MID-TERM ASSIGNMENT

Course: DATA WAREHOUSING AND DATA MINING[A]

Name: Sami, Sadman Kabir

ID: 18-37378-1

SUBMITTED TO-
Dr. Md. Mahbub Chowdhury Mishu
Assistant Professor and Head in Charge (UG)
Dept. of Computer Science
Faculty of Science and Technology
American International University-Bangladesh

Title Of Dataset:

Blood Transfusion Service Center Data Set.

Source:

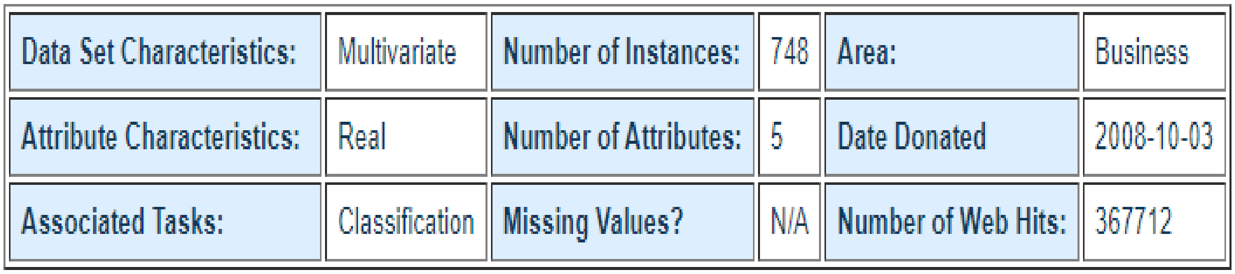
I collected this Dataset from UCI.

Give the names, email locations, organizations, and other contact data of the benefactors and makers of the information set. The unique dataset variant was gathered by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan) and made accessible by them on FigShare under the Attribution 4.0 International (CC BY 4.0: opportunity to share and adjust the material) copyright in July 2017. The flow rendition of the dataset was explained by Davide Chicco (Krembil Research Institute, Toronto, Canada) and given to the University of California Irvine Machine Learning Repository under a similar Attribution 4.0 International (CC BY 4.0) copyright in January 2020. Davide Chicco can be reached at <davidechicco '@' davidechicco.it>

Link: https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

Abstract:

In my Dataset, there is a huge amount of 748 donor data. Each case carries 5 attributes. There are various techniques for prediction but as the Decision tree algorithm gives me the most accuracy (specifically on this dataset) on this dataset so we are going to use the Decision tree algorithm for prediction in our work.



Data Set Characteristics:	Multivariate	Number of Instances:	748	Area:	Business
Attribute Characteristics:	Real	Number of Attributes:	5	Date Donated	2008-10-03
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	367712

Dataset Information:

To show the RFMTC promoting model (a changed variant of RFM), this review embraced the giver data set of Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The middle passes their blood bonding administration transport to one college in Hsin-Chu City to accumulate blood gave about like clockwork. To construct an FRMTC model, we chose 748 contributors at irregular from the benefactor information base. This 748 benefactor information, everyone included R (Recency - months since the last gift), F (Frequency - absolute number of gift), M (Monetary - complete blood gave in c.c.), T (Time - months since the first gift), and a double factor addressing whether he/she gave blood in March 2007 (1 represent giving blood; 0 represents not giving blood).

Attribute Information:

Given are the variable name, variable sort, the estimation unit, and a short depiction. The "Blood Transfusion Service Center" is an arrangement issue. The request for this posting compares to the request for numerals along the lines of the information base. R- (Recency - months since the last gift),

F - (Frequency - all out the number of gifts),

M- (Monetary - complete blood gave in c.c.),

T (Time - months since the first gift), and a paired variable addressing whether he/she gave blood in March 2007 (1 represents giving blood; 0 represents not giving blood).

