Adekitan *et al. J Big Data* (2019) 6:11
https://doi.org/10.1186/s40537-019-0176-5

Journal of Big Data

RESEARCH Open Access

Data mining approach for predicting the daily Internet data traffic of a smart university

Aderibigbe Israel Adekitan*, Jeremiah Abolade and Olamilekan Shobayo

*Correspondence: aderibigbe.adekitan@covenantuniversity.edu.ng
Department of Electrical and Information Engineering, Covenant University, Ota, Ogun State, Nigeria

Abstract

Internet traffic measurement and analysis generate dataset that are indicators of usage trends, and such dataset can be used for traffic prediction via various statistical analyses. In this study, an extensive analysis was carried out on the daily internet traffic data generated from January to December, 2017 in a smart university in Nigeria. The dataset analysed contains seven key features: the month, the week, the day of the week, the daily IP traffic for the previous day, the average daily IP traffic for the two previous days, the traffic status classification (TSC) for the download and the TSC for the upload internet traffic data. The data mining analysis was performed using four learning algorithms: the Decision Tree, the Tree Ensemble, the Random Forest, and the Naïve Bayes Algorithm on KNIME (Konstanz Information Miner) data mining application and kNN, Neural Network, Random Forest, Naïve Bayes and CN2 Rule Inducer algorithms on the Orange platform. A comparative performance analysis for the models is presented using the confusion matrix, Cohen's Kappa value, the accuracy of each model, Area under ROC Curve, etc. A minimum accuracy of 55.66% was observed for both the upload and the download IP data on the KNIME platform while minimum accuracies of 57.3% and 51.4% respectively were observed on the Orange platform.

Keywords: Machine learning, Data mining, Nigerian university, Internet data traffic, Network operations monitoring, Pattern recognition models

Introduction

In this data era where data is the new oil, internet data traffic is growing significantly each year [1, 2]. With the advent of state of the art technologies on data transmission and processing in the last decade, the internet has witnessed an increase in the intensity and the volume of internet activities globally [3]. User-generated dataset contains useful statistics and information that can be harnessed for learning but this may be challenged by privacy issues [4, 5]. Internet activities generate data traffic of various kinds; during both data download and upload. Monitoring and analysis of internet traffic is becoming more challenging daily due to sheer increase in the volume of the internet data traffic and the large capacity of connection trunks [2].

Internet traffic measurement and management is vital to the operations of Internet Service Providers for predicting future demands [6], and traffic monitoring can be achieved using flow statistics tools. Internet traffic measurement is typically deployed

Springer Open

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

by capturing process packet at a particular data monitoring point using high performance central servers and specialized tools such as Flowscan and Coralreef [7]. Internet traffic monitoring over a large network, e.g. a state-wide computer network, produces huge volume of data which may be time intensive to analyse especially in cases of global, worm and virus attacks. Hence, it is vital to ensure that an optimal methodology is deployed for traffic monitoring [8], and for generating flow statistics [2, 9, 10] using flow aggregation and packet sampling methodologies in place of continuous sampling. An innovative approach for analysing internet traffic flow using timestamp data which generates traffic analysis cookie was developed by [11].

The study by Kim et al. [12] emphasized the importance of traffic classification for uniquely identifying data traffic of certain types that ought to be blocked toward ensuring network security, and also, for preventing malicious activities [13, 14] and programs [15, 16]. In the study, machine learning algorithms using WEKA application were applied in carrying out the performance evaluation of the seven most commonly used learning algorithms for traffic classification. According to K. Claffy and Monk [17], and Kim et al. [12] there is no industry norm or standard format for comparing the performance of a network with another and neither is there a defined, best traffic classification method to apply, and as such, for the success of commercial internet the only baseline available through which organisations may be able to calibrate the performance of their network is by referencing past network performance data. This therefore emphasises the need to monitor and log internet data traffic for a comparative network performance analysis.

Apart from the analysis of internet traffic for network security reasons, internet data traffic carries a lot of useful information about the originating network. The daily volumetric variation of internet traffic creates usage pattern that can be deployed for predictive analysis which will help network engineers in preparing the network adequately for anticipated heavy internet traffic so as to ensure optimal quality of service [18–20]. Also, the quality of packet traffic may be impaired by packet losses [21–24]. The peak and off-peak internet usage periods can be determined from monitored networks and such information is vital for planning. Likewise, the capacity of the network to meet rising traffic demand can be easily observed, and this will help the network managers to respond proactively to likely future network issues due to network overloading by excessive traffic [25] and appropriate mitigations, control and possible network expansion can be deployed in a timely manner.

The study by Tokuyama et al. [26] proposed the use of day of the week and time as features for improving network prediction accuracy using Recurrent Neural Network. In [27], deep neural network was applied for predicting internet traffic by analysing the aggregated traffic data logged over a year period. The feasibility of using non-linear time series analysis for internet traffic prediction was demonstrated in the extensive study by [28]. Extracting useful information from data traffic can take different forms such as time series models, regression analysis, machine learning, and so forth. In this study, data mining algorithms were deployed for a classification analysis of the internet data traffic of a smart-community compliant private university in Nigeria for a period of one year ranging from January to December 2017.

