
Group Assignment

---

**TDM 3301**
(Data Mining)

# Classification of the KDD Cup 1999 Dataset

Prepared by


| | | | | |
|---|---|---|---|---|
| Samuel Wong Ing Shiing | 1151103687 | TT02 | 1151103687@student.mmu.edu.my | 0168706324 |
| Wayile Jialade | 1151102347 | TT02 | 1151102347@student.mmu.edu.my | 0146865506 |

# TABLE OF CONTENTS

# 1.0 INTRODUCTION

The dataset used in this data mining assignment is a part of the KDD 1999 dataset. The KDD 1999 dataset is the dataset used for The Third International Knowledge Discovery and Data Mining Tools Competition that was held together with the KDD-99 The Fifth International COnference on Knowledge Discovery and Data Mining. The main objective of the competition was to build a network intrusion detector as a predictive model which is able to distinguish between "bad" connections and "good" regular connections. The database contains a set of data to be audited, which includes a wide variety of intrusions that were simulated in a military network environment.

The dataset contains 42 attributes, each represents a feature defined for the connection records that are contained within the dataset. The following is the list of attributes contained within the dataset.

**TABLE 1.0 Attribute Properties**

| Attribute No. | Attribute Name | Type of Attribute Value |
|---|---|---|
| 1 | Type of connection | Symbolic Value |
| 2 | Duration | Continuous Value |
| 3 | Protocol Type | Symbolic Value |
| 4 | Service | Symbolic Value |
| 5 | Flag | Symbolic Value |
| 6 | Src bytes | Continuous Value |
| 7 | Dst bytes | Continuous Value |
| 8 | Land | Continuous Value |
| 9 | Wrong fragment | Continuous Value |
| 10 | Urgent | Continuous Value |
| 11 | Hot | Continuous Value |
| 12 | Num failed logins | Continuous Value |
| 13 | Logged in | Symbolic Value |
| 14 | Num compromised | Continuous Value |

| | | |
|---|---|---|
| 15 | Root shell | Continuous Value |

| | | |
|---|---|---|
| 16 | Su attempted | Continuous Value |
| 17 | Num root | Continuous Value |
| 18 | Num file | Continuous Value |
| 19 | Num shells | Continuous Value |
| 20 | Num access files | Continuous Value |
| 21 | Num outbound cmds | Continuous Value |
| 22 | Is host login | Symbolic Value |
| 23 | Is guest login | Symbolic Value |
| 24 | Count | Continuous Value |
| 25 | Srv count | Continuous Value |
| 26 | Serror rate | Continuous Value |
| 27 | Srv serror rate | Continuous Value |
| 28 | Rerror rate | Continuous Value |
| 29 | Srv rerror rate | Continuous Value |
| 30 | Same srv rate | Continuous Value |
| 31 | Diff srv rate | Continuous Value |
| 32 | Srv diff host rate | Continuous Value |
| 33 | Dst host count | Continuous Value |
| 34 | Dst host src count | Continuous Value |
| 35 | Dst host same src rate | Continuous Value |
| 36 | Dst host diff srv rate | Continuous Value |
| 37 | Dst host same src port rate | Continuous Value |
| 38 | Dst host srv diff host rate | Continuous Value |

| | | |
|---|---|---|
| 39 | Dst host serror rate | Continuous Value |
| 40 | Dst host srv serror rate | Continuous Value |
| 41 | Dst host rerror rate | Continuous Value |
| 42 | Dst host srv rerror rate | Continuous Value |

The dataset contains both quantitative data as well as qualitative data. Both of them are represented by continuous values and symbolic values respectively. There is no missing values or outliers that can be identified. The objective that is wished to be accomplished by the time of the completion of this assignment is to effectively classify records in the test dataset into either an attack class or a non-attack class using different algorithms that are available as well as effective.

A research published in 2009 by Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani (2009) has landed its focus on the KDD Cup 99 dataset in aim to produce a detailed analysis on the dataset. The researchers have tested 7 types of classification algorithms on the dataset that has been partly extracted as the dataset for this specific assignment. The researchers have utilised these classification algorithms in the effort to learn the overall behaviours of the dataset to ultimately provide labels for each record. The result of the researchers' test is illustrated in the table below.