We chose 500 pieces of information indiscriminately as the preparation set and the rest 248 as the testing set. Variable Data Type Measurement Description min max mean sexually transmitted disease Recency quantitative Months Input 0.03 74.4 9.74 8.07 Recurrence quantitative Times Input 1 50 5.51 5.84

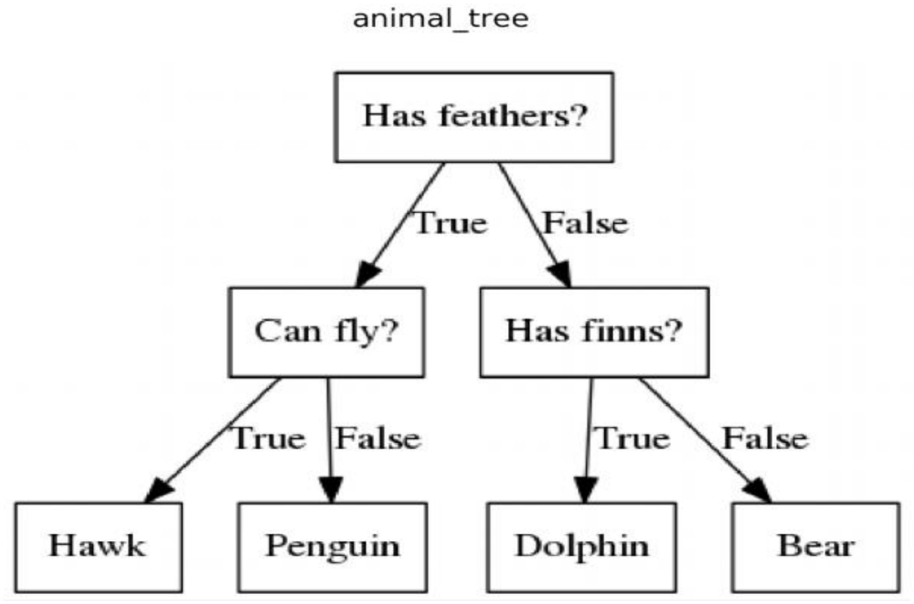
Financial quantitative c.c. blood Input 250 12500 1378.68 1459.83 Time quantitative Months Input 2.27 98.3 34.42 24.32 Regardless of whether he/she gave blood in March 2007 parallel 1=yes 0=no Output 0 1 (24%) 0 (76%).

Why I Selected This Dataset:

It was discussed before that my dataset is about a blood transfusion service center. There are lots of datasets on the internet my I selected this as my ideal dataset for so many particular reasons. Today, most clinical consideration relies upon a consistent stockpile of blood from givers, as one out of seven individuals entering the clinic needs blood. For Moffitt, as disease care increments, so does the interest for blood and platelet gifts. A sufficient measure of blood is required in all medical services offices to meet the critical requirement for patients confronting injury and other lifesaving methods, for example, blood bondings – which saves a great many lives every year. Giving blood is a straightforward method that should be possible in not over 60 minutes.

Used Technique For The Classification(Decision Tree-j48) :

We learned KNN, Naeve Bayes, Decision tree. I chose the Decision tree calculation methods. Let's talk a little about what a decision tree actually is. A choice tree is a directed learning procedure that has a pre-characterized target variable and is regularly utilized in characterization issues. This tree can be applied to either clear cut or constant info and yield factors. The preparation cycle takes after a stream diagram, with each inner (non-leaf) hub a trial of quality, each branch is the result of that test, and each leaf (terminal) hub contains a class name. The most noteworthy center in the tree is known as the root center point. In the choice interaction, the example (populace) is parted into at least two sub-populaces sets of maximal, which is chosen by the main splitter or differentiator in the info factors. A definitive objective is to make a prescient model that can take perceptions about an example (the branches) and make exact decisions about the example's objective worth (the leaves).