Studies on internet traffic over time have provided various methods for improving internet traffic flow monitoring statistics and computation time [7] with focus often on traffic analysis, trend monitoring, traffic classification for categorising traffic types, and for identifying threats and malicious traffic [6, 29]. Traffic volume prediction is another vital aspect of network monitoring which is often analysed using time series (linear and non-linear), regression analysis, decomposition methods, hybrid methods etc. [26, 30, 31], and this provides an opportunity for further related studies using alternative methods, tools and features. Traffic data can be tracked and analysed over specific time interval, e.g. in minutes or hourly over a 24-h period. The focus of this study is on predictive analysis of internet traffic data using aggregated daily upload and download IP traffic over a year. Internet traffic data can be examined using tools such as neural network, time series statistics, deep learning, etc. In this study, predictive data mining models will be developed using the interactive data pipeline workflows and visual programming on KNIME and Orange platforms [32, 33]. This paper is a case study analysis that is focused on identifying unique internet traffic data trends within a university environment, and this provides an opportunity for enhancing the quality of daily service through anticipated traffic prediction. The study implements data mining analysis using the latest visual programming tools that does not demand rigorous coding, and as such, it demonstrates an alternative approach to the traditional extensive code-based data mining methods, and this can be easily implemented by network engineers for predicting daily internet traffic using well defined traffic status classification.

Data acquisition and methodology

Valuable information that can guide decision making, and the efficiency and productivity of operational processes can be extracted from historical dataset of systems and processes by applying data mining methodologies. Databases are rich sources of historical information, and as such, useful knowledge can be obtained by analysing the accumulated dataset [34, 35]. Data mining entails the use of computer applications for applying various learning algorithms that identify patterns within the dataset [36]. Data mining is a broad field that encompasses computer science and statistics. In the study by Auld et al. [6], the use of supervised learning for classifying internet traffic data was demonstrated using a trained Bayesian Neural Network, and accuracies of 95% and 98% respectively were achieved for the cases considered. Naıve Bayes classifier was applied by [37] in the basic form for classifying internet traffic, and an accuracy of 65% was achieved, also sophisticated refinements were proposed for improving the predictive accuracy. An untrained classifier was applied by McGregor et al. [38] for identifying classes of traffic with similar properties for clustering into unique groups [39]. In the study by Soule et al. [40], data flow analysis was carried out by classifying traffic into elephant flows and non-elephant flows for estimating the probability of flow-membership.

In this study, the internet traffic data of Covenant University in Nigeria over a period of one year was evaluated and analysed using predictive data mining algorithms. The data was logged using Mikrotik Hotspot Manager and FreeRADIUS, Radius Manager Web application deployed on LINUX platform as implemented by Adeyemi et al. [18] through the SmartCU cluster. The dataset logged contains the Upload (in GigaBytes) and the download (in GigaBytes) internet traffic data from the 1st of January to the

19th of December when the school closed for the year in 2017. During data preparation, the actual day of the week (Monday to Sunday) was captured to allow the model to identify any hidden unique data usage pattern for each day of the week within a specific week and month. Covenant University runs a stable academic calendar which is fixed for each year, and as such there is a high tendency that specific academic activities within the university might be causal factors influencing internet traffic for each day. Hence, if such unknown, regular, daily-activity driven internet usage patterns were identified, it would be easy using the acquired knowledge to forecast the anticipated data usage for any specific day and date in the next academic year. This forecast information will help network engineers prepare adequately towards maintaining top-notch quality of service. To achieve this goal, an extensive methodology was deployed to process the dataset, and this comprises data cleaning, data sorting, extraction of descriptive statistics, data normalization and coding, and implementation of classification algorithms to train and classify the data and evaluate the performance of the algorithms.

From the yearlong dataset, four unique quarters were identified as shown in Figs. 1 and 2. The quarters are based on the minimum, the lower quartile, the median, the upper quartile and the maximum values of each parameter. Based on the quartiles, the internet traffic for each day was classified into four categories as shown in Table 1.

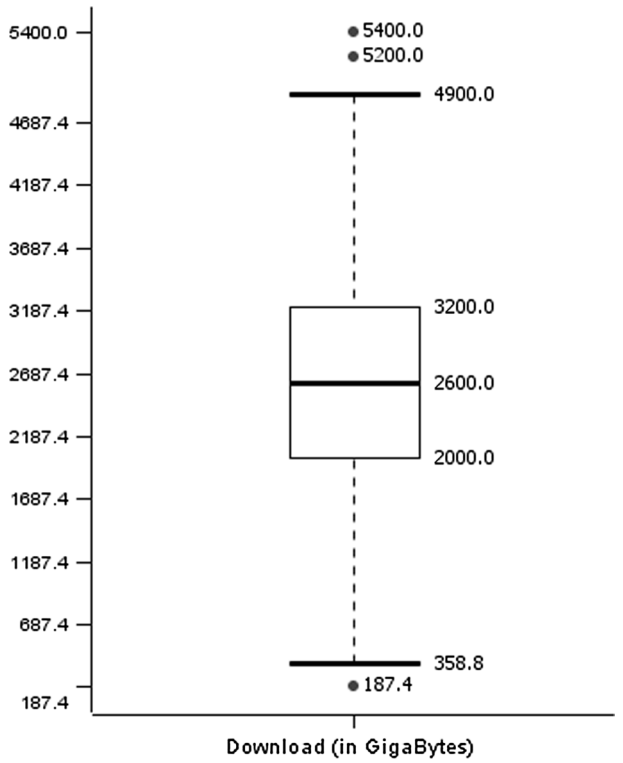

Fig. 1 Box plot of the download internet traffic. The boxplot shows the variation in the daily download internet traffic for a year across four quartiles