**TABLE 1.1 Performance Of Classification Algorithms**

| **Classification Algorithm** | **Accuracy (%)** |
|---|---|
| J48 | 93.82 |
| Naive Bayes | 81.66 |
| NB Tree | 93.51 |
| Random Forest | 92.79 |
| Random Tree | 92.53 |
| Multi-layer Perceptron | 92.26 |
| SVM | 65.01 |

Based on the results of the research, the algorithms with the 3 highest accuracy are J48, NB Tree and Random Forest with accuracies of 93.82, 93.51 and 92.79 respectively.

## 2.0 FORMULATING THE LEARNING PROBLEM

The type of problem that has been described by the dataset provided in this assignment is classification. The dataset contains a number of attributes representing a network connection that is being simulated as well as the class of the record connection which the it is categorised under. The classes of the records that are to be used in the categorisation include attack class, representing an attack connection and also non-attack class representing a normal connection in this assignment. In short, the dataset contains a large number records representing the simulated connections, each records contain a number of attributes about the connection as well as the category of either attack or non-attack for each record.

The first problem is to use data mining techniques, specifically classification algorithms in this case, to learn from the training dataset on the classification of either attack or non-attack for each records and ultimately test the accuracy of the models on the test dataset that is also included in this assignment.

The second problem is the fact that after converting the class label to only two classes, attack and non-attack, there exists an imbalance in the classes such that one class is more represented than the other one. This type of imbalance within datasets can produce highly unreliable measures of classification performance such as classification accuracy. In the dataset provided for this assignment after the conversion mentioned above, there are 97,278 instances for the non-attack class but 396,743 instances for the attack class. The imbalance is roughly presented at a ratio of 1 : 4.078.

The third problem is the fact that there exists redundancy of records contained within the dataset. These redundant records at a certain degree can cause learning algorithms to be biased towards more frequent records, at the same time learn less on the less frequent records.

## 3.0 UNDERSTAND YOUR DATA

The KDD Cup 99 Dataset used in this assignment for training consists of exactly 494,021 records and the test dataset consists of exactly 311,029 records. Each of these records represents a connection which is essentially a sequence of TCP packets starting and ending at some fixed times, between which data flows to and from a source IP address as well as a target IP address based on some defined protocol. Each of these connections has been labelled as either attack or normal with the exact attack type specified. Every connection record is about 100 bytes.

By generating scatter plots of every single attribute other than the class label in relation to the different types of attack classes as well as normal connection, a brief analysis on the attributes of the given dataset can be formed. The table containing the analysis description for

each of the attributes is shown in Table 1.0. All of the analysis included in the table are based on the plots shown in Figure 1.0, Figure 1.1 and Figure 1.2.

**TABLE 2.0 Attribute Analysis**

| Attribute | Attribute Type | Description of Analysis |
|---|---|---|
| 1 | Numerical (0-57715) | Most of attack/normal classes are highly concentrated at attribute value 0. |
| 2 | Nominal (3 types) | Most of the attack/normal classes are concentrated on tcp while pod dos. and smurf. Attack classes concentrate on icmp. |
| 3 | Nominal (65 types) | Most of the attack/normal classes are concentrated on half of the attribute value where attack class neptune. is relatively evenly scattered throughout the attribute values. |
| 4 | Nominal (11 types) | Most of the attack/normal classes are correlated with the attribute value SF where attack class neptune. is more concentrated on REJ, S0 and RST0. RSTOS0 and OTH only have minimal correlation with the portsweep. attack class. |
| 5 | Numerical (0-62825648) | Most of the attack/normal classes are concentrated on the attribute value 0. |
| 6 | Numerical (0-5203179) | |
| 7 | Numerical (0-1) | |
| 8 | Numerical (0-3) | |
| 9 | Numerical (0-3) | |
| 10 | Numerical (0-101) | |
| 11 | Numerical (0-4) | |
| 12 | Numerical (0-1) | Most of the attack/normal classes are concentrated on either the attribute value of 0 or 1 with the concentration on the value 0 slightly higher than that of 1. |
| 13 | Numerical (0-796) | Most of the attack/normal classes are concentrated on the attribute value 0. |
| 14 | Numerical (0-1) | |

