics
College of Public Health

4400 University Dr, Fairfax, VA 22030

I. INTRODUCTION

Diabetes is a chronic metabolic disease characterized by elevated blood glucose (or blood sugar) levels that over time lead to severe damage to the heart, blood vessels, eyes, kidneys, and nerves. Type 2 diabetes is the most common and is usually seen in adults and occurs when the body becomes resistant or does not produce enough insulin [(1)]. According to the 2022 National Diabetes Statistics available by the Centers for Disease Control and Prevention (CDC), there are more than 130 million persons in the US who have diabetes or pre-diabetes [(2)]. Adults with diabetes are at a 2-4 times high risk to develop vascular diseases than healthy adults [(3)]. Chronic complications of DM include retinopathy, neuropathy, nephropathy, increased risk of cardiovascular disease, and serious cardiac events such as myocardial infarction and stroke. Due to the high prevalence of DM and its complications, it is a common comorbidity in hospitalized patients. This results in longer hospital stays (LOS) in DM patients, increased in-hospital complications, and frequent procedural and procedural hospitalizations reported to have increased mortality. According to reports, hospitalized patients with DM had 30-day readmission rates that range between 14.4 and 22.7%, which is much higher than the average rate for hospitalized patients (8.5-13.5%) [(4)].

OBJECTIVES

Primary Objective: To predict the readmission rates in diabetic patients using data mining techniques.
Secondary Objective: To predict the presence of circulatory disease in diabetic patients.

LITERATURE REVIEW

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records", *BioMed Research International*, vol. 2014, Article ID 781670, 11 pages, 2014. https://doi.org/10.1155/2014/781670.

The importance of protecting the body from hyperglycemia cannot be underestimated; the direct and indirect effects on the human vascular tree are the major source of morbidity and mortality in both types of diabetes3. Diabetes mellitus is a common comorbid illness among hospitalized patients due to its high prevalence. Governmental organizations and healthcare systems have recently placed a greater emphasis on 30-day readmission rates to assess the complexity of their patient populations and to raise standards of care5. The early diagnosis and treatment of type 2 diabetes are among the most relevant actions to prevent further development and complications. A sensitivity analysis of USA data proved a 25% relative reduction in diabetes-related complication rates for a 2-year earlier diagnosis. Consequently, many researchers have endeavored to develop predictive models of type 2 diabetes. The first models were based on classic statistical learning techniques, e.g., linear regression. Recently, a wide variety of machine learning techniques has been added to the toolbox. The number of studies developed in the field creates two main challenges for researchers and developers aiming to build type 2 diabetes predictive models. First, there is considerable heterogeneity in previous studies regarding machine learning techniques used, making it challenging to identify the optimal one. Second, there is a lack of transparency about the features used to train the models, which reduces their interpretability, a feature utterly relevant to the doctor [(10)].

II. MATERIALS AND METHODS

DATA SOURCE

The data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grants UL1 TR00058 and a recipient of the CERNER data [(5)]. The data is de-identified from the Health Facts Database (Cerner Corporation, Kansas City, MO)

DATA DICTIONARY

This dataset represents 10 years (1999-2008) of clinical care in 130 US hospitals and integrated delivery networks. Includes over 50 features to plot patient and hospital outcomes. Information was extracted from the database for encounters meeting the following criteria:

It is an inpatient (hospitalized). Diabetes Encounter, i.e., Diabetes of some type entered the system as a diagnosis. The period of stay is between 1 day and 14 days. Laboratory tests were conducted during the encounter. Medication was administered during the encounter.

Data includes patient ID, race, gender, age, type of hospitalization, duration of hospitalization, a specialty of hospitalized physician, number of laboratory tests performed, HbA1c test result, diagnosis, number of medications, diabetes medication, hospitalization Number of outpatients, hospitalizations, emergency consultations, etc. in the previous year

DATA PREPROCESSING

Data preprocessing is an iterative process that is used to transform the raw data into understandable and usable data. The data we obtained from the source was not collected in a standardized way, so it is important to transform the data into the best form that is suited to evaluate the objectives of the study. The preprocessed then obtained using SQL is used to perform the predictive analysis.

Preprocessing Steps

Aggregation: Original data consists of 3 diagnoses which are primary, secondary, and tertiary coded with ICD9. All three attributes are combined into a single attribute in accordance with the objective.

Data Reduction: Removed the attributes with higher missing values such as weight (97%), payer code (40%), and medical specialty (47%) to improve the processing time and memory and reduce noise in the data.

Feature Creation: New attributes “Is Diabetic” and “Is Circulatory” is created to allow maximum information gain from the data.

Discretization: Nominal attributes “Is Diabetic and Is Circulatory” are binarized.

Mapping: The original data with the ICD9 codes are mapped to the description tables to obtain the diagnosis in the form of a description instead of the ICD9 which is difficult to understand. The Mapping is done using Python.

Outlier Detection: All the outliers are detected in the data and removed for better outcomes.

FEATURE SELECTION

Feature selection is used to select the attributes to be used to train the machine learning models. In WEKA, we used different attribute selection methods both supervised and unsupervised methods. The correlational method along with the ranker produced a better result than other methods such as Info Gain and Wrapper method. The Correlational ranker method gives high correlation features which are more linearly dependent and hence virtually equally affect the dependent variable. It produced 14 best attributes out of all the attributes in the original data shown in Figure 3.

DESCRIPTIVE ANALYSIS

Dependent Variables: Diabetic patients readmitted to the hospital is a categorical variable in which subjects with less than, and greater than 30 days are recorded as ‘< 30-day readmission’ and ‘> 30-day readmission’ respectively shown in Figure 3. Patients with no readmission are recorded as ‘NO’ in the readmission. Is circulatory is the dependent categorical variable of the secondary objective.

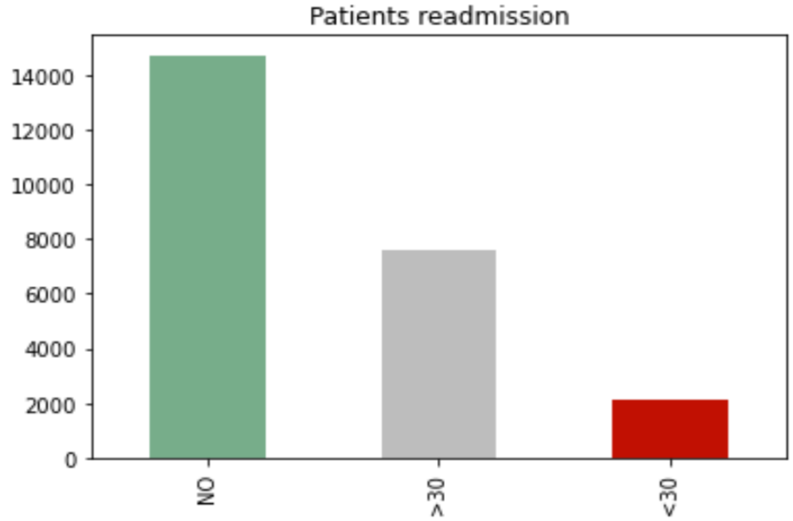


Figure 1. Distribution of Diabetic Patient's Readmission

Independent Variables: Age, race, gender, serum maximum glucose level, change in medication, time in hospital, HbA1c, insulin status, and admission type are the independent variables. Age, race, gender, HbA1c, change in medication, insulin status, and admission type are categorical variables.