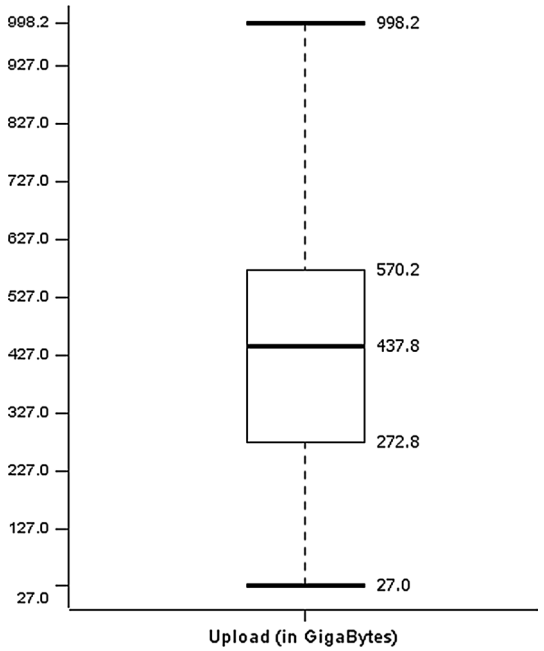

Fig. 2 Box plot of the upload internet traffic. The boxplot shows the variation in the daily upload internet traffic for a year across four quartiles

Table 1 Internet data traffic classification

Traffic status classification (TSC)	Quartile	Coding	Download (GB)	Upload (GB)
Heavy data traffic	Max	HDT	> 3200	> 570.2
Moderate data traffic	Q3	MDT	3200 ≤ TSC > 2600	570.2 ≤ TSC > 437.8
Slight data traffic	Median	SDT	2600 ≤ TSC > 2000	437.8 ≤ TSC > 272.8
Low data traffic	Q1	LDT	TSC ≤ 2000	TSC ≤ 272.8

The model

Six features were analysed in the data mining model for predicting the IP download traffic and these are: month, week (week 1 to week 51), the day of the week (Monday to Sunday), the daily download traffic for the previous day, the average daily download traffic for the two previous days, and the TSC for the download internet traffic data. Likewise, for the IP upload traffic the following features were considered: month, week, the day of the week, the daily upload traffic for the previous day, the average daily upload traffic for the two previous days, and the TSC for the upload internet traffic data. The data mining analysis was performed using four learning algorithms: Tree Ensemble, Decision Tree, Random Forest, and Naïve Bayes learner and predictor nodes on KNIME data mining application, and K-nearest neighbour (kNN), Random Forest, Neural Network, Naïve Bayes and CN2 Rule Inducer on the Orange data mining platform. The KNIME and Orange data mining platforms were combined in this study for an extensive analysis, and to identify significant variations in result between the two platforms, if any.

For the whole year, internet traffic data samples were captured and analysed for 353 days. 70% of the data samples were used for training the learning algorithm while the remaining 30% was applied for evaluating the performance of the trained model. The dataset was imported into the model using the Excel Reader. The numeric parameters were normalized to prevent size-based bias at the learning stage. The processed dataset was applied to the configured learner algorithms and the model results were exported for evaluation. The KNIME-based model showing the data workflow is available in Appendix as Fig. 18.

Based on the confusion matrix generated by each predictive data mining algorithm; model performance measures such as the accuracy, the F-measure, etc. can be determined using Eqs. 1 to 6 [41–43]. Given that the correctly predicted positive samples are referred to as True Positive (TP), the incorrect positive predictions as False Positive (FP), correctly predicted negative samples as True Negative (TN), and incorrect negative predictions as False Negative (FN). The accuracy of the machine learning algorithm as expressed in Eq. (1) is the percentage of the correct predictions made by the model with respect to the total number of predictions.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP) + (TN + FN)} \tag{1}$$

A dataset is said to be unbalanced when the number of instances is significantly unequal among the classes or when a particular instance is not observed at all. Imbalance ratio varies from dataset to dataset, and it may create a bias towards the majority class. The use of accuracy as a performance measure is inadequate for unbalanced dataset. For such cases, the balanced accuracy is more suitable as defined in Eq. (2).

$$\text{Balanced Accuracy} = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{2}$$

For each class, the precision is the number of correctly classified samples out of the total samples classified in that particular class. It is mathematically defined in Eq. (3).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

For each class, the recall is the number of correctly classified samples out of the total samples that are truly in that particular class. It is mathematically defined in Eq. (4).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

The F-measure or F-score is the harmonic mean of the recall and the precision as defined in Eq. (5).

$$\text{F - measure} = \left(\frac{recall^{-1} + precision^{-1}}{2} \right)^{-1} \tag{5}$$

The error rate of the machine learning algorithm is defined by Eq. (6)

$$\text{Error} = \frac{(FP + FN)}{(TP + FP) + (TN + FN)} \tag{6}$$

Table 2 Descriptive statistics of the internet traffic data for the year 2017

	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Sum
Download (in GigaBytes)	187.4	5400	2482.567	926.8042	858,966.1	−0.22454	−0.07779	876,346.2
Upload (in GigaBytes)	27	998.2	420.1399	189.4003	35,872.46	−0.02044	−0.71202	148,309.4

Table 3 Logistic distribution fitting model parameters for the internet download traffic

	Parameter estimate	Standard error	Estimated covariance of parameter estimates	
			Mean	Scale
Mean	2523.67	49.1392	2414.67	−53.8113
Scale	528.401	23.3857	−53.8113	546.892
Log likelihood	−2915.46			
Domain	−Inf < y < Inf			
Variance	918,558			