| | | |
|---|---|---|
| 15 | Numerical (0-2) | |
| 16 | Numerical (0-878) | |
| 17 | Numerical (0-100) | |
| 18 | Numerical (0-5) | |
| 19 | Numerical (0-4) | |
| 20 | Numerical (-0.05 - 0.05) | |
| 21 | Numerical (0-1) | |
| 22 | Numerical (0-1) | |
| 23 | Numerical (0-511) | Most of the attack/normal classes are concentrated on attribute value 0 with the exception of normal connection and attack classes like neptune., smurf., teardrop., portsweep. and satan. extending out from attribute value 0 to 511. |
| 24 | Numerical (0-511) | Most of the attack/normal classes are concentrated on attribute value 0 with the exception of normal connection and attack classes like smurf., extending out from attribute value 0 to 511. |
| 25 | Numerical (0-1) | Most of the attack/normal classes are concentrated on attribute value 0 with minimal concentration on the attribute value 1. |
| 26 | Numerical (0-1) | |
| 27 | Numerical (0-1) | |
| 28 | Numerical (0-1) | |
| 29 | Numerical (0-1) | Most of the attack/normal classes are concentrated on attribute value 1 with minimal concentration on the attribute value 0. |
| 30 | Numerical (0-1) | Most of the attack/normal classes are concentrated on attribute value 0 with minimal concentration on the attribute value 1. |
| 31 | Numerical (0-1) | |
| 32 | Numerical (0-255) | Most of the attack/normal classes are concentrated on attribute value 255 with minimal concentration on the attribute value 0. |

| | | |
|---|---|---|
| 33 | Numerical (0-255) | Most of the attack/normal classes are relatively evenly spreaded across the range of the attribute value. |
| 34 | Numerical (0-1) | |
| 35 | Numerical (0-1) | Most of the attack/normal classes are concentrated on attribute value 0 with minimal concentration on the attribute value 1. |
| 36 | Numerical (0-1) | |
| 37 | Numerical (0-1) | |
| 38 | Numerical (0-1) | |
| 39 | Numerical (0-1) | |
| 40 | Numerical (0-1) | |
| 41 | Numerical (0-1) | |