Ages are divided into 10 groups with the highest percent 22.08 in 60-70 ages. Most of the people are between the ages of 50 to 80 years Table (1). About 72.41 percent of subjects belong to the Caucasian race, 22.41 percent of subjects are African- Americans and only a few percent are Asian and Hispanic of the total subjects Table (2). Female subjects are 53.29 slightly greater than that of the male percent which is 46.71. Serum maximum glucose levels are divided into three categories: none, >300, and > 200, most percent of subjects have normal serum glucose levels which are marked as "none".

Table 1. Age Distribution among Subjects.

Age	Counts	Percent
0-10	60	0.25
10-20	299	1.22
20-30	724	2.96
30-40	1518	6.20
40-50	3279	13.39
50-60	5085	20.77
60-70	5407	22.08
70-80	5074	20.72
80-90	2659	10.86
90-100	380	1.55

Table 2. Race Distribution among the Subjects

Race	Counts	Percent
African American	5486	22.41
Asian	166	0.66
Caucasian	17729	72.41
Hispanic	659	2.69
Other	450	1.84

Admission type is categorized into 4 categories which are urgent care, elective, newborn, and trauma. Insulin status is divided into 3 categories on steady insulin, decreased insulin input, and not on insulin. Time in the hospital is a continuous variable varying from 2 days to > 30 days.
Is diabetic being a categorical independent variable.

III. RESULTS

Using WEKA, different data mining techniques are used on the training data in predicting the outcome variable of the two objectives of the study.

The Association mining Apriori Algorithm is used to find the best association rules using the frequent item sets. The data used produced 10 best association rules a minimum confidence of 0.9 and a minimum support of 0.7. But the rules obtained had similar attributes serum maximum glucose, diabetic medications, race, and HbA1c are repeated in the rules having a correlation with the diabetic attributes and vice versa as shown in the figure 4.
Methods used for the primary objective which is to predict the readmission rate in diabetic patients are Naïve Bayes, Random Forest, J48, and Logistic Regression.

Naïve Bayes is the best-suited algorithm while predicting outcome variables with multiple independent variables and it performs better with categorical input variables hence this model is selected. Since the assumption of independent variables holds true, Naïve Bayes performs better in comparison to other models. The results of the Naïve Bayes for the readmission rate >30 days have a ROC area value of 0.577 and that of the readmission rate < 30 days has a value of 0.604. The weighted average true positive is 0.597. F-score is 0.504 for the weighted average of the three groups in the dependent variable shown in Figure 5. The weighted average true positive value is 0.597 which means the model can predict a 59.7% outcome of the positive class correctly.

Random Forest is a machine-learning technique that can solve both regression and classification problems. This algorithm consists of many decision trees where prediction is made from the average or the mean of the output from these trees. It reduces the over fitting of the datasets and increases the precision. The results obtained from the random forest algorithm have a ROC area value of 0.526 for readmission rate > 30 days which is less than that obtained from the Naïve Bayes and the values for the readmission rate < 30 days is 0.526 which is also less than the value of Naïve Bayes. F-score is 0.501 for the weighted average of the three groups in the dependent variable shown in Figure 6. The true positive for the model is 0.540 which is less than that of the Naïve Bayes.

J48 is one of the best machine learning algorithms to examine the data categorically and continuously which can occupy more memory space increasing its performance and accuracy, especially in medical data. F- score of this model was not obtained by the algorithm because of extremely poor prediction of WEKA for some classes of the dataset. In Figure 7, the confusion matrix has only values for the second group of the dependent variable which is for no readmission but not for the other two groups showing the unavailability of the data to perform the predictive metrics. The true positive value is 0.605 as the algorithm considered only the patients with the readmission but not the other two groups which make it inefficient to be compared to the other predictive model.

Logistic Regression is a supervised machine learning algorithm used to predict the probability of a binary event occurring. It is used to predict whether a patient with diabetes has readmission. The F-score of the model is 0.482 which is slightly less than that of the other two models, but the true positive value is 0.598 which is greater than the other two predictive models shown in Figure 8.

Overall findings of the analysis:
We may infer from the ROC observations discussed above that the model's accuracy does not differ significantly. Logistic regression followed by Naive Bayes has higher accuracy levels compared to other selected models based on ROC values while Naive Bayes has better precision than Logistic regression shown in Table 3.

Table 3: Comparison Table of the findings of Analysis I

Model	Precision	Recall	ROC
Naive Bayes	0.508	0.597	0.593
Random Forest	0.483	0.540	0.536
J48	?	0.605	0.500
Logistic Regression	0.483	0.598	0.594

Methods used for the secondary objective which is to predict the patient with circulatory disease alongside diabetes are Naïve Bayes, Random Tree, and Sequential Minimal Optimization.

Naïve Bayes is the best algorithm when predicting the outcome variable with multiple independent variables, and it works better with categorical input variables, so this model was chosen. Naive Bayes outperforms other models because the independent variable assumptions hold. The results obtained from this model are shown in Figure 9.

Random Tree is used to utilize bootstrap sampling to select a subset of predictors at random and use the best one to forecast tree nodes. Unlike other models, the random tree classifier can handle both categorical and numerical values. We used the random tree model as the data include categorical variables. However, the random tree classifier is less computationally intensive, but it performs faster and better with large datasets. The model has the same results as the Naïve Bayes shown in Figure 10.

Sequential Minimal Optimization: SMO works by splitting the quadratic problem into little problems, and SMO trains the support vector machine to solve it. SMO stores memory in a linear manner, allowing it to train big data sets while avoiding expensive matrix computation. Most of the articles referred to data mining techniques in diabetes prediction preferred to use SMO, results are shown in Figure 11.

Though three different algorithms were used to predict the presence of circulatory disease in diabetic patients, the results obtained are the same with all three models. The only difference that can be seen is the time taken to train the model and evaluate the test data supplied to the model.

Overall findings of the analysis:

In the model comparison shown in Table 4., we can see that the ROC values are the same for all models selected which is 0.50 indicating that models have the same accuracy.