The traffic status classification of the aggregated IP traffic flow $Q(n)$ for $day(n)$ in the university under study is mapped using data mining classification as a function of knowledge acquired from five key variables (day, week, month, the traffic for the previous day, and the average daily traffic for the two previous days) as expressed in Eqs. (7) and (8) for the daily upload and download internet traffic respectively, where $Q(n-1) \rightarrow Q(n) \rightarrow Q(n+1)$ implies daily traffic variation.

$$TSC\ Q_u(n) = \left[day(n), week(n), month(n), Q_u(n-1), \frac{Q_u(n-1) + Q_u(n-2)}{2}\right] \tag{7}$$

$$TSC\ Q_d(n) = \left[day(n), week(n), month(n), Q_d(n-1), \frac{Q_d(n-1) + Q_d(n-2)}{2}\right] \tag{8}$$

Descriptive statistics of the dataset

The statistical properties of the dataset are summarized in this section. Table 2 presents the descriptive statistics of the internet traffic data while Table 3 presents the parameters of the Logistic Distribution model which was used to fit the internet download traffic data. Table 4 shows the Logistic Distribution model parameters for fitting the internet upload traffic data. The Internet traffic variations across the 51 weeks is presented in Fig. 3 for the download traffic and in Fig. 4 for the upload traffic. The average, weekly internet traffic size for the download and upload IP traffic is presented in Fig. 5. Figures 6 and 7 show the probability density plot and the cumulative probability plot of the internet download traffic data while Figs. 8 and 9 show the probability density plot and the cumulative probability plot of the internet upload traffic data.

Table 4 Logistic distribution fitting model parameters for the internet upload traffic

	Parameter estimate	Standard error	Estimated covariance of parameter estimates	
			Mean	Scale
Mean	422.681	10.5987	112.332	− 1.01822
Scale	112.28	4.86845	− 1.01822	23.7018
Log likelihood	− 2362.51			
Domain	− Inf < y < Inf			
Variance	41,474.7			

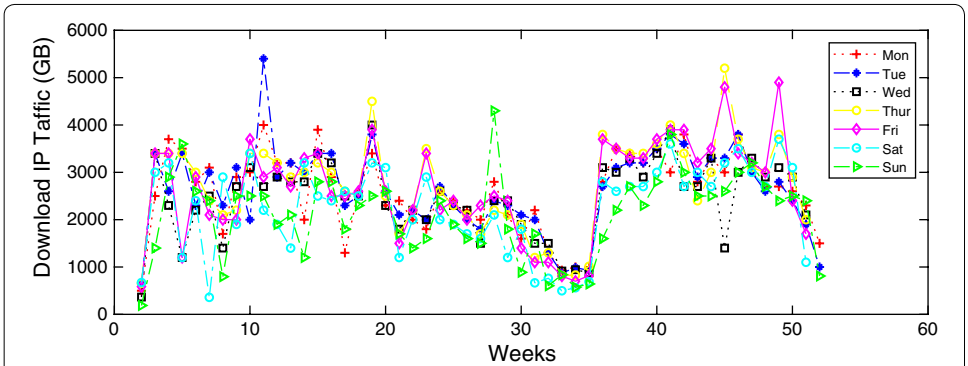

Fig. 3 Internet download traffic variations across the 51 weeks

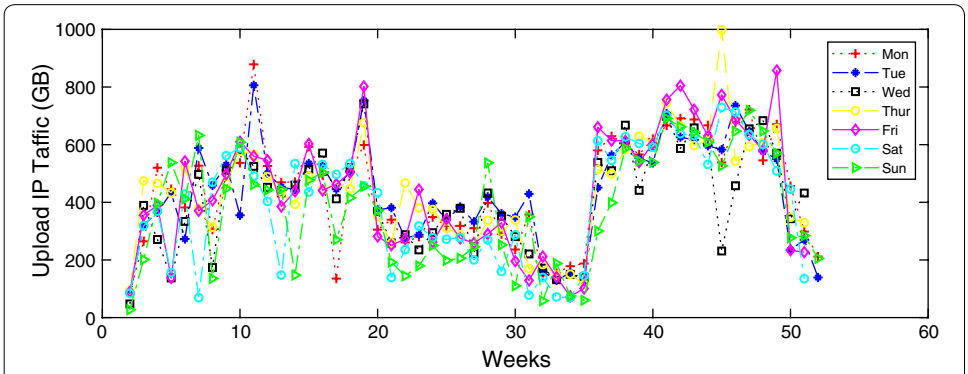

Fig. 4 Internet upload traffic variations across the 51 weeks

Results and discussion

The Decision Tree, the Tree Ensemble, the Random Forest, and the Naïve Bayes learners on KNIME platform were trained using 70% of the dataset. On the Orange platform; the kNN, Neural Network, Random Forest, Naïve Bayes and CN2 Rule Inducer data mining algorithms were trained using 70% random sampling with stratified shuffle split which ensures that the percentage of the samples for each class is preserved in the training and testing data divisions. The result of the predictive model evaluation using the remaining 30% of the data is presented in this section. The predictive