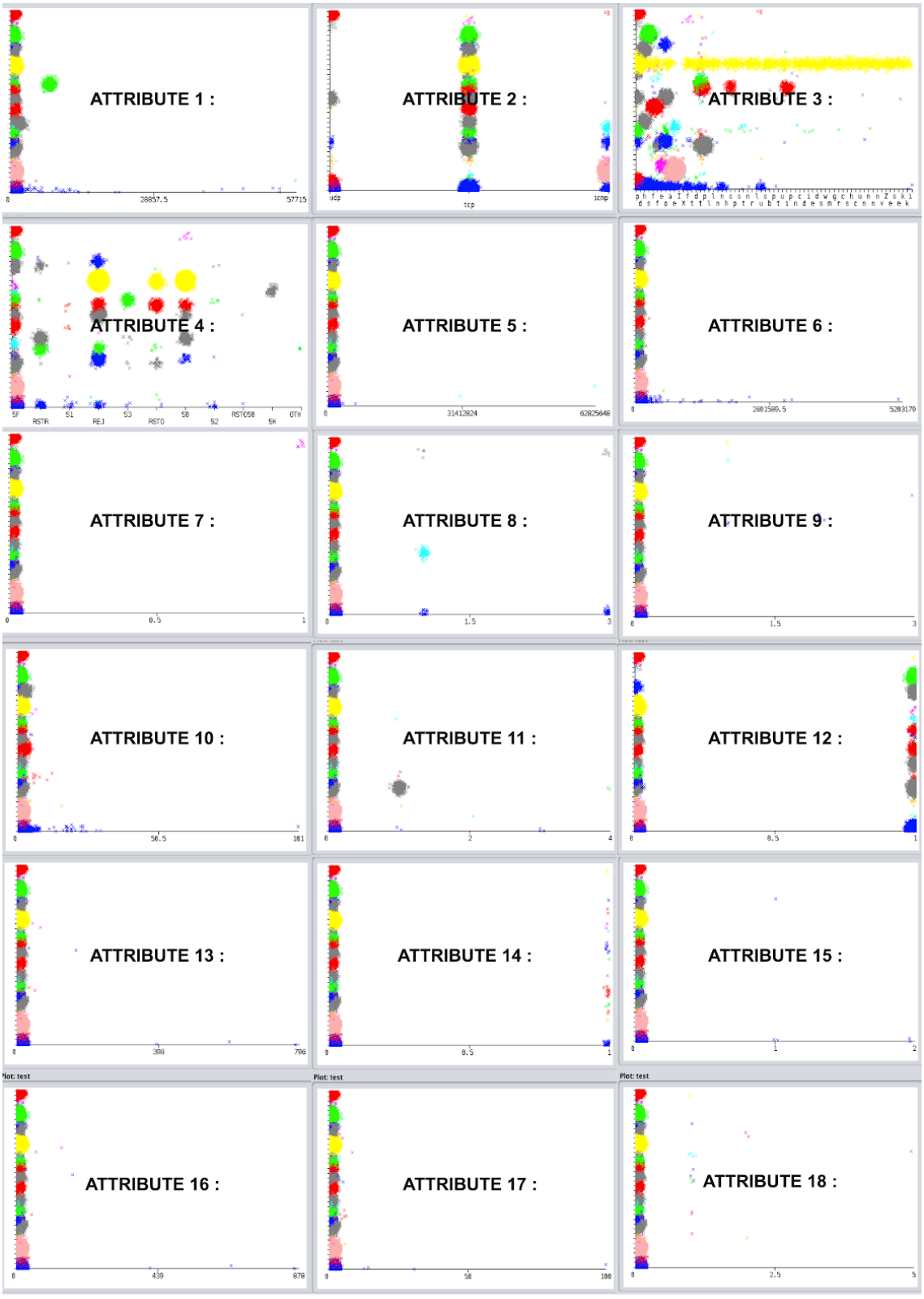


**FIGURE 2.0 Plots For All Attributes Within The Dataset (Part1)**

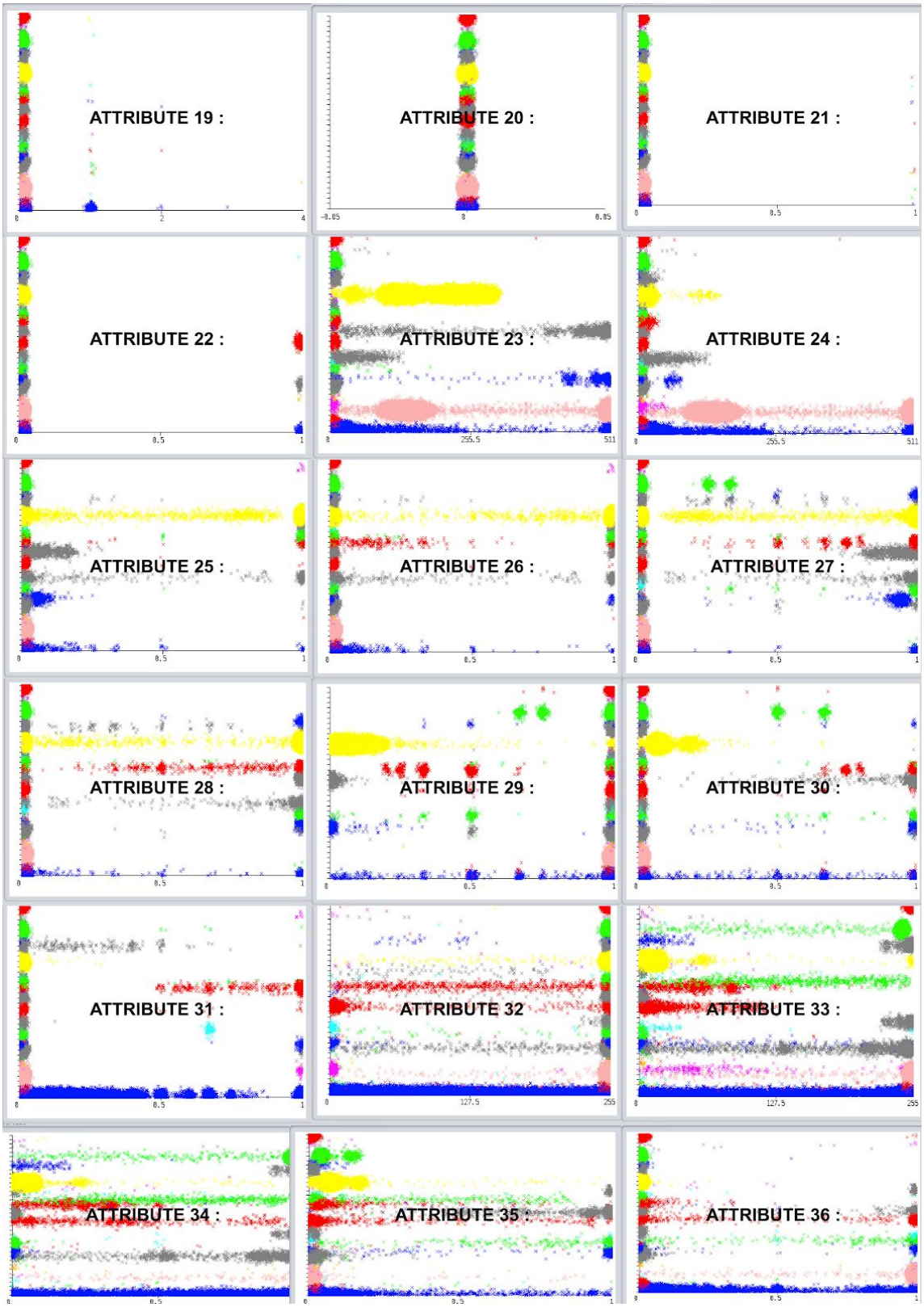


**FIGURE 2.1 Plots For All Attributes Within The Dataset (Part 2)**

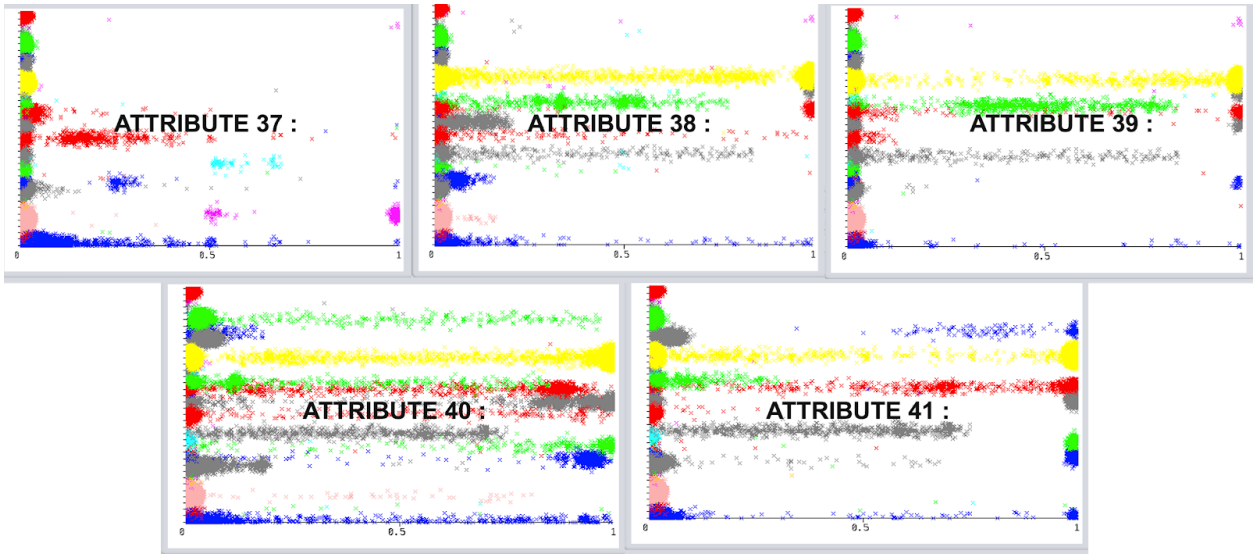


**FIGURE 2.2 Plots For All Attributes Within The Dataset (Part 3)**

## 4.0 PREPROCESSING

In this assignment, the team has decided to convert the class label for all the records into only 2 classes that are non-attack as well as attack classes for the motive of not to overcomplicate the already complicated assignment. This is done by converting all of the different types of attack classes into one class, that is attack class. In turn, the normal connection class will be convert into a non-attack class.