Table 4: Comparison Table of Findings of Analysis II

Model	ROC
Naive Bayes	0.50
Random Tree	0.50
SMO	0.50

IV. DISCUSSION

The primary objective of this study is based on a single diagnosis whereas the existing approach of some studies was based on the primary diagnosis and HbA1c measurement. We are aware that the results of the current research are preliminary and have inherent limitations due to the sheer volume of the health information. This study is constrained by a nonrandomized study design in addition to the drawbacks of working with big clinical datasets. It may not come as a surprise that patients with main diagnoses of diabetes mellitus received less attention to their diabetes management than those with admitting diagnoses of circulatory disease. However, our results strongly imply that giving these high-risk patients' diabetes treatment more attention while they are in the hospital may have a major impact on readmission.

Strengths and Limitations: All the major missing data columns are removed to ensure precision and accuracy. The data is divided into training and tested randomly to train the models which reduce the over fitting. Dataset is huge to obtain maximum information. The dataset used has valuable but heterogenous difficult data in terms of missing, incomplete, inconsistent, and complex attributes. The type of information collected was not under control. Race in the study sample was not uniformly distributed; hence there may be bias of the dominant racial group among the sample. The relationship between the attributes and the outcome variables was not statistically significant. Lack of expertise in data mining and dataset selection. While preprocessing, we didn't use feature subset selection which could have improved the model performance.

V. CONCLUSION

The readmission rate of diabetic patients has been increasing around the world; the patients with diabetes are also prone to develop circulatory diseases. Our study used the data obtained from 130 hospitals that include information needed to predict the readmission rate and to see the presence of circulatory disease in diabetic patients. The use of machine learning techniques improves prediction. Different predictive models or algorithms are used to predict the readmission rate in diabetic patients. Of all the models, logistic regression showed better results compared to Naïve Bayes and the Random Forest for the primary objective. The secondary objective which is the prediction of the presence of circulatory disease in the diabetes patient showed the same results for all the predictive models. From the results, we can conclude that more significant attributes may produce better results for the primary analysis, and an increase in the number of attributes may increase the precision and accuracy of the models in the secondary analysis.

REFERENCE

1. *WHO. (2022). Diabetes. World Health Organization.*https://www.who.int/health topics/diabetes#tab=tab_1

2. *National DPP Customer Service Center. (2022). Cdc.gov. https://nationaldppcsc.cdc.gov/s/article/CDC-2022-National-Diabetes-Statistics-Report*

3. *Diabetes and Heart Disease*. (n.d.). Www.hopkinsmedicine.org. https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-heart-disease

4. Fowler, M. J. (2008, April 1). *Microvascular and Macrovascular Complications of Diabetes*. American Diabetes Association. https://diabetesjournals.org/clinical/article/26/2/77/1823/Microvascular-and-Macrovascular-Complications-of

5. *Krinsley JS. Association between hyperglycemia and increased hospital mortality in heterogeneous population of critically ill patients.* https://pubmed.ncbi.nlm.nih.gov/11889147/

6. *Ostling, S., Wyckoff, J., Ciarkowski, S. L., Pai, C.-W., Choe, H. M., Bahl, V., & Gianchandani, R. (2017). The relationship between diabetes mellitus and 30-day readmission rates. Clinical Diabetes and Endocrinology, 3(1). https://doi.org/10.1186/s40842-016-0040-x*

7. *UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008#*

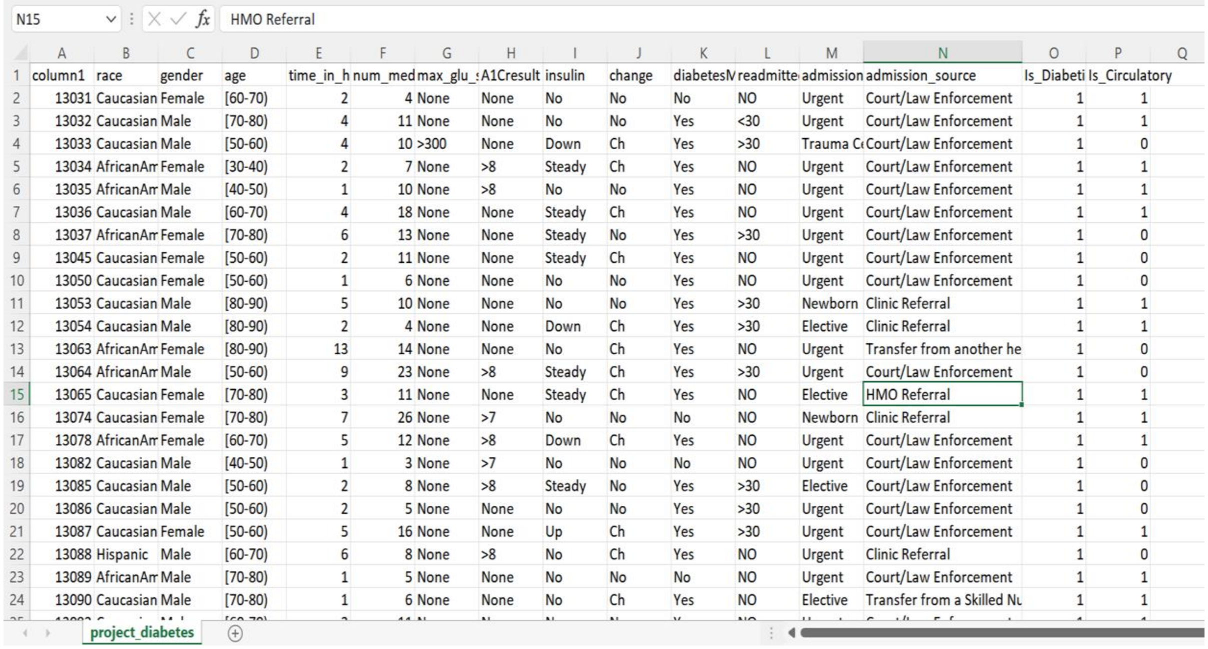
8. *Introduction to Random Forest in Machine Learning*. (n.d.-a). Engineering Education (EngEd) Program | Section. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

9. *International Journal of Computational ... - periyar university*. (n.d.). Retrieved December 10, 2022, from https://www.periyaruniversity.ac.in/ijcii/issue/marnew/2_mar_18.pdf

10. Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetology & Metabolic Syndrome, 13(1). https://doi.org/10.1186/s13098-021-00767-9

APPENDIX

Figure 2. Data obtained from the Preprocessing.