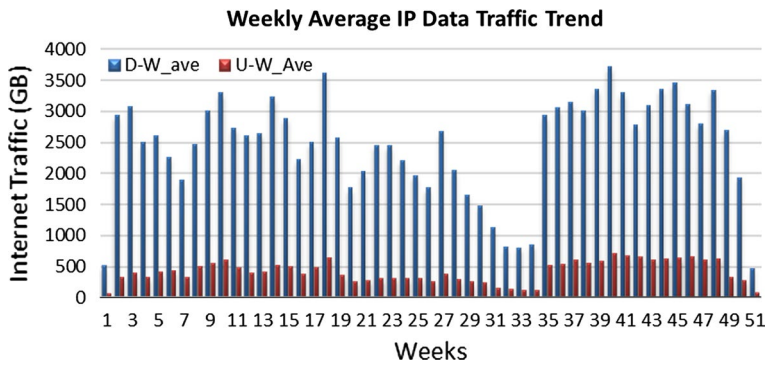

Fig. 5 Average weekly internet traffic size for the download and upload data

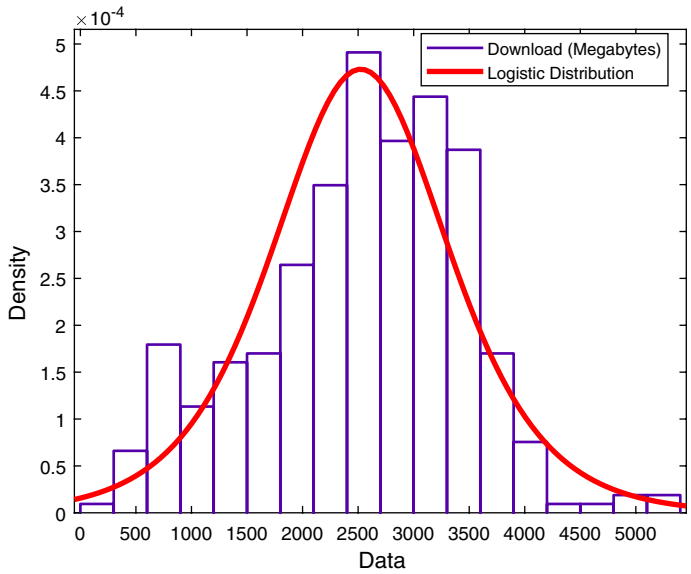

Fig. 6 Probability density plot of the internet download traffic

analysis was carried out in two parts: for the download and the upload traffic data using the four predictive learners for each as presented in the following sections. The KNIME workflow implemented for the classification analysis is presented in the Appendix as Fig. 18.

Results for the KNIME based model

A. Internet download traffic data

i. The Ensemble Tree Algorithm

The Ensemble Tree learner was able to accurately predict the Traffic Status Classification (TSC) for 62.264% of the test samples. The confusion matrix for the Ensemble Tree predictor is presented in Table 5.

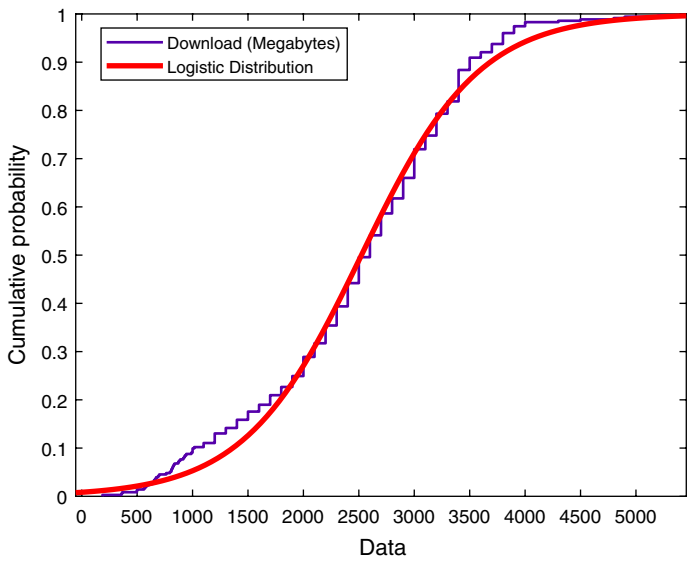

Fig. 7 Cumulative probability plot of the internet download traffic

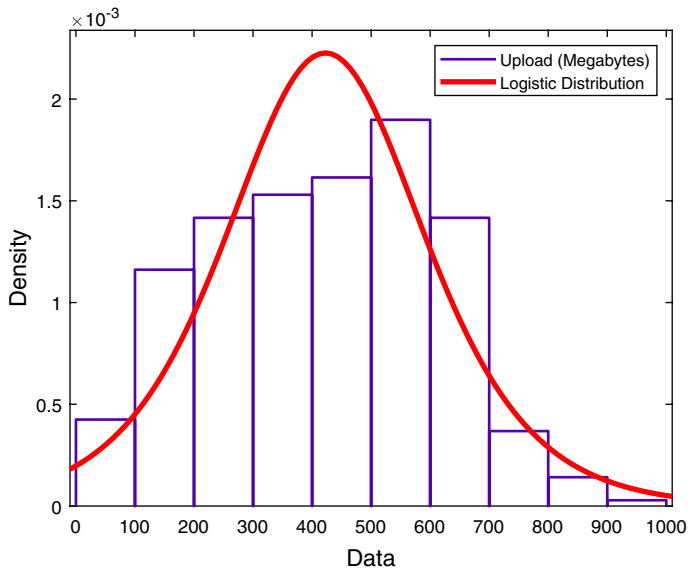

Fig. 8 Probability density plot of the internet upload traffic

ii. Decision Tree Algorithm
The Decision Tree learner was able to accurately predict the Traffic Status Classification (TSC) for 55.66% of the test samples. The confusion matrix for the Decision Tree predictor is presented in Table 6.

iii. Random Forest Algorithm
The Random Forest learner was able to accurately predict 60.377% of the model evaluation test samples with a Cohen's Kappa (k) value of 0.465. The confusion matrix for the Random Forest predictor is presented in Table 7.