The next task is to remove redundant records within the data for the reason that has been discussed in the learning problem. This is done by using the unsupervised removeRuplicates filter (weka.filters.unsupervised.instance.RemoveDuplicates). After the application of the filter, the dataset has been reduced in size to 145,585 total instances, 87,832 instances being in the non-attack class and 57,753 instances being in the attack class (Figure 4.1), from a previous total instances of 494,021; 97,278 for non-attack class and 396,743 for attack class (Figure 4.0). It is a huge reduction in size.

The following preprocessing task includes resampling the test dataset into a smaller one where the imbalance issue discussed in the learning problems can be resolved. The resampling process involves the use of supervised resampling filter (weka.filters.supervised.instance. resample). The parameters used are biasToUniformClass = 1, sampleSizePercent = 60.0. The resulting dataset has 87,350 instances, 43,675 non-attack instances and 43,675 attack instances, which is being successfully balanced.

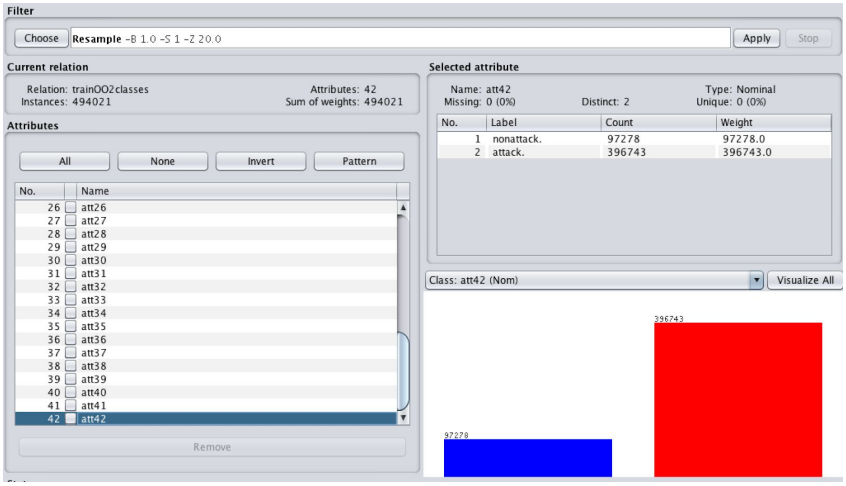


**FIGURE 4.0 Before Redundancy Reduction**

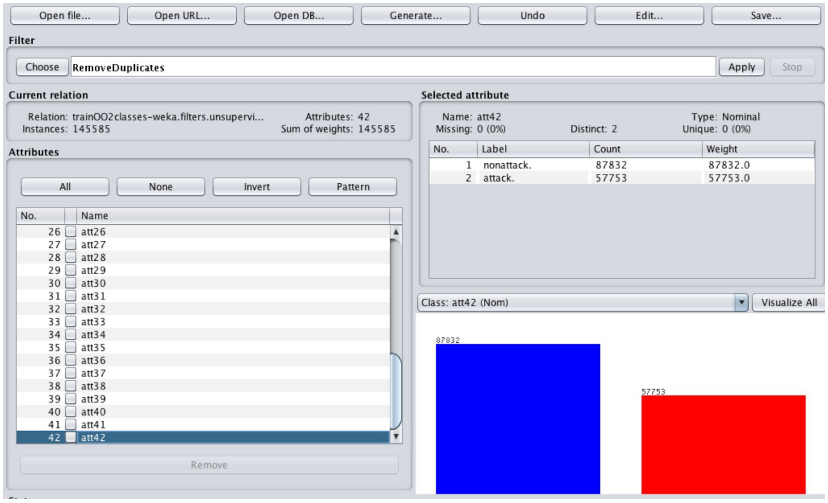


**FIGURE 4.1 After Redundancy Reduction**

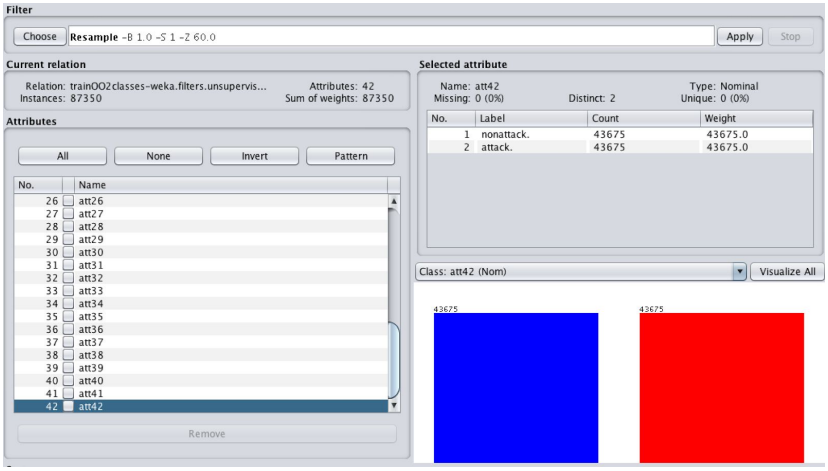


**FIGURE 4.2 : After Resampling**

Being researched by a handful of researches around the globe, it is proven that discretization can be very beneficial towards classification algorithms. To test on its impact on the three classification algorithms to be tested in this assignment, discretization is being applied to all numeric attributes through Discretize (weka.filters.supervised.attribute.Discretize).

When the task comes to identifying the most appropriate set of features or as known as attributes, the team decided to utilised three different feature selection algorithms named, CfsSubsetEval and ClassifierSubsetEval. For ClassifierSubsetEval, 3 different classifiers are chosen as the test classifier for the feature selection model, that are J48, Naive Bayes and Bayesian Network. The features selected by the different models are to be paired with the respective classifier used later in the process. The attributes selected by both of the attribute selection algorithm are shown in Table 4.0.