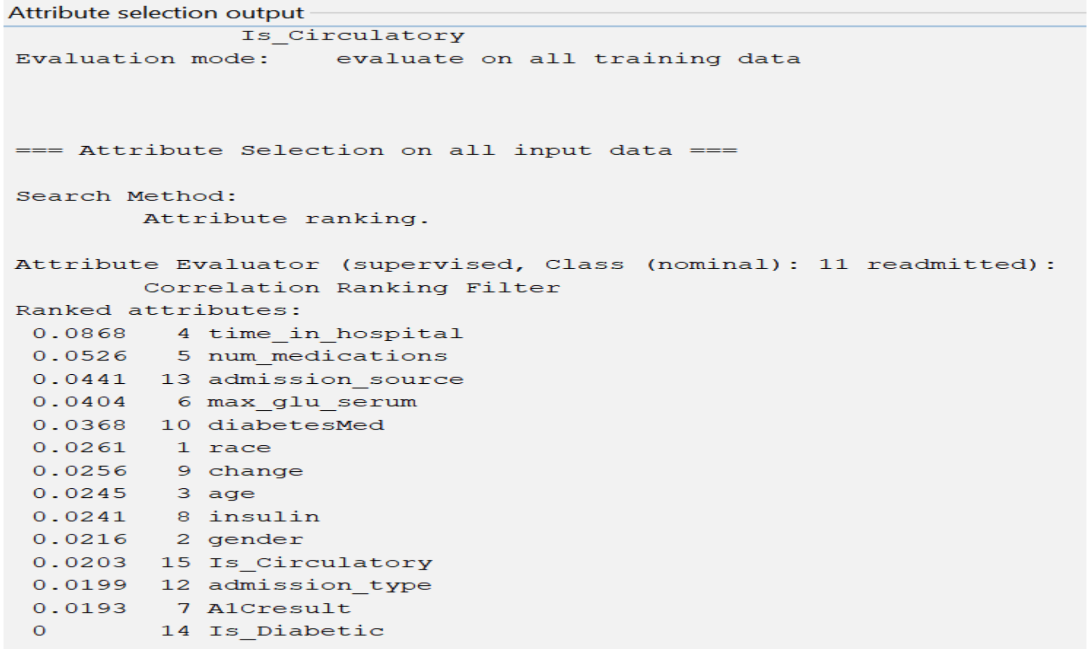


N15 HMO Referral

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	column1	race	gender	age	time_in_h	num_med	max_glu_:	A1Cresult	insulin	change	diabetesM	readmitte	admission	admission_source	Is_Diabeti	Is_Circulatory	
2	13031	Caucasian	Female	[60-70)	2	4	None	None	No	No	No	NO	Urgent	Court/Law Enforcement	1	1	
3	13032	Caucasian	Male	[70-80)	4	11	None	None	No	No	Yes	<30	Urgent	Court/Law Enforcement	1	1	
4	13033	Caucasian	Male	[50-60)	4	10	>300	None	Down	Ch	Yes	>30	Trauma C(	Court/Law Enforcement	1	0	
5	13034	AfricanAr	Female	[30-40)	2	7	None	>8	Steady	Ch	Yes	NO	Urgent	Court/Law Enforcement	1	1	
6	13035	AfricanAr	Male	[40-50)	1	10	None	>8	No	No	Yes	NO	Urgent	Court/Law Enforcement	1	1	
7	13036	Caucasian	Male	[60-70)	4	18	None	None	Steady	Ch	Yes	NO	Urgent	Court/Law Enforcement	1	1	
8	13037	AfricanAr	Female	[70-80)	6	13	None	None	Steady	No	Yes	>30	Urgent	Court/Law Enforcement	1	0	
9	13045	Caucasian	Female	[50-60)	2	11	None	None	Steady	Ch	Yes	NO	Urgent	Court/Law Enforcement	1	0	
10	13050	Caucasian	Female	[50-60)	1	6	None	None	No	No	Yes	NO	Urgent	Court/Law Enforcement	1	0	
11	13053	Caucasian	Male	[80-90)	5	10	None	None	No	No	Yes	>30	Newborn	Clinic Referral	1	1	
12	13054	Caucasian	Male	[80-90)	2	4	None	None	Down	Ch	Yes	>30	Elective	Clinic Referral	1	1	
13	13063	AfricanAr	Female	[80-90)	13	14	None	None	No	Ch	Yes	NO	Urgent	Transfer from another he	1	0	
14	13064	AfricanAr	Male	[50-60)	9	23	None	>8	Steady	Ch	Yes	>30	Urgent	Court/Law Enforcement	1	0	
15	13065	Caucasian	Female	[70-80)	3	11	None	None	Steady	Ch	Yes	NO	Elective	HMO Referral	1	1	
16	13074	Caucasian	Female	[70-80)	7	26	None	>7	No	No	No	NO	Newborn	Clinic Referral	1	1	
17	13078	AfricanAr	Female	[60-70)	5	12	None	>8	Down	Ch	Yes	NO	Urgent	Court/Law Enforcement	1	1	
18	13082	Caucasian	Male	[40-50)	1	3	None	>7	No	No	No	NO	Urgent	Court/Law Enforcement	1	0	
19	13085	Caucasian	Male	[50-60)	2	8	None	>8	Steady	No	Yes	>30	Elective	Court/Law Enforcement	1	0	
20	13086	Caucasian	Male	[50-60)	2	5	None	None	No	No	Yes	>30	Urgent	Court/Law Enforcement	1	0	
21	13087	Caucasian	Female	[50-60)	5	16	None	None	Up	Ch	Yes	>30	Urgent	Court/Law Enforcement	1	1	
22	13088	Hispanic	Male	[60-70)	6	8	None	>8	No	Ch	Yes	NO	Urgent	Clinic Referral	1	0	
23	13089	AfricanAr	Male	[70-80)	1	5	None	None	No	No	No	NO	Urgent	Court/Law Enforcement	1	1	
24	13090	Caucasian	Male	[70-80)	1	6	None	None	No	Ch	Yes	NO	Elective	Transfer from a Skilled Nu	1	1	

