KULLIYYAH OF INFORMATION & COMMUNICATION TECHNOLOGY
SEMESTER 2, 2021/2022

INFO 4313 DATA MINING

SECTION 01

"GROUP ASSIGNMENT"

PREPARED BY:

NAME	MATRIC NO.
MD ABDUR RAHMAN	1639233
MAMMI FARJANA	1912190
MOHAMMAD RAIHANUL ISLAM	1825891
HASAN MD TANVIR	1716763
MD SAJIBUR RAHMAN	1715205

LECTURER ATIKAH BALQIS BINTI BASRI

DUE 03 JUNE 2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

The process of extracting and identifying hidden patterns and relationships in large volumes of data is known as data mining. It's a powerful data analysis process that combines machine learning and artificial intelligence to extract the relevant data that helps analysts better understand their objectives and make better decisions for increasing profits, lowering expenses, boosting interconnections across diverse sets of data, and much more.

One of the strategies used in data mining is classification, which assists in the organization of data sets. The purpose of classification techniques is to identify patterns, movements, and groups in massive amounts of data and transform that information into more accurate knowledge for better decision-making. We used the Blood Transfusion Service Centre dataset in Hsin-Chu City, Taiwan, to run multiple Machine Learning algorithms and afterwards compare the finest and worst performers in classification techniques.

2.0 DATA SOURCE

This information was gathered from the Blood Transfusion Service Centre's donor database in Hsin-Chu City, Taiwan. The clinic sends its transfusion services bus to a university in Hsin-Cho every three months to collect donated blood. A total of 748 donors were chosen at random for the study.

3.0 CONCEPTS OF CLASSIFICATION

The process of classifying items based on their attributes is known as classification. The purpose of classification is to predict a specific outcome based on the data or input given. It consists of two basic stages: the learning phase, in which the classification algorithm is learned, and the classification phase, in which the algorithm labels fresh data.

4.0 USED OF METHOD

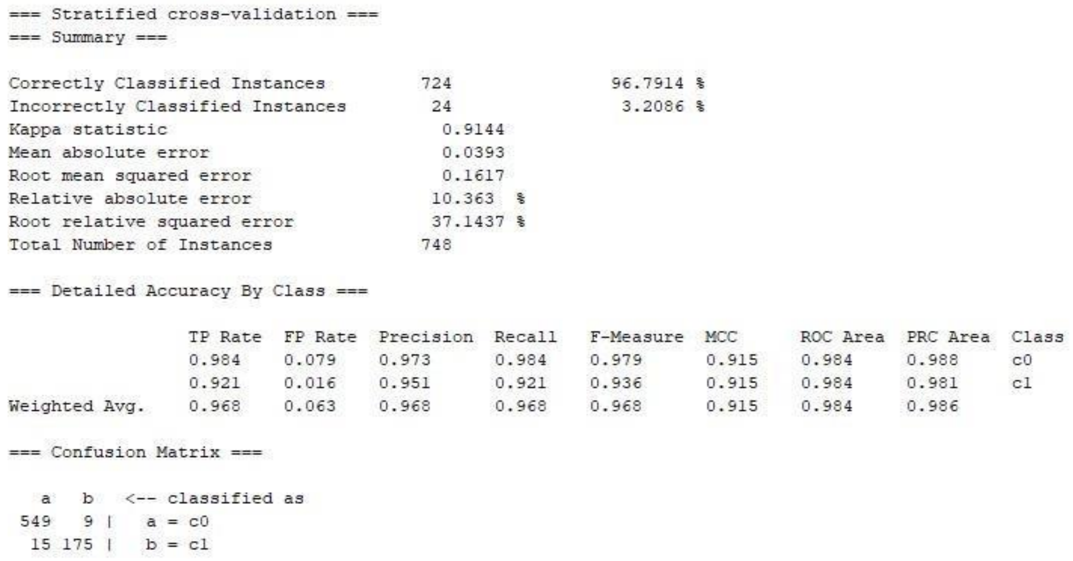
Weka has lots of ML algorithms. For our project we used data from Blood Transfusion Service Centre and our group is using four types of ML algorithm **(Neural Network, Logistic regression, Tree.J48 and Naive Bayes)** to classify the data by using 10-fold cross validation.

4.1 NEURAL NETWORK / MULTILAYERPERCEPTRON

According to the theory, a neural network is an AI system approach that indicates computers analyze data in a way inspired by the human brain. However, it is a supervised learning machine learning approach that employs linked nodes or neurons in a layered structure that resembles the human brain. Furthermore, this method provides an adaptive system that allows computers to learn from their mistakes and continue improving. As a result, artificial neural networks aim to solve complex tasks with increased precision, such as summarizing papers or identifying faces.