**TABLE 4.0 Table oF Attributes Selected by Algorithms**

| **CfsSubsetEval** | **ClassifierSubsetEval** | | |
|---|---|---|---|
| | With J48 classifier | With Naive Bayes classifier | With Bayesian Network classifier |
| Att3<br>Att4<br>Att6<br>Att8<br>Att12<br>Att26<br>Att29<br>Att30 | Att1<br>Att2<br>Att3<br>Att4<br>Att5<br>Att12<br>Att25<br>Att29<br>Att32<br>Att33<br>Att34 | Att3<br>Att4<br>Att8<br>Att23<br>At33<br>Att37 | Att3<br>Att5<br>Att8<br>Att10<br>Att13<br>Att31<br>Att36 |

## 5.0 ALGORITHM CHOICE & PERFORMANCE

Three different classification algorithms that have been used in this assignment are J48, Naive Bayes and Bayesian Network. A baseline classification performance has been tested and recorded for all three algorithms without any preprocessing procedure being applied to the datasets. The result of the classification performance is illustrated in Table 5.0.

**TABLE 5.0 Initial Classifier Performance**

| Algorithm | Accuracy | TP Rate | TN Rate | ROC Area |
|---|---|---|---|---|
| J48 | 92.648% | Overall: 0.926 | Overall: 0.022 | Overall: 0.956 |
| | | Attack: 0.910 | Attack: 0.006 | Attack: 0.956 |
| | | Non-Attack: 0.994 | Non-Attack: 0.090 | Non-Attack: 0.956 |
| Naive Bayes | 91.4677% | Overall: 0.915 | Overall: 0.039 | Overall: 0.963 |
| | | Attack: 0.900 | Attack: 0.024 | Attack: 0.959 |
| | | Non-Attack: 0.976 | Non-Attack: 0.100 | Non-Attack: 0.963 |
| Bayesian Network | 91.6484% | Overall: 0.916 | Overall: 0.029 | Overall: 0.959 |
| | | Attack: 0.899 | Attack: 0.011 | Attack: 0.961 |
| | | Non-Attack: 0.989 | Non-Attack: 0.101 | Non-Attack: 0.959 |

The experiment is proceeded with one of the preprocessing procedure, removing duplicates within the dataset. After applying the filter and testing the dataset on all three algorithms, none of them produced higher classification accuracy.