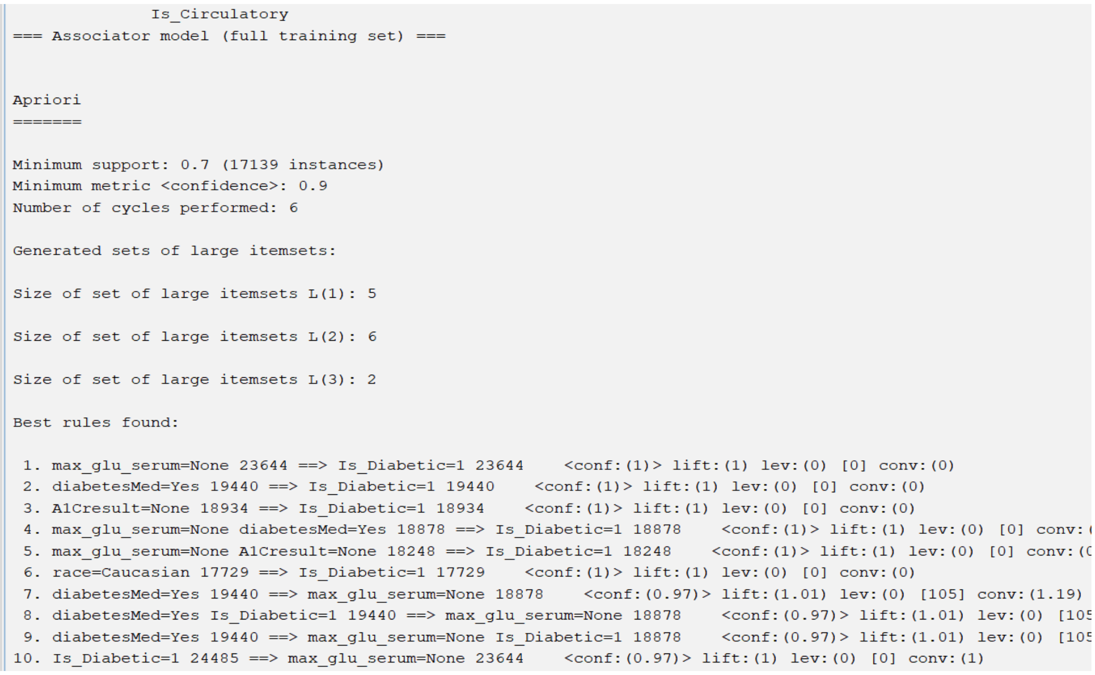
FEATURE SELECTION:

Figure 3. Attribute Selection by Correlational + Ranker



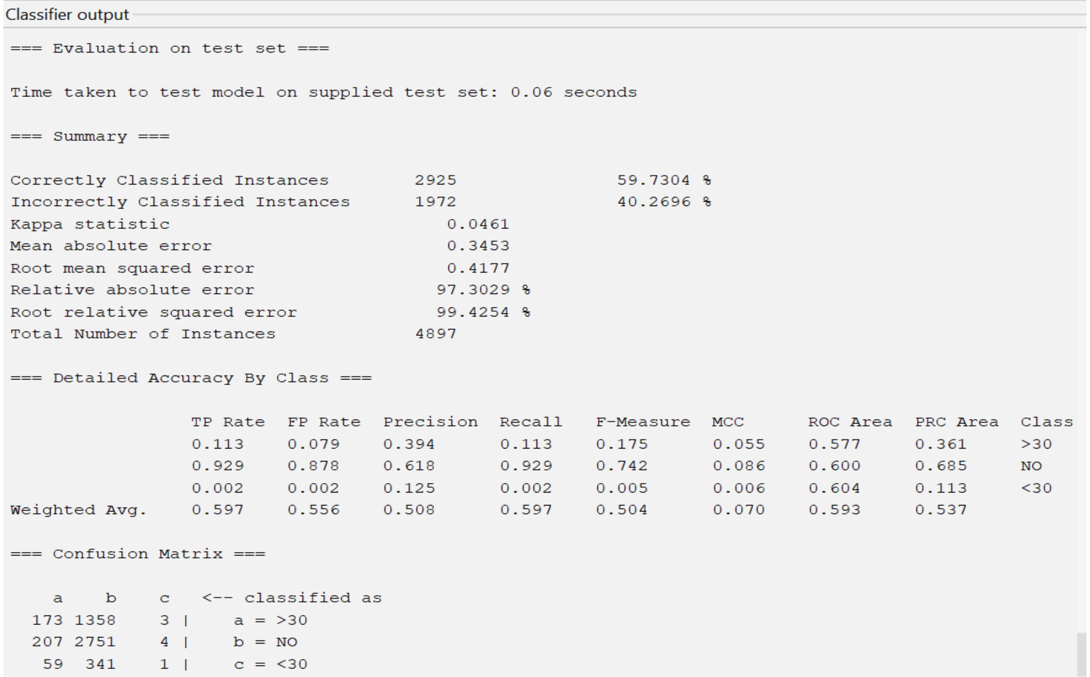
```
Attribute selection output
                    Is_Circulatory
Evaluation mode:     evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 11 readmitted):
        Correlation Ranking Filter
Ranked attributes:
 0.0868    4 time_in_hospital
 0.0526    5 num_medications
 0.0441   13 admission_source
 0.0404    6 max_glu_serum
 0.0368   10 diabetesMed
 0.0261    1 race
 0.0256    9 change
 0.0245    3 age
 0.0241    8 insulin
 0.0216    2 gender
 0.0203   15 Is_Circulatory
 0.0199   12 admission_type
 0.0193    7 A1Cresult
 0         14 Is_Diabetic
```

Figure 4. Apriori Algorithm used for Association Mining:



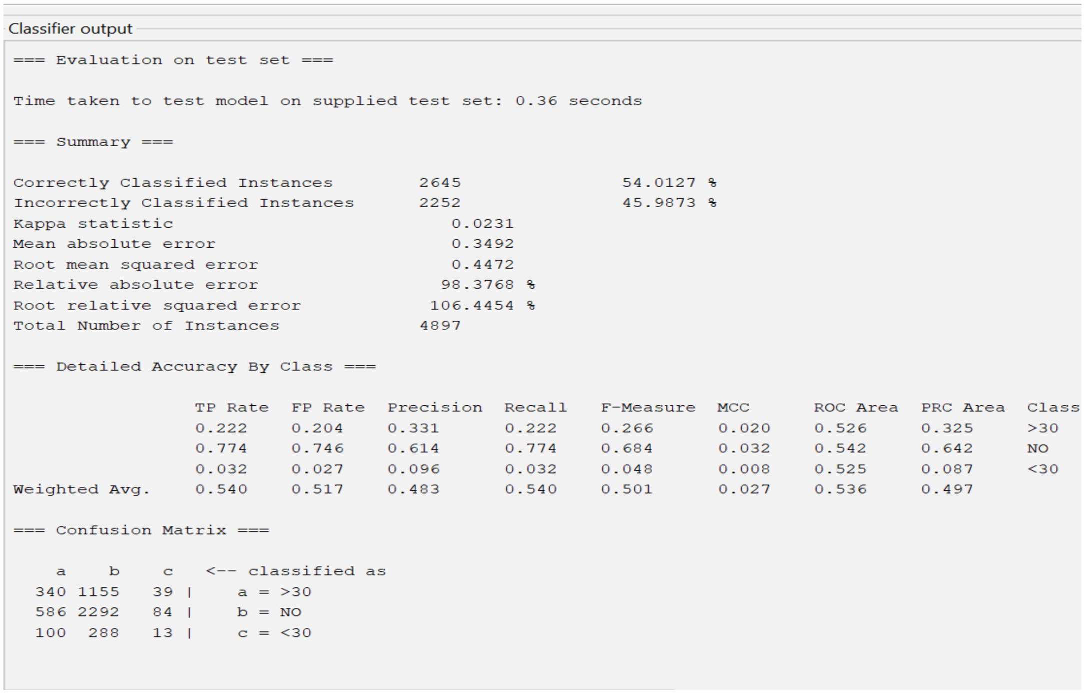
```
              Is_Circulatory
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.7 (17139 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 6

Size of set of large itemsets L(3): 2

Best rules found:

 1. max_glu_serum=None 23644 ==> Is_Diabetic=1 23644    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. diabetesMed=Yes 19440 ==> Is_Diabetic=1 19440    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. A1Cresult=None 18934 ==> Is_Diabetic=1 18934    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. max_glu_serum=None diabetesMed=Yes 18878 ==> Is_Diabetic=1 18878    <conf:(1)> lift:(1) lev:(0) [0] conv:(
 5. max_glu_serum=None A1Cresult=None 18248 ==> Is_Diabetic=1 18248    <conf:(1)> lift:(1) lev:(0) [0] conv:(
 6. race=Caucasian 17729 ==> Is_Diabetic=1 17729    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. diabetesMed=Yes 19440 ==> max_glu_serum=None 18878    <conf:(0.97)> lift:(1.01) lev:(0) [105] conv:(1.19)
 8. diabetesMed=Yes Is_Diabetic=1 19440 ==> max_glu_serum=None 18878    <conf:(0.97)> lift:(1.01) lev:(0) [105
 9. diabetesMed=Yes 19440 ==> max_glu_serum=None Is_Diabetic=1 18878    <conf:(0.97)> lift:(1.01) lev:(0) [105
10. Is_Diabetic=1 24485 ==> max_glu_serum=None 23644    <conf:(0.97)> lift:(1) lev:(0) [0] conv:(1)
```