As far as we know, a neural network is a scientific principle predicated on the idea of biological neural systems. However, it does imitate the human comprehension process. It also consists of a collection of artificial neurons that analyze data that is sent into it while also establishing a link between the inputted data and the network's "memory," which accumulates during network training and, in certain networks, while the program is run.
(Sinkov et al. 2016).

4.1.1 RESULT:



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

Correctly Classified Instances         724               96.7914 %
Incorrectly Classified Instances        24                3.2086 %
Kappa statistic                          0.9144
Mean absolute error                      0.0393
Root mean squared error                  0.1617
Relative absolute error                 10.363  %
Root relative squared error             37.1437 %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.984    0.079    0.973      0.984    0.979      0.915    0.984     0.988     c0
                 0.921    0.016    0.951      0.921    0.936      0.915    0.984     0.981     c1
Weighted Avg.    0.968    0.063    0.968      0.968    0.968      0.915    0.984     0.986

=== Confusion Matrix ===

   a   b   <-- classified as
 549   9 |   a = c0
  15 175 |   b = c1
```

Figure 1: Weka Explorer- classification using Multilayer Perceptron

After using the Multilayer Perception or neural network algorithm we can see **(figure 1)** that correctly classified instances are 724 data which is 96.7914% and incorrectly classified instances are 24 data which is 3.2086%. From the confusion matrix we can classified like this, Let, A= Yes and B= No instead of a=0 and b=1

As we know, we can easily measure the performance of a classification problem through confusion matrix. A confusion matrix contains a table with two dimensions which are "Actual" and "Predicted". Both dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", "False Negatives (FN)".

Confusion matrix from (figure 1),

```
   a   b   <-- classified as
 549   9 |   a = c0
  15 175 |   b = c1
```

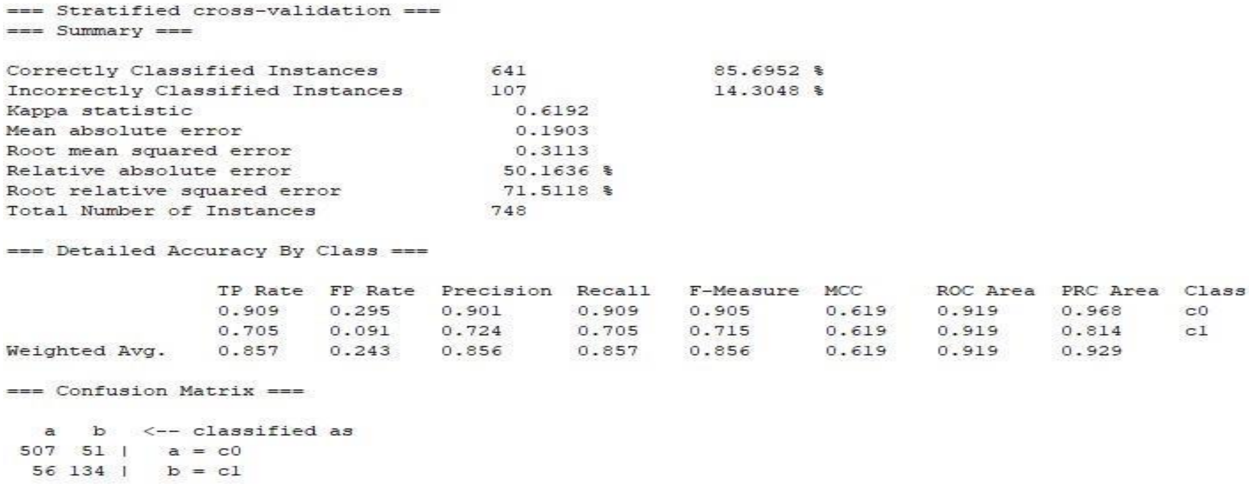
True Positive	False Positive	True Negative	False Negative
549	9	175	15

Table 1: Confusion matrix of Neural Network

4.2 LOGISTIC REGRESSION

Logistic regression is a statistical analytic approach for predicting a binary result, such as yes or no. A logistic regression algorithm analyzes the connection between one or more existing independent variables to predict a dependent data variable. For example, a logistic regression could be used to predict whether a political candidate will win or lose and go or out an election, rainy or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives. In fact, logistic regression is one of the commonly used algorithms in machine learning as well as data mining for binary classification problems, which are problems with two class values, including predictions such as "this or that," "yes or no," and "A or B."

4.2.1 RESULT:



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

Correctly Classified Instances         641               85.6952 %
Incorrectly Classified Instances       107               14.3048 %
Kappa statistic                          0.6192
Mean absolute error                      0.1903
Root mean squared error                  0.3113
Relative absolute error                 50.1636 %
Root relative squared error             71.5118 %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.909    0.295    0.901      0.909    0.905      0.619    0.919     0.968     c0
                 0.705    0.091    0.724      0.705    0.715      0.619    0.919     0.814     c1
Weighted Avg.    0.857    0.243    0.856      0.857    0.856      0.619    0.919     0.929

=== Confusion Matrix ===

   a   b   <-- classified as
 507  51 |   a = c0
  56 134 |   b = c1
```

Figure 2: Weka Explorer- classification using function Logistic

After using the logistic regression algorithm, we can see **(figure 2)** that correctly classified instances are 85.6952% and incorrectly classified instances are 14.3048%. From the confusion matrix we can classified like this, Let, A= Yes and B= No instead of a=0 and b=1

As we know, we can easily measure the performance of a classification problem through confusion matrix. A confusion matrix contains a table with two dimensions which are "Actual" and "Predicted". Both dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", "False Negatives (FN)".