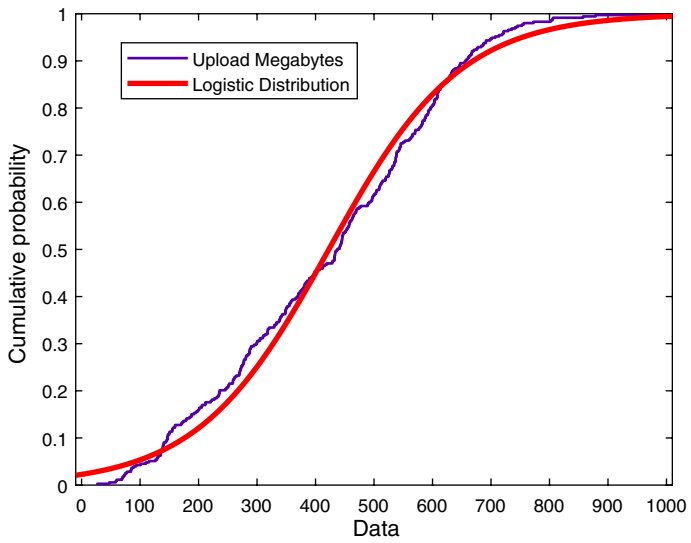

Fig. 9 Cumulative probability plot of the internet upload traffic

Table 5 Confusion matrix for the Tree Ensemble predictor

	LDT	SDT	MDT	HDT
LDT	27	3	0	0
SDT	9	15	1	2
MDT	4	1	13	4
HDT	4	6	6	11

Table 6 Confusion matrix for the Decision Tree Predictor

	LDT	SDT	MDT	HDT
LDT	23	4	1	2
SDT	5	16	0	6
MDT	2	5	11	4
HDT	4	10	4	9

Table 7 Confusion matrix for the Random Forest Predictor

	LDT	SDT	MDT	HDT
LDT	27	3	0	0
SDT	8	15	1	3
MDT	6	1	12	3
HDT	4	9	4	10

iv. Naïve Bayes Algorithm

The Naïve Bayes Algorithm is a probabilistic classifier which applies the Bayes theorem with naïve independence assumptions among the classified features. The

Naïve Bayes Algorithm accurately predicted 59.434% of the total test samples with a Cohen's Kappa value of 0.454. The confusion matrix for the Naïve Bayes predictor is presented in Table 8.

B. Internet upload traffic data

i. The Ensemble Tree Algorithm

Similar to the prediction for the internet download traffic analysis, the Ensemble Tree Algorithm was able to accurately predict the Traffic Status Classification for 62.264% of the model evaluation test samples. The confusion matrix for the Ensemble Tree predictor is presented in Table 9. A comparison of Tables 5 and 9 for the Ensemble Tree Algorithm shows that although the accuracy for both the internet upload and download traffic prediction are the same but the items misclassified in both cases are different.

ii. Decision Tree Algorithm
The Decision Tree learner for the upload IP traffic had a predictive accuracy of 55.66%. The confusion matrix for the Decision Tree predictor is presented in Table 10.

iii. Random Forest Algorithm
The Random Forest learner was able to accurately predict 63.208% of the model evaluation test samples with a Cohen's Kappa (k) value of 0.51. The confusion matrix for the Random Forest predictor is presented in Table 11.

Table 8 Confusion matrix for the Naïve Bayes Predictor

	LDT	SDT	MDT	HDT
LDT	26	4	0	0
SDT	7	16	1	3
MDT	3	1	8	10
HDT	3	3	8	13

Table 9 Confusion matrix for the Ensemble Tree Predictor

	LDT	SDT	MDT	HDT
LDT	16	9	1	0
SDT	11	13	2	0
MDT	3	3	19	2
HDT	2	1	6	18

Table 10 Confusion matrix for the Decision Tree Predictor

	LDT	SDT	MDT	HDT
LDT	14	10	2	0
SDT	12	12	2	0
MDT	1	10	14	2
HDT	0	2	6	19

Table 11 Confusion matrix for the Random Forest Predictor

	LDT	SDT	MDT	HDT
LDT	18	7	1	0
SDT	9	14	3	0
MDT	4	2	17	4
HDT	2	1	6	18

Table 12 Confusion matrix for the Naïve Bayes Predictor

	LDT	SDT	MDT	HDT
LDT	15	10	1	0
SDT	8	15	2	1
MDT	3	1	15	8
HDT	0	1	5	21

Table 13 The confusion analysis for the four machine learning algorithms on KNIME platform

	Tree Ensemble		Decision Tree		Random Forest		Naïve Bayes	
	True positives	False positives	True positives	False positives	True positives	False positives	True positives	False positives
Internet download traffic								
LDT	27	17	23	11	27	18	26	13
SDT	15	10	16	19	15	13	16	8
MDT	13	7	11	5	12	5	8	9
HDT	11	6	9	12	10	6	13	13
Overall	66	40	59	47	64	42	63	43
Internet upload traffic								
LDT	16	16	14	13	18	15	15	11
SDT	13	13	12	22	14	10	15	12
MDT	19	9	14	10	17	10	15	8
HDT	18	2	19	2	18	4	21	9
Overall	66	40	59	47	67	39	66	40

iv. Naïve Bayes Algorithm
The Naïve Bayes Algorithm accurately predicted 62.264% of the test samples with a Cohen's Kappa value of 0.497. The confusion matrix for the Naïve Bayes predictor is presented in Table 12.