The team proceed with the resampling of the dataset on top of the duplication removal that is done previously. After applying the filter and testing the dataset on all three algorithms, only the J48 algorithm produces higher classification accuracy. The result is shown in Table 5.1.

**TABLE 5.1 Improved Accuracy of the J48 Algorithm**

| Algorithm | Accuracy | TP Rate | TN Rate | ROC Area |
|---|---|---|---|---|
| J48 | 93.0154% | Overall: 0.930 | Overall: 0.022 | Overall: 0.954 |
| | | Attack: 0.915 | Attack: 0.006 | Attack: 0.954 |
| | | Non-Attack: 0.994 | Non-Attack: 0.085 | Non-Attack: 0.954 |

Discretization as another preprocessing procedure is being applied to the dataset which is then being tested on all three of the classification algorithms again. This time around, the Naive Bayes algorithm produces a higher than before classification accuracy. The result is displayed in Table 5.2.

**TABLE 5.2 Improved Accuracy of the Naive Bayes Algorithm**

| Algorithm | Accuracy | TP Rate | TN Rate | ROC Area |
|---|---|---|---|---|
| Naive Bayes | 91.5403% | Overall: 0.915 | Overall: 0.024 | Overall: 0.967 |
| | | Attack: 0.896 | Attack: 0.004 | Attack: 0.967 |
| | | Non-Attack: 0.996 | Non-Attack: 0.104 | Non-Attack: 0.967 |

Lastly, feature selection algorithms are being put to test by removing attributes that are not selected by the two feature selection algorithms mentioned previously. Both results of the two algorithms are being applied individually to all three of the classification algorithms. This results in a higher classification accuracy for the J48 algorithm. The result can be found in Table 5.3.

**TABLE 5.3 Improved Accuracy of the J48 Algorithm**

| Algorithm | Accuracy | TP Rate | TN Rate | ROC Area |
|---|---|---|---|---|
| J48 | 93.536% | Overall: 0.935 | Overall: 0.021 | Overall: 0.957 |
| | | Attack: 0.921 | Attack: 0.007 | Attack: 0.957 |
| | | Non-Attack: 0.993 | Non-Attack: 0.079 | Non-Attack: 0.957 |

As a reference, the highest classification accuracy achievement for all three of the classification algorithms is included in Table 5.4.

**TABLE 5.4 Highest Accuracy Achieved by All 3 Algorithms**

| Algorithm | Accuracy | TP Rate | TN Rate | ROC Area |
|---|---|---|---|---|
| J48 | 93.536% | Overall: 0.935 | Overall: 0.021 | Overall: 0.957 |
| | | Attack: 0.921 | Attack: 0.007 | Attack: 0.957 |

| | | | | |
|---|---|---|---|---|
| | | Non-Attack: 0.993 | Non-Attack: 0.079 | Non-Attack: 0.957 |
| Naive Bayes | 91.5403% | Overall: 0.915 | Overall: 0.024 | Overall: 0.967 |
| | | Attack: 0.896 | Attack: 0.004 | Attack: 0.967 |
| | | Non-Attack: 0.996 | Non-Attack: 0.104 | Non-Attack: 0.967 |
| Bayesian Network | 91.6484% | Overall: 0.916 | Overall: 0.029 | Overall: 0.959 |
| | | Attack: 0.899 | Attack: 0.011 | Attack: 0.961 |
| | | Non-Attack: 0.989 | Non-Attack: 0.101 | Non-Attack: 0.959 |

## 6.0 REFERENCES

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. doi:10.1109/cisda.2009.5356528

Dataset taken from:
Kdd Cup 99 Dataset. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>