Confusion matrix from (figure 2),

```
   a   b   <-- classified as
 507 51 |   a = c0
  56 134 |   b = c1
```

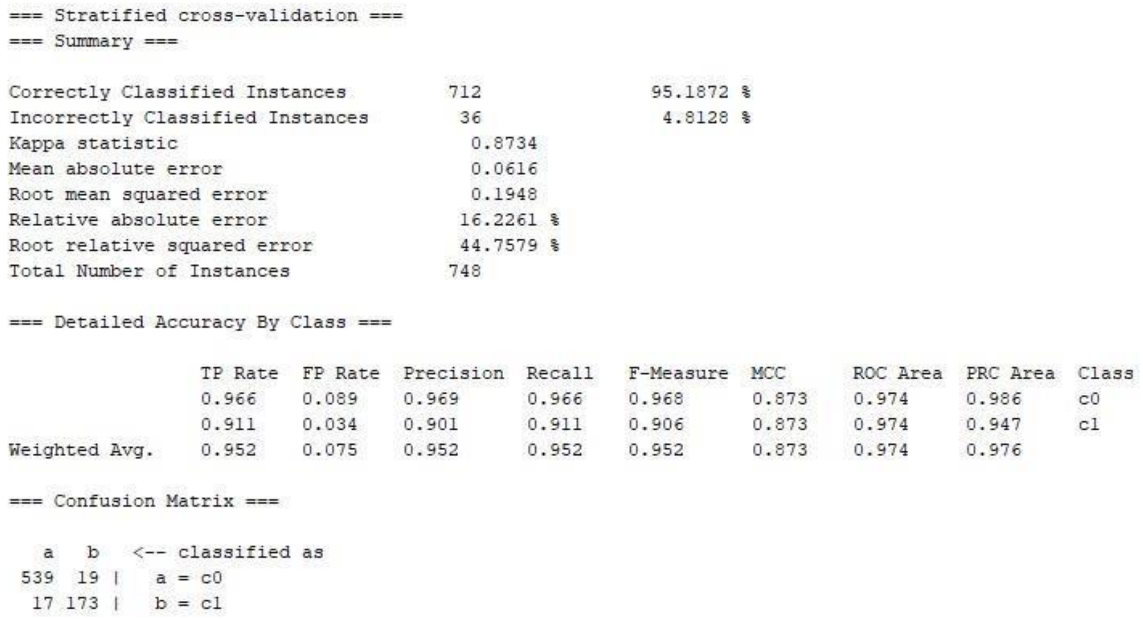
True Positive	False Positive	True Negative	False Negative
507	51	134	56

Table 2: Confusion matrix of Logistic Regression

4.3 TREES.J48

An algorithmic strategy that can slice the information in various ways based on different variables can be used to create decision trees. The most powerful algorithms in the domain of supervised algorithms are decision trees. Every feature of the data must be divided into small subsets to make a choice. J48 analyzes the unified data gain to see whether the findings are indeed the outcomes of splitting the data by attribute. To conclude, extreme standardized data obtained is used as a characteristic. The algorithm will generate the minor subsets. If a subset has a location with a comparable class in all cases, the split methods come to an end. J48 creates a decision node based on the class's expected predictions (Venkatesan, 2015).

4.3.1 RESULT



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

Correctly Classified Instances         712               95.1872 %
Incorrectly Classified Instances        36                4.8128 %
Kappa statistic                          0.8734
Mean absolute error                      0.0616
Root mean squared error                  0.1948
Relative absolute error                 16.2261 %
Root relative squared error             44.7579 %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.089    0.969      0.966    0.968      0.873    0.974     0.986     c0
                 0.911    0.034    0.901      0.911    0.906      0.873    0.974     0.947     c1
Weighted Avg.    0.952    0.075    0.952      0.952    0.952      0.873    0.974     0.976

=== Confusion Matrix ===

   a   b   <-- classified as
 539  19 |   a = c0
  17 173 |   b = c1
```

Figure 3: Weka Explorer- classification using Trees.J48

For J48, correctly classified instances are 712 data which is 95.1872% and incorrectly classified instances are 36 data which is 4.8128%. From the confusion matrix we can classified like this, Let, A= Yes and B= No instead of a=0 and b=1

Confusion matrix from (figure 3),