The comparison of the performances of the KNIME based Decision Tree, Tree Ensemble, the Random Forest, and the Naïve Bayes learners is presented as a summary in Tables 13 and 14. The F-measure statistics is presented in Table 15.

Results for the Orange data mining platform

Orange is an open source data mining and machine learning software for explorative data analysis using visual programming. According to the developers, Orange is a fruitful and fun way of deploying data mining interactively for fast qualitative data analysis.

Table 14 Comparison of the performance of the four data mining algorithms on KNIME platform

	Tree Ensemble	Decision Tree	Random Forest	Naïve Bayes
Internet download traffic				
Correct classified	66	59	64	63
Accuracy	62.264%	55.66%	60.377%	59.434%
Cohen's Kappa (k)	0.492	0.403	0.465	0.454
Wrong classified	40	47	42	43
Error	37.736%	44.34%	39.623%	40.566%
Internet upload traffic				
Correct classified	66	59	67	66
Accuracy	62.264%	55.66%	63.208%	62.264%
Cohen's Kappa (k)	0.492	0.409	0.51	0.497
Wrong classified	40	47	39	40
Error	37.736%	44.34%	36.792%	37.736%

Table 15 Comparison of the F-measure statistics

	Tree Ensemble	Decision Tree	Random Forest	Naïve Bayes
Internet download traffic				
LDT	0.7297	0.7188	0.7200	0.7536
SDT	0.5769	0.5161	0.5455	0.6275
MDT	0.6190	0.5789	0.6154	0.4103
HDT	0.5000	0.3750	0.4651	0.4906
Internet upload traffic				
LDT	0.5517	0.5283	0.6102	0.5769
SDT	0.5000	0.4000	0.5600	0.5660
MDT	0.6909	0.5490	0.6296	0.6000
HDT	0.7660	0.7917	0.7347	0.7368

Five machine learning algorithms were applied on the Orange platform to explore the upload and download IP traffic data and these are: kNN, Random Forest, Neural Network, Naïve Bayes and CN2 Rule Inducer algorithm. The samples were randomly selected using stratified shuffle split and the result of the analysis is presented in the following sections using the average over classes. The performance of the algorithms is compared using the Classification Accuracy (CA), Area under ROC Curve (AUC), the Precision rate, the Recall, and the F1 score. The Orange workflow is presented in the Appendix section as Fig. 19.

Internet Download Traffic Data

Table 16 shows a comparative performance analysis for the five machine learning algorithms deployed on the Orange platform for analysing the download internet traffic data. For a visual appreciation of the variation in the performance of each of the machine learning algorithms on the Orange platform, the AUC is presented using the receiver operating characteristic (ROC) curve which is a probability curve that plots sensitivity; that is, the true positive rate on the y-axis against the false positive rate (1-specificity).

Table 16 Comparative evaluation of the performance of the data mining algorithms using Orange software

Algorithm	AUC	F1	CA	Recall	Precision
kNN	0.752	0.507	0.514	0.514	0.505
Random Forest	0.812	0.564	0.566	0.566	0.563
Neural Network	0.823	0.605	0.606	0.606	0.607
Naive Bayes	0.838	0.586	0.588	0.588	0.587
CN2 rule inducer	0.731	0.549	0.546	0.54	0.555

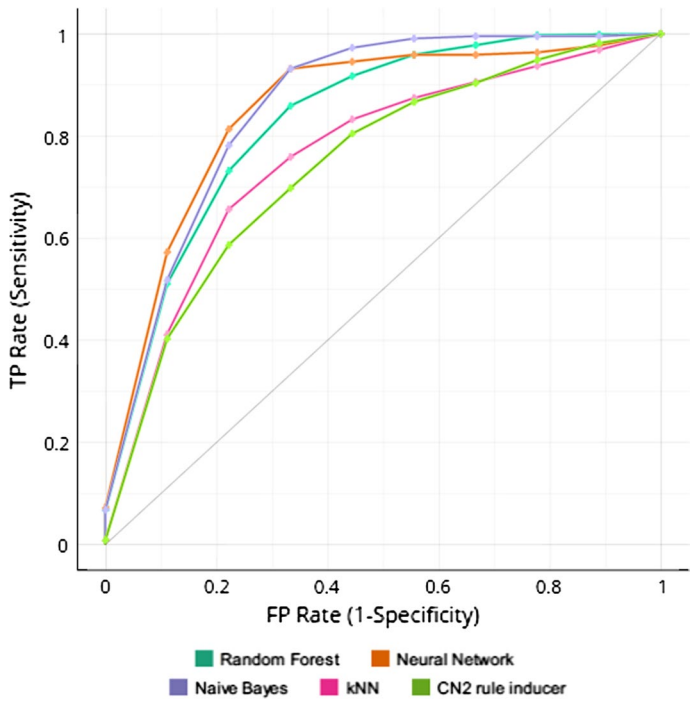

Fig. 10 ROC for the HDT IP download TSC

The ROC curve is plotted in Fig. 10 for the heavy data traffic (HDT) internet download, IP traffic status classification while Fig. 11 shows the ROC curve for the moderate data traffic (MDT) internet download, IP traffic status classification. Figures 12 and 13 present the ROC curve for the internet download, IP traffic status classification for the slight data traffic (SDT) and low data traffic (LDT) respectively.

Internet upload traffic data

Table 17 shows a comparative performance analysis for the five data mining algorithms deployed on the Orange platform for the upload IP traffic. For the internet upload IP traffic, the ROC curve is plotted in Fig. 14 for the HDT internet upload, IP traffic status classification while Fig. 15 shows the ROC curve for the MDT internet upload IP traffic status classification. Figures 16 and 17 present the ROC curve for the SDT and LDT respectively.