Example: Animal Tree

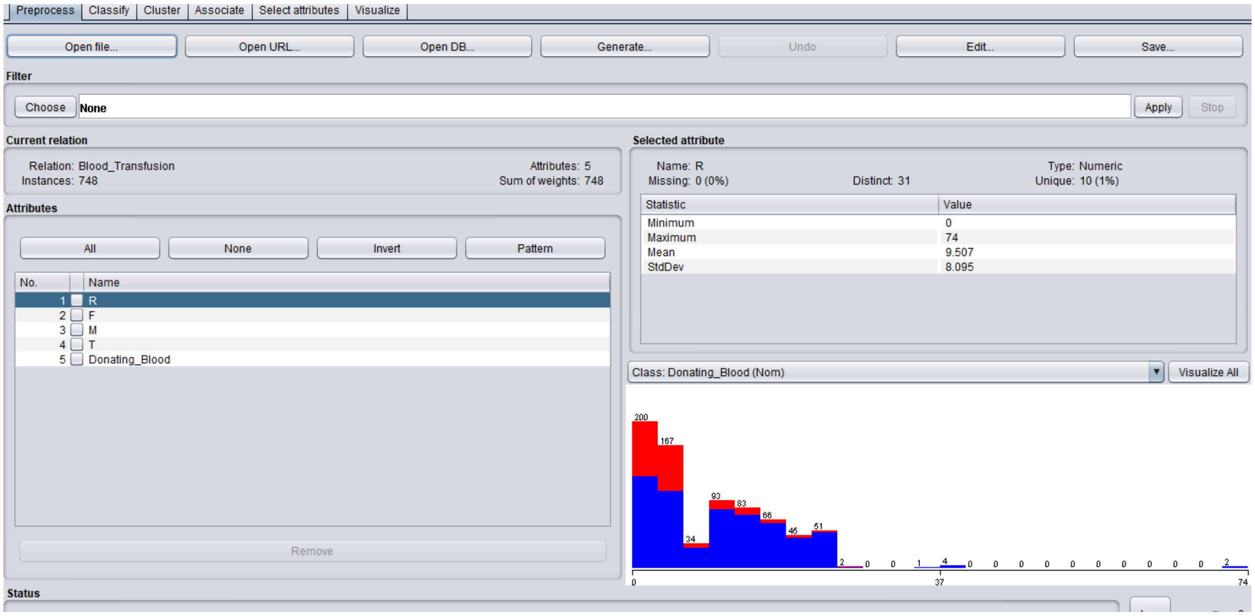
Why I Selected Decision Tree:

Right off the bat we utilized the disarray grid for discovering the exactness in our dataset. We tried every technique (KNN, Naeve Byes, Decision tree) on our dataset. After using KNN we found 71% accuracy. By using Naeve Bayes we got 75.4% accuracy but after using the Decision tree we have got a whopping 77.8% accuracy. There are also many reasons for selecting this technique.

1. Contrasted with different calculations choice trees require less exertion for information arrangement during pre-handling.
2. It doesn't need standardization of information.

3. It doesn't need scaling of information too.
4. Missing qualities in the information likewise don't influence the method involved with building a choice tree to any significant degree.
5. A Decision tree model is extremely natural and simple to disclose to specialized groups just as partners.

Simulation In Weka:



Filter: None

Current relation

Relation: Blood_Transfusion
Instances: 748
Attributes: 5
Sum of weights: 748

Attributes

No.		Name
1	☐	R
2	☐	F
3	☐	M
4	☐	T
5	☐	Donating_Blood

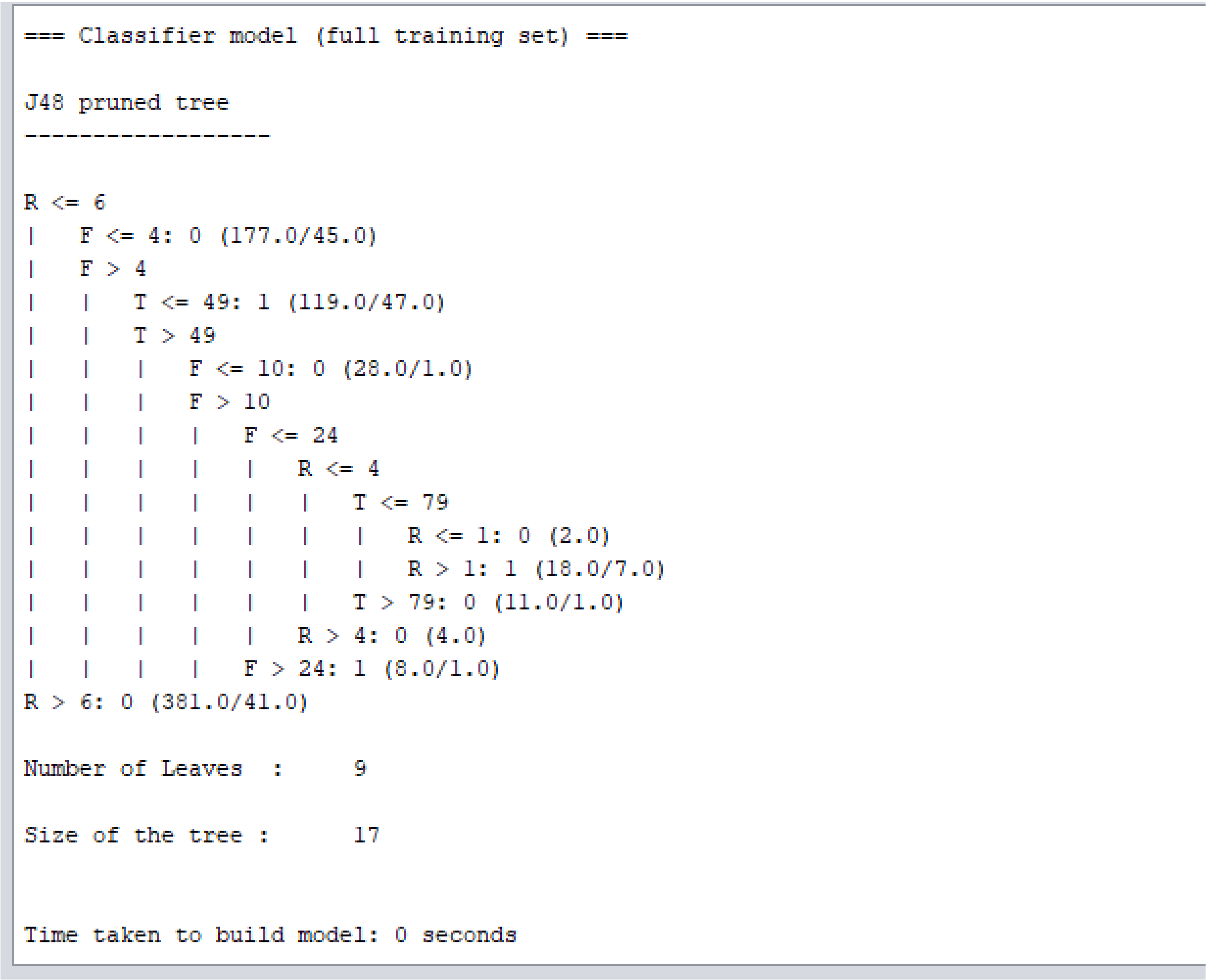
Selected attribute

Name: R
Missing: 0 (0%)
Distinct: 31
Type: Numeric
Unique: 10 (1%)

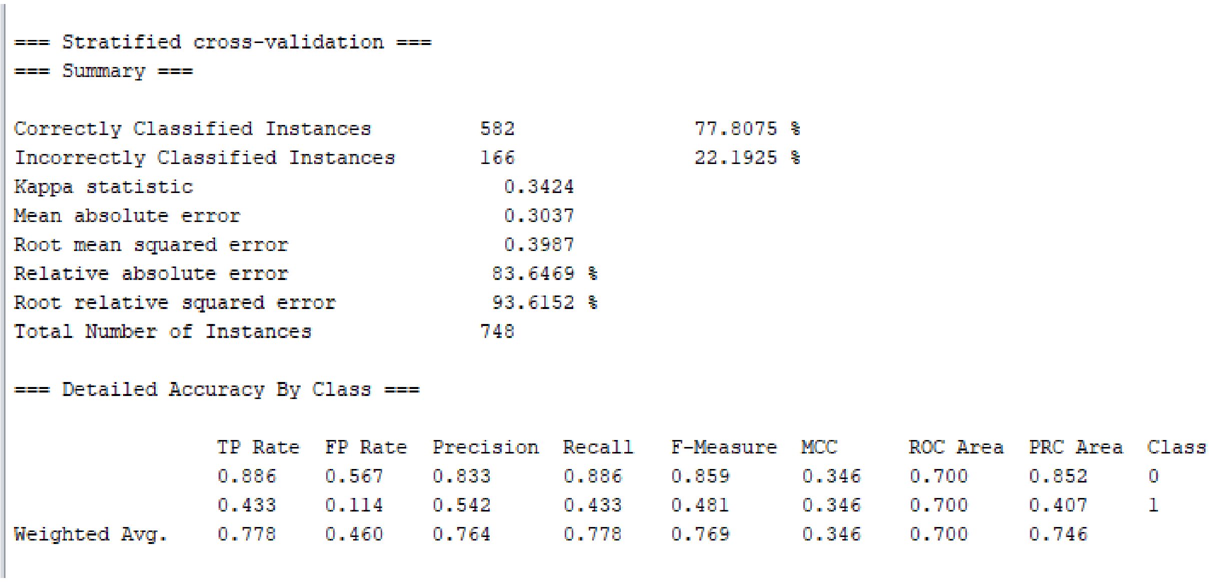
Statistic	Value
Minimum	0
Maximum	74
Mean	9.507
StdDev	8.095

Class: Donating_Blood (Nom)