```
   a   b   <-- classified as
539 19 |   a = c0
 17 173 |   b = c1
```

True Positive	False Positive	True Negative	False Negative
539	19	173	17

Table 3: Confusion matrix of J48

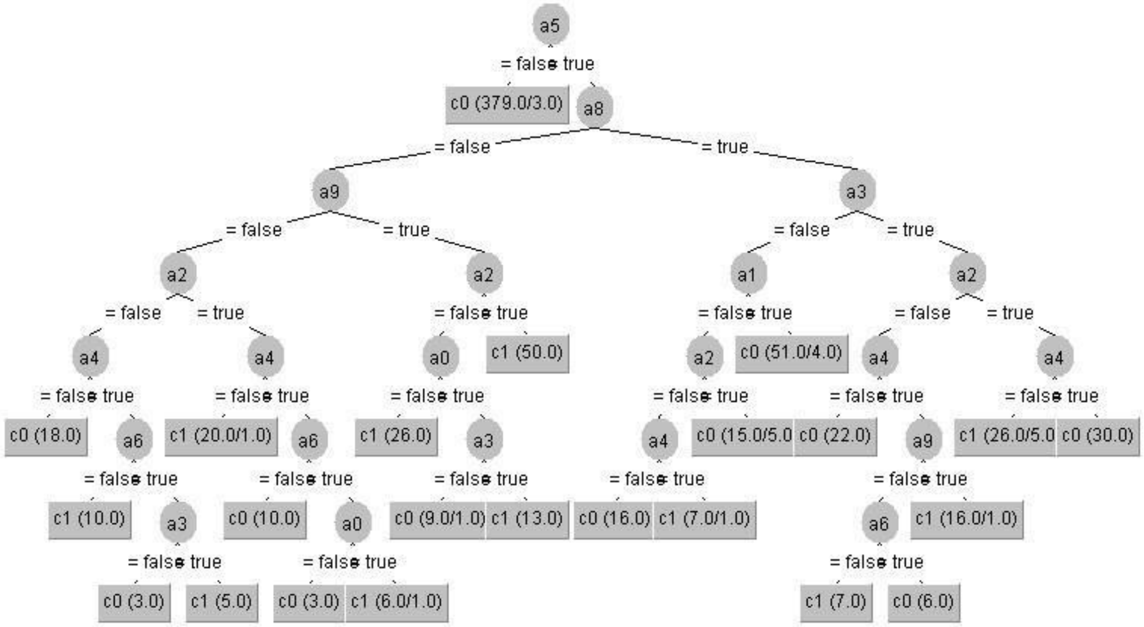
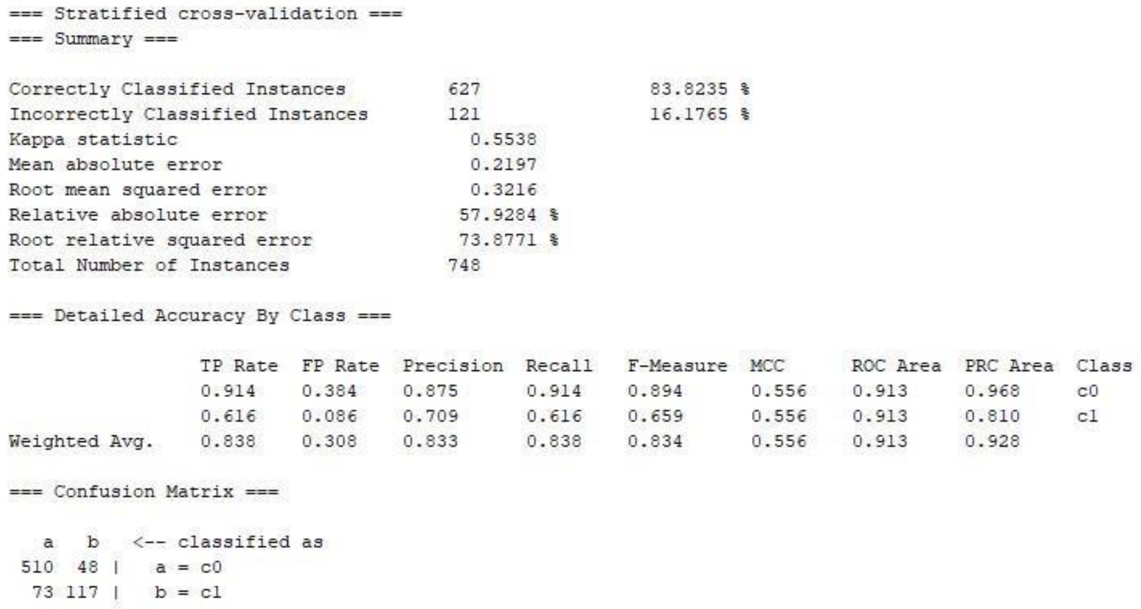


Figure 4: View of Trees.J48

4.4 NAIVE BAYES

In the modern world one of the most significant data mining disciplines is Nave Bayes in machine learning techniques. One of the best-supervised classification techniques in data Mining is Naive Bayes classification. Naive Bayes classification is good at predicting outcomes and often outperforms other classification techniques. (kalyan et al.2015)

4.4.1 RESULT



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

Correctly Classified Instances         627               83.8235 %
Incorrectly Classified Instances       121               16.1765 %
Kappa statistic                          0.5538
Mean absolute error                      0.2197
Root mean squared error                  0.3216
Relative absolute error                 57.9284 %
Root relative squared error             73.8771 %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.914    0.384    0.875      0.914    0.894      0.556    0.913     0.968     c0
                 0.616    0.086    0.709      0.616    0.659      0.556    0.913     0.810     c1
Weighted Avg.    0.838    0.308    0.833      0.838    0.834      0.556    0.913     0.928

=== Confusion Matrix ===

   a   b   <-- classified as
 510  48 |   a = c0
  73 117 |   b = c1
```