Figure 5. Naïve Bayes Predictive Model Results of Analysis I:



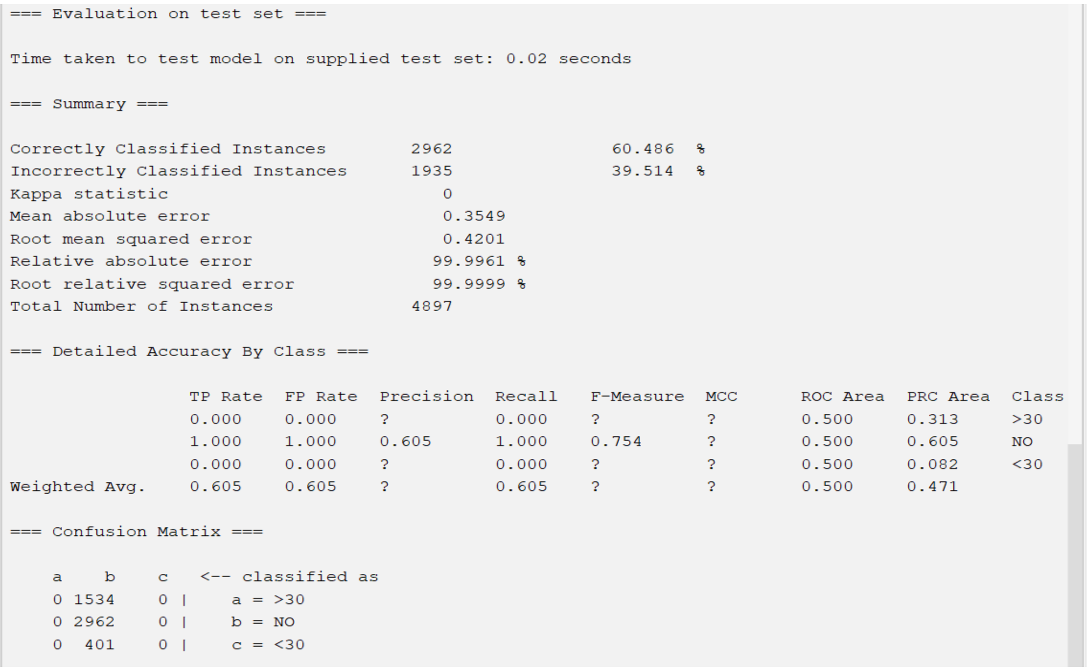
```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.06 seconds

=== Summary ===

Correctly Classified Instances        2925               59.7304 %
Incorrectly Classified Instances      1972               40.2696 %
Kappa statistic                          0.0461
Mean absolute error                      0.3453
Root mean squared error                  0.4177
Relative absolute error                 97.3029 %
Root relative squared error             99.4254 %
Total Number of Instances             4897

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.113    0.079    0.394      0.113    0.175      0.055    0.577     0.361     >30
                 0.929    0.878    0.618      0.929    0.742      0.086    0.600     0.685     NO
                 0.002    0.002    0.125      0.002    0.005      0.006    0.604     0.113     <30
Weighted Avg.    0.597    0.556    0.508      0.597    0.504      0.070    0.593     0.537

=== Confusion Matrix ===

    a    b    c   <-- classified as
  173 1358    3 |    a = >30
  207 2751    4 |    b = NO
   59  341    1 |    c = <30
```

Figure 6. Random Forest Predictive Model Results of Analysis I:



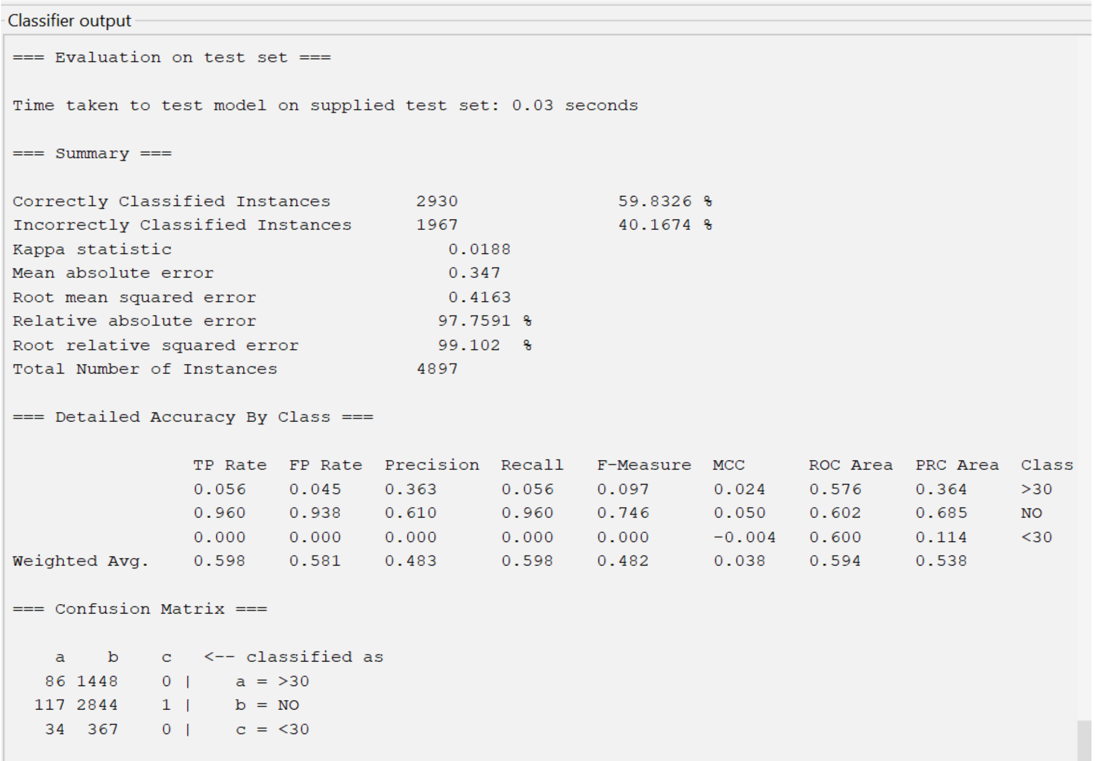
```
Classifier output
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.36 seconds

=== Summary ===

Correctly Classified Instances        2645               54.0127 %
Incorrectly Classified Instances      2252               45.9873 %
Kappa statistic                          0.0231
Mean absolute error                      0.3492
Root mean squared error                  0.4472
Relative absolute error                 98.3768 %
Root relative squared error            106.4454 %
Total Number of Instances             4897

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.222    0.204    0.331      0.222    0.266      0.020    0.526     0.325     >30
                 0.774    0.746    0.614      0.774    0.684      0.032    0.542     0.642     NO
                 0.032    0.027    0.096      0.032    0.048      0.008    0.525     0.087     <30
Weighted Avg.    0.540    0.517    0.483      0.540    0.501      0.027    0.536     0.497

=== Confusion Matrix ===

    a    b    c   <-- classified as
  340 1155   39 |    a = >30
  586 2292   84 |    b = NO
  100  288   13 |    c = <30
```

Figure 7. J48 Predictive Model Results of Analysis I:



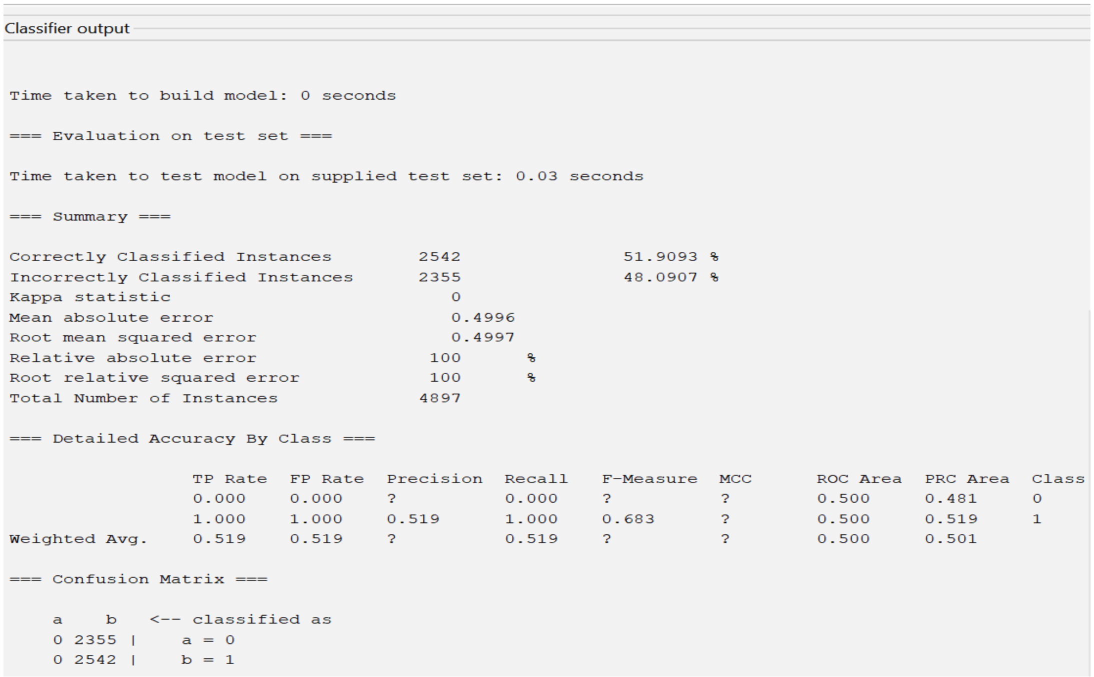
```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances        2962               60.486  %
Incorrectly Classified Instances      1935               39.514  %
Kappa statistic                          0
Mean absolute error                      0.3549
Root mean squared error                  0.4201
Relative absolute error                 99.9961 %
Root relative squared error             99.9999 %
Total Number of Instances             4897

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.313     >30
                 1.000    1.000    0.605      1.000    0.754      ?        0.500     0.605     NO
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.082     <30
Weighted Avg.    0.605    0.605    ?          0.605    ?          ?        0.500     0.471

=== Confusion Matrix ===

    a    b    c   <-- classified as
    0 1534    0 |    a = >30
    0 2962    0 |    b = NO
    0  401    0 |    c = <30
```

Figure 8. Logistic Regression Predictive Model Results of Analysis I:



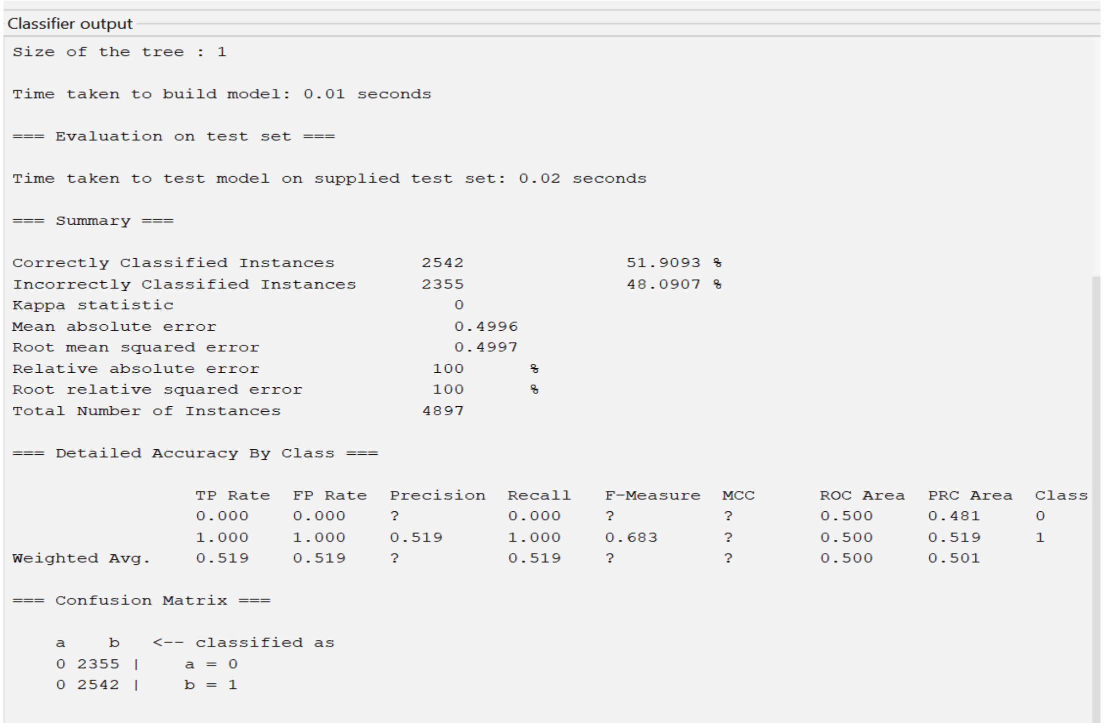
```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        2930               59.8326 %
Incorrectly Classified Instances      1967               40.1674 %
Kappa statistic                          0.0188
Mean absolute error                      0.347
Root mean squared error                  0.4163
Relative absolute error                 97.7591 %
Root relative squared error             99.102  %
Total Number of Instances             4897

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.056    0.045    0.363      0.056    0.097      0.024    0.576     0.364     >30
                 0.960    0.938    0.610      0.960    0.746      0.050    0.602     0.685     NO
                 0.000    0.000    0.000      0.000    0.000      -0.004   0.600     0.114     <30
Weighted Avg.    0.598    0.581    0.483      0.598    0.482      0.038    0.594     0.538

=== Confusion Matrix ===

    a    b    c   <-- classified as
   86 1448    0 |    a = >30
  117 2844    1 |    b = NO
   34  367    0 |    c = <30
```

Figure 9. Naive Bayes Predictive Model Results of Analysis II:



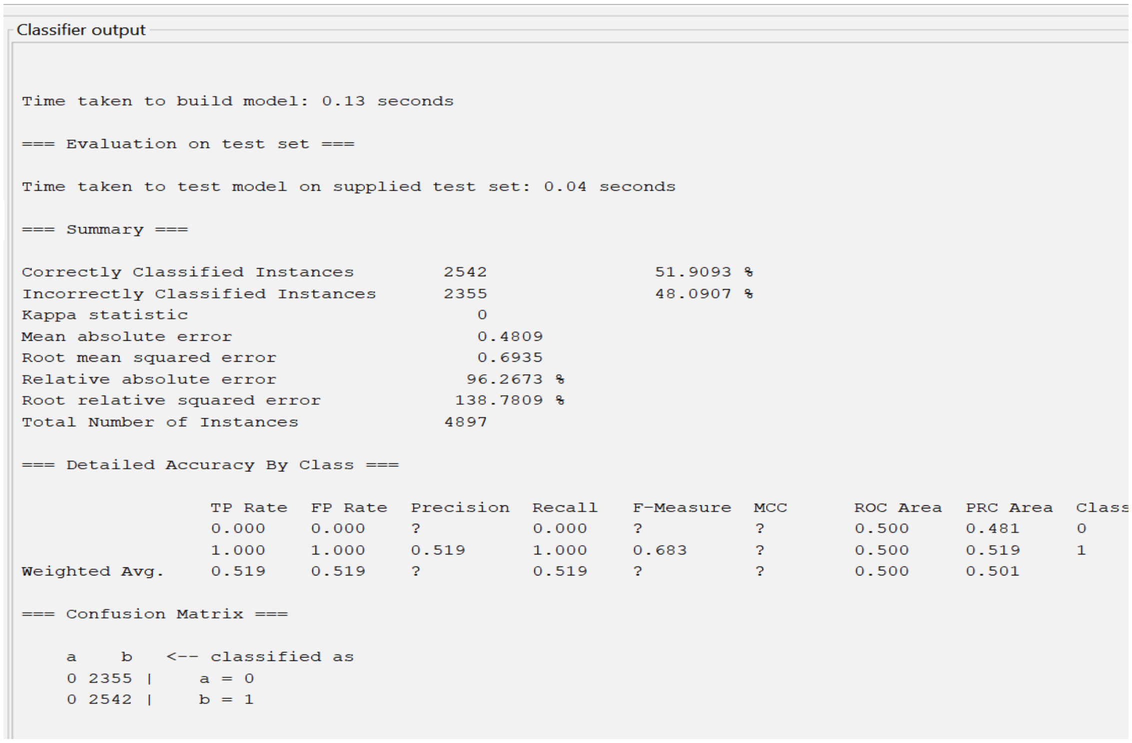
```
Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        2542               51.9093 %
Incorrectly Classified Instances      2355               48.0907 %
Kappa statistic                          0
Mean absolute error                      0.4996
Root mean squared error                  0.4997
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             4897

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.481     0
                 1.000    1.000    0.519      1.000    0.683      ?        0.500     0.519     1
Weighted Avg.    0.519    0.519    ?          0.519    ?          ?        0.500     0.501

=== Confusion Matrix ===

    a    b   <-- classified as
    0 2355 |    a = 0
    0 2542 |    b = 1
```

Figure 10. Random Tree Predictive Model Results of Analysis II:



```
Size of the tree : 1

Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances        2542               51.9093 %
Incorrectly Classified Instances      2355               48.0907 %
Kappa statistic                          0
Mean absolute error                      0.4996
Root mean squared error                  0.4997
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             4897

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.481     0
                 1.000    1.000    0.519      1.000    0.683      ?        0.500     0.519     1
Weighted Avg.    0.519    0.519    ?          0.519    ?          ?        0.500     0.501

=== Confusion Matrix ===

    a    b   <-- classified as
    0 2355 |    a = 0
    0 2542 |    b = 1
```

Figure 11. Sequential Minimum Optimization Predictive Model Results of Analysis II:



```
Time taken to build model: 0.13 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        2542               51.9093 %
Incorrectly Classified Instances      2355               48.0907 %
Kappa statistic                          0
Mean absolute error                      0.4809
Root mean squared error                  0.6935
Relative absolute error                 96.2673 %
Root relative squared error            138.7809 %
Total Number of Instances             4897

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.481     0
                 1.000    1.000    0.519      1.000    0.683      ?        0.500     0.519     1
Weighted Avg.    0.519    0.519    ?          0.519    ?          ?        0.500     0.501

=== Confusion Matrix ===

    a    b   <-- classified as
    0 2355 |    a = 0
    0 2542 |    b = 1
```