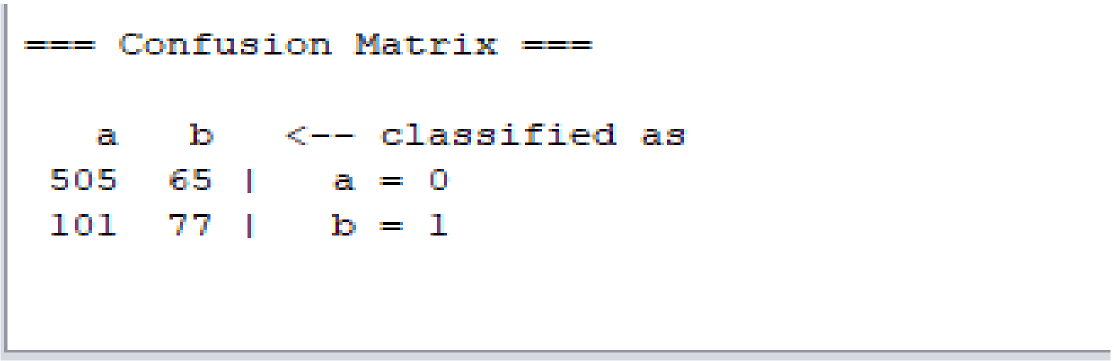
```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Blood_Transfusion
Instances:    748
Attributes:   5
              R
              F
              M
              T
              Donating_Blood
Test mode:    10-fold cross-validation
```



```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

R <= 6
|   F <= 4: 0 (177.0/45.0)
|   F > 4
|   |   T <= 49: 1 (119.0/47.0)
|   |   T > 49
|   |   |   F <= 10: 0 (28.0/1.0)
|   |   |   F > 10
|   |   |   |   F <= 24
|   |   |   |   |   R <= 4
|   |   |   |   |   |   T <= 79
|   |   |   |   |   |   |   R <= 1: 0 (2.0)
|   |   |   |   |   |   |   R > 1: 1 (18.0/7.0)
|   |   |   |   |   |   T > 79: 0 (11.0/1.0)
|   |   |   |   |   R > 4: 0 (4.0)
|   |   |   |   F > 24: 1 (8.0/1.0)
R > 6: 0 (381.0/41.0)

Number of Leaves  :     9

Size of the tree :      17


Time taken to build model: 0 seconds
```



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

Correctly Classified Instances         582               77.8075 %
Incorrectly Classified Instances       166               22.1925 %
Kappa statistic                          0.3424
Mean absolute error                      0.3037
Root mean squared error                  0.3987
Relative absolute error                 83.6469 %
Root relative squared error             93.6152 %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.886    0.567    0.833      0.886    0.859      0.346    0.700     0.852     0
                 0.433    0.114    0.542      0.433    0.481      0.346    0.700     0.407     1
Weighted Avg.    0.778    0.460    0.764      0.778    0.769      0.346    0.700     0.746
```



```
=== Confusion Matrix ===

   a   b   <-- classified as
 505  65 |   a = 0
 101  77 |   b = 1
```

	Predicted No a	Predicted Yes b
Actual No	TN=505	FP=65
Actual Yes	FN=101	TP=77

Accuracy = $\frac{\mathrm{TP+TN}}{\mathrm{Total}}$

$= \frac{77+505}{748}$

= 0.778

= 77.8%

Sensitivity = $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$

$= \frac{77}{77+101}$

= 0.432

= 43.2%

Specificity = $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$

$= \frac{505}{505+65}$

= 0.885

=88.5%

Decision Tree(Types):

Root Node: It is the hub that begins the chart. A typical choice tree assesses the variable that best divides the information.

Intermediate nodes: These are hubs where factors are assessed however these are not the last hubs where forecasts are made.

Leaf Nodes:

These are the last hubs of the tree, where the expectations of a classification or a mathematical worth are made.

Conclusion:

As we discussed before in KNN we got 71% accuracy. We got 75.40% accuracy with naeve bayers. Above all, we got a big 77.8% accuracy using the decision tree technique on our dataset where sensitivity is 43.2% and specificity is 88.5%. We clearly can see that we are getting the best output using the decision tree technique. Decision trees are valuable administration instruments that assist with formalizing your point of view and give a graphical portrayal of how various components may impact your arrangements. They unmistakably spread out potential ways from the choice to every single imaginable outcome, with the goal that the expense and advantage of every way can be thought of. By judging all these reasons, facilities we selected this decision tree technique to get a good result.