Figure 5: Weka Explorer- classification using Naive Bayes

For Naive Bayes, correctly classified instances are 627 data which is 83.8235% and incorrectly classified instances are 121 data which is 16.1765%. From the confusion matrix we can classified like this, Let, A= Yes and B= No instead of a=0 and b=1

Confusion matrix from (figure 5),

a b <-- classified as

510 48 **|** a = c0

73 117 **|** b = c1

True Positive	False Positive	True Negative	False Negative
510	48	117	73

Table 4: Confusion matrix of naive Bayes

5.0 DISCUSSION

Our group has used four algorithms to classification. So, we have chosen confusion matrix to identify the best and the worst classification. From the confusion matrix we are going to use accuracy formula to prove the best and worst classification.

- ***Multilayer Perceptron***

True Positive	False Positive	True Negative	False Negative
549	9	175	15

TP = 549

FP = 9

TN = 175

FN = 15

As we know,

Accuracy **=** TP + TN / TP + FP + FN + TN

= 549 + 175 / 549 + 9 + 15 + 175

= 724 / 748

= 0.97

- ***Function. Logistic***

True Positive	False Positive	True Negative	False Negative
507	51	134	56

TP = 507 FP

= 51 TN =

134 FN = 56

As we know,

Accuracy **=** TP + TN **/** TP + FP + FN + TN

= 507 + 134 / 507 + 51 + 56 + 134

= 641 / 748

= 0.86

- ***Trees.J48***

True Positive	False Positive	True Negative	False Negative
539	19	173	17

TP = 539 FP

= 19 TN =

173 FN = 17

As we know,

Accuracy **=** TP + TN **/** TP + FP + FN + TN

= 539 + 173 / 539 + 19 + 17 + 173

= 712 / 748

= 0.95

- ***Naive Bayes***

True Positive	False Positive	True Negative	False Negative
510	48	117	73

TP = 510 FP = 48 TN = 117 FN = 73

As we know,

Accuracy **=** TP + TN **/** TP + FP + FN + TN

= 510 + 117 / 510+ 48 + 73 + 117

= 627 / 748

= 0.84

After using the accuracy formula our group has found the accuracy value as follows:

Multilayer Perception is = **0.97 or 97%**

Function. Regression is = 0.86 or 86%

Tree.J48 is = 0.95 or 95%

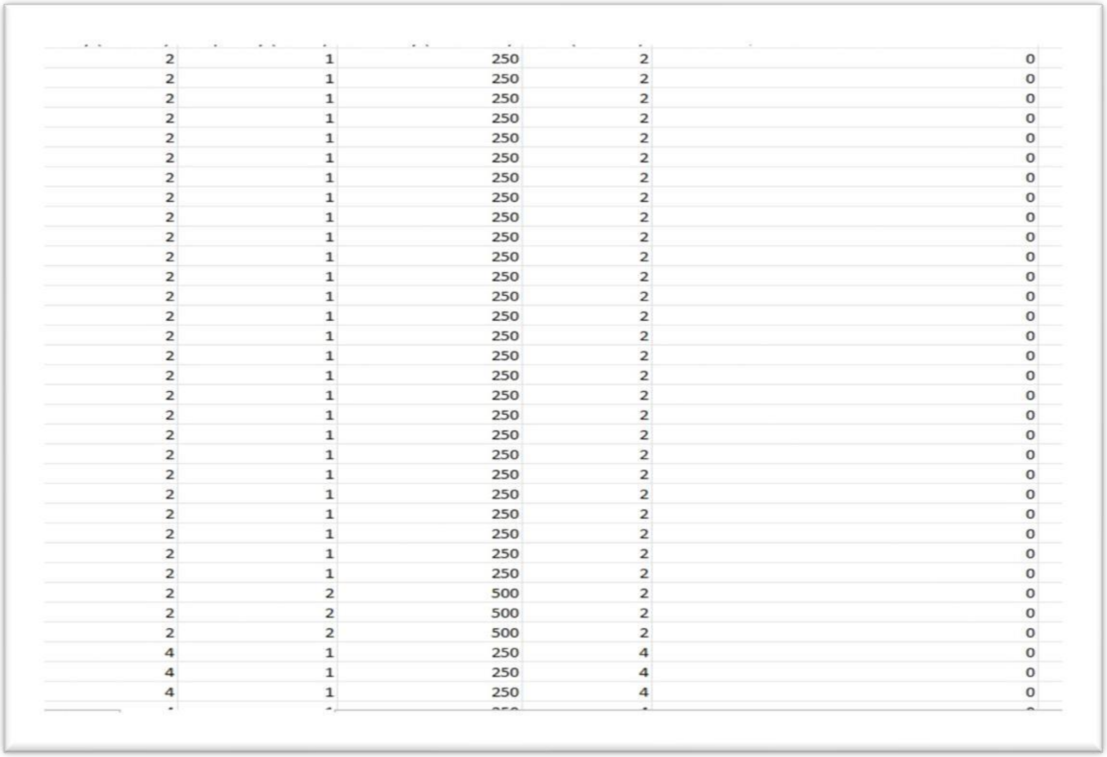
Naive Bayes is = **0.84 or 84%**

According to the calculation above, we can decide that the Multilayer Perception method is the best (where the accuracy value is **97%)** and Naive Bayes method is the worst (where the accuracy value is **84%)** classification for the blood transfusion dataset.

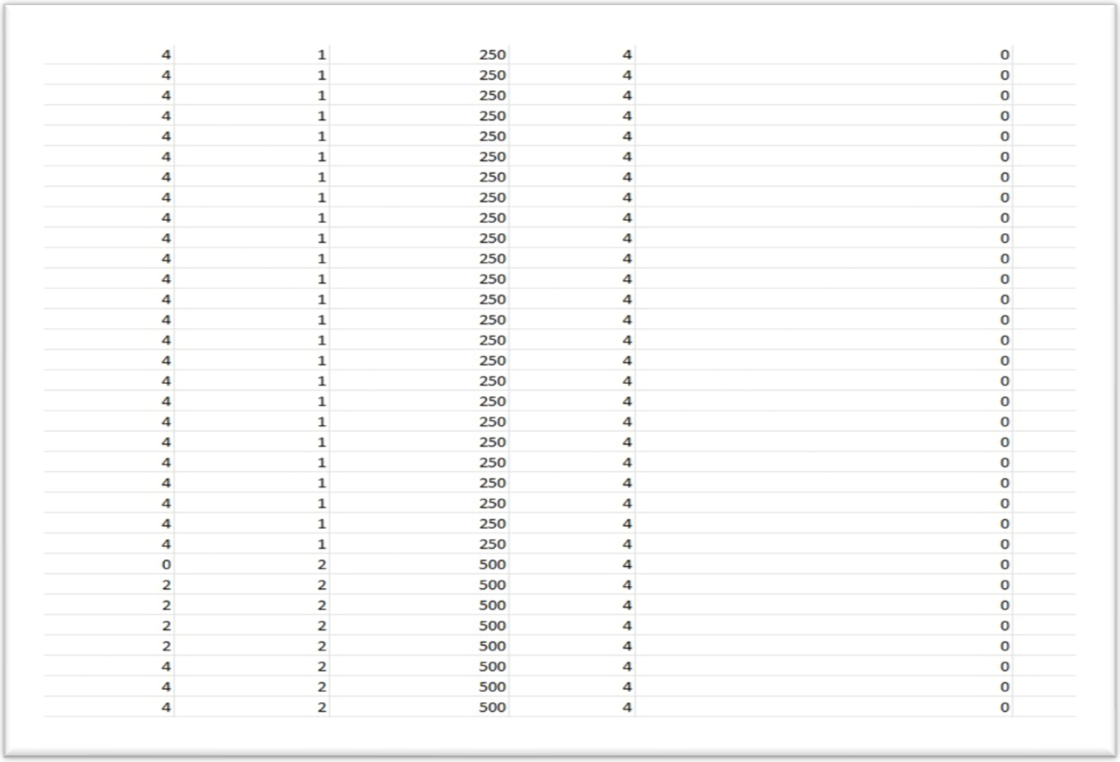
6.0 CONCLUSION

To conclude, data mining is a technique for uncovering hidden patterns in massive volumes of data. This hidden data can be utilized to forecast future behavior and aid in improved decision-making. Upon that donor dataset Blood Transfusion Service Center, we applied multiple Machine Learning algorithms in this study. The goal of experimenting with different algorithms is to see which one has the best level of accuracy while using WEKA. The algorithms that have been used are Multilayer Perceptron, Logistic regression, Trees.J48 and Naive Bayes. After calculating the accuracy of each algorithm, we concluded that Multilayer Perception has a better performance compared to the rest of the algorithm, while Naive Bayes method is the worst. However, different dataset gives different results, so Multilayer Perceptron performs better in our research, but another dataset may have different results.

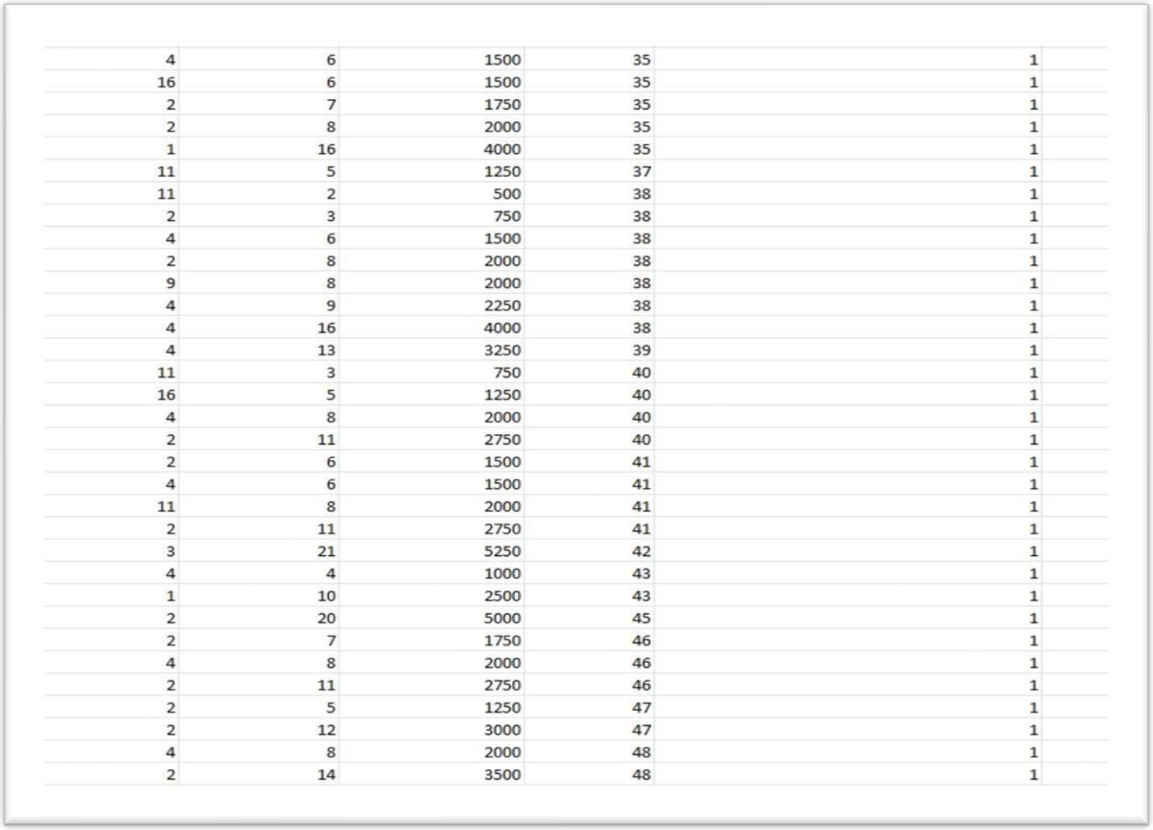
7.0 APPENDIX



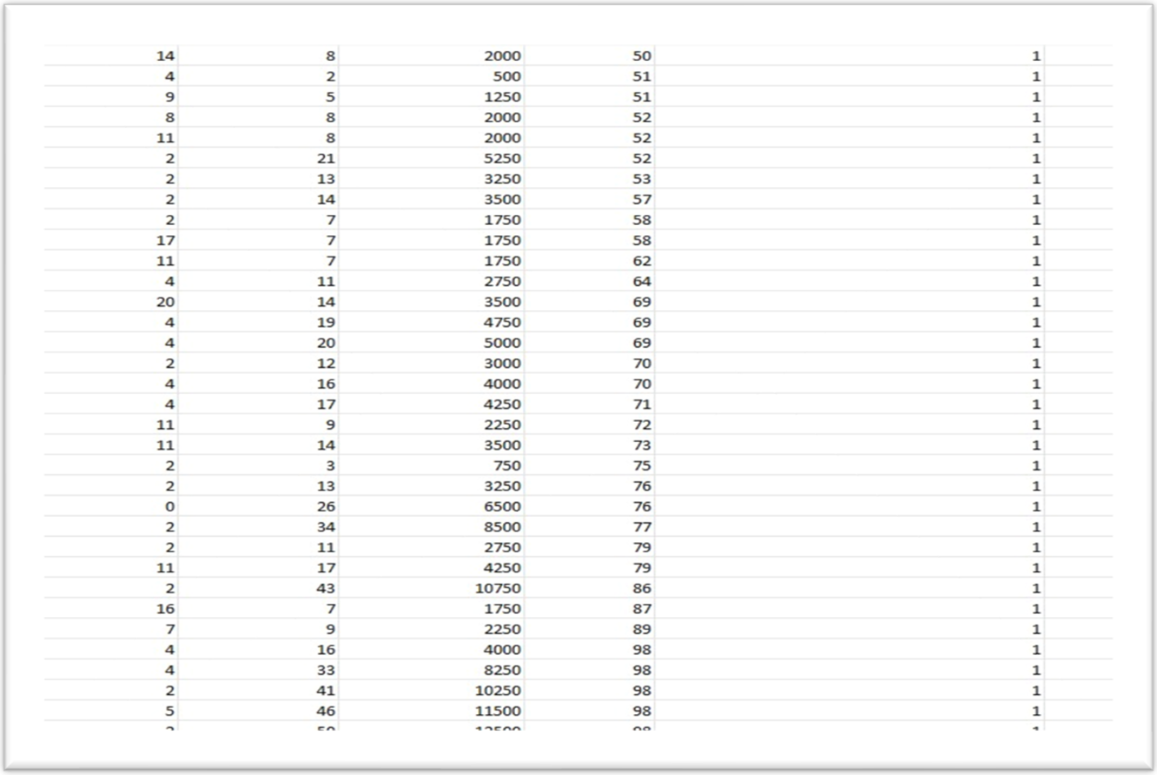
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	1	250	2	0
2	2	500	2	0
2	2	500	2	0
2	2	500	2	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0



4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
4	1	250	4	0
0	2	500	4	0
2	2	500	4	0
2	2	500	4	0
2	2	500	4	0
2	2	500	4	0
4	2	500	4	0
4	2	500	4	0
4	2	500	4	0



4	6	1500	35	1
16	6	1500	35	1
2	7	1750	35	1
2	8	2000	35	1
1	16	4000	35	1
11	5	1250	37	1
11	2	500	38	1
2	3	750	38	1
4	6	1500	38	1
2	8	2000	38	1
9	8	2000	38	1
4	9	2250	38	1
4	16	4000	38	1
4	13	3250	39	1
11	3	750	40	1
16	5	1250	40	1
4	8	2000	40	1
2	11	2750	40	1
2	6	1500	41	1
4	6	1500	41	1
11	8	2000	41	1
2	11	2750	41	1
3	21	5250	42	1
4	4	1000	43	1
1	10	2500	43	1
2	20	5000	45	1
2	7	1750	46	1
4	8	2000	46	1
2	11	2750	46	1
2	5	1250	47	1
2	12	3000	47	1
4	8	2000	48	1
2	14	3500	48	1



14	8	2000	50	1
4	2	500	51	1
9	5	1250	51	1
8	8	2000	52	1
11	8	2000	52	1
2	21	5250	52	1
2	13	3250	53	1
2	14	3500	57	1
2	7	1750	58	1
17	7	1750	58	1
11	7	1750	62	1
4	11	2750	64	1
20	14	3500	69	1
4	19	4750	69	1
4	20	5000	69	1
2	12	3000	70	1
4	16	4000	70	1
4	17	4250	71	1
11	9	2250	72	1
11	14	3500	73	1
2	3	750	75	1
2	13	3250	76	1
0	26	6500	76	1
2	34	8500	77	1
2	11	2750	79	1
11	17	4250	79	1
2	43	10750	86	1
16	7	1750	87	1
7	9	2250	89	1
4	16	4000	98	1
4	33	8250	98	1
2	41	10250	98	1
5	46	11500	98	1
2	50	12500	98	1

9.0 REFERENCES

Download Limit Exceeded. (n.d.). Citeseerx.ist.psu.edu. Retrieved June 1, 2022, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.640.136&rep=rep1&type=pdf

Netti, K., & Radhika, Y. (2015). A novel method for minimizing loss of accuracy in Naive Bayes classifier. *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).* https://doi.org/10.1109/iccic.2015.7435801

Sinkov, A., Asyaev, G., Mursalimov, A., & Nikolskaya, K. (2016). Neural networks in data mining. *2016 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*. https://doi.org/10.1109/icieam.2016.7911596

UCI Machine Learning Repository: Blood Transfusion Service Center Data Set. (2022). Uci.edu. https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

Venkatesan, E. V. (2015). Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Indian Journal of Science and Technology.