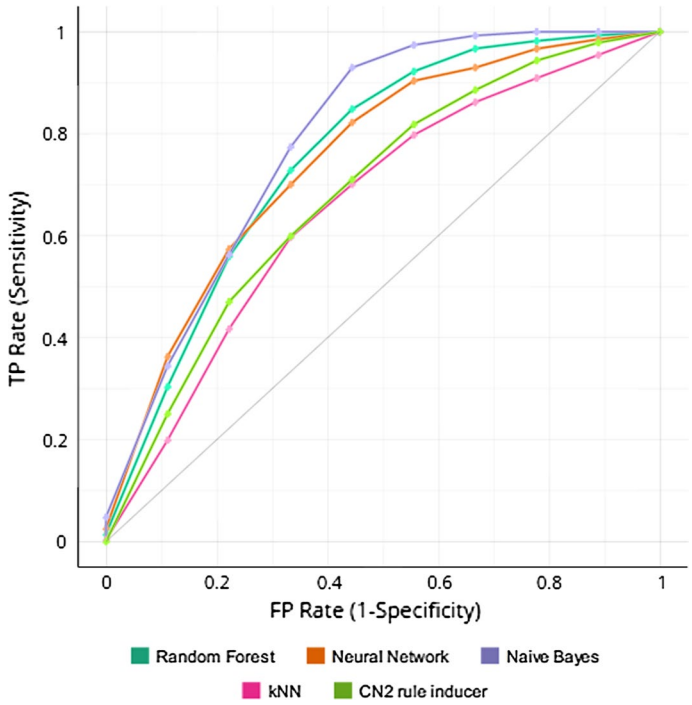

Fig. 11 ROC for the MDT IP download TSC

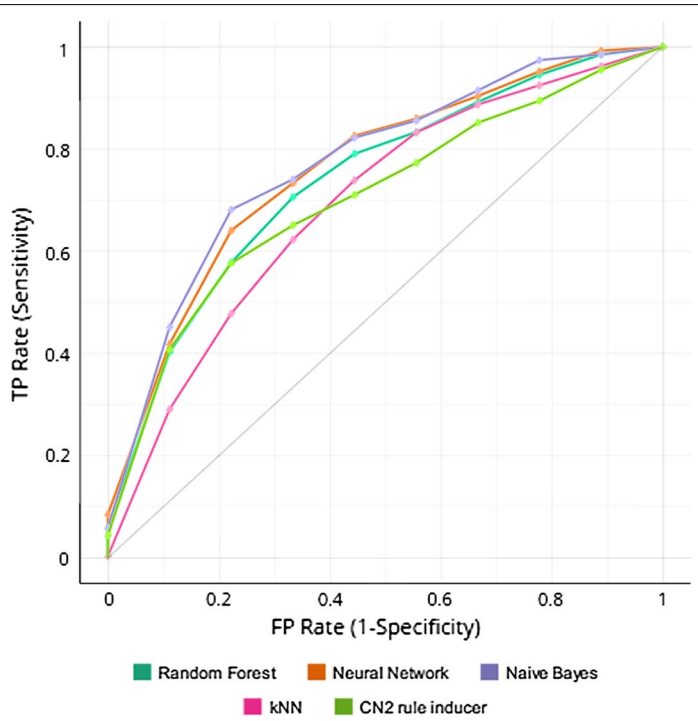

Fig. 12 ROC for the SDT IP download TSC

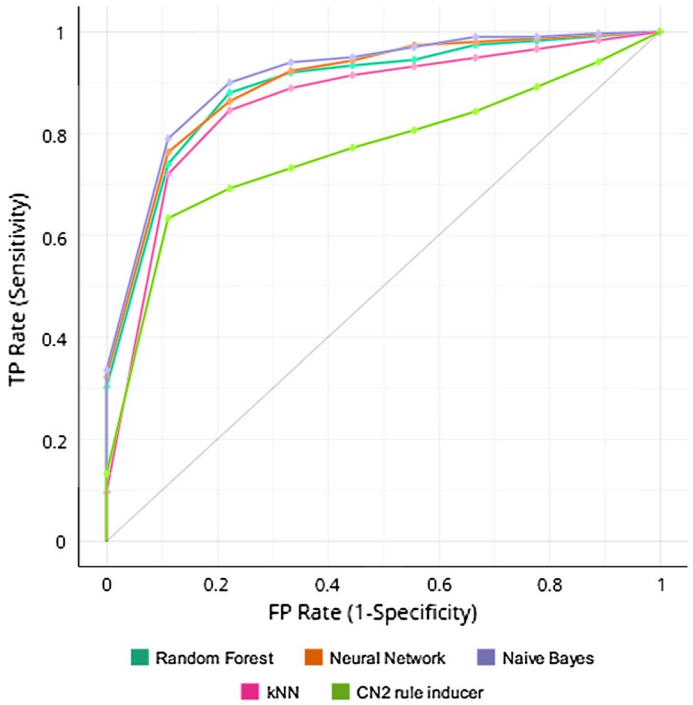

Fig. 13 ROC for the LDT IP download TSC

Table 17 Comparative evaluation of the performance of the data mining algorithms using Orange software

Algorithm	AUC	F1	CA	Recall	Precision
kNN	0.796	0.566	0.573	0.573	0.563
Random Forest	0.848	0.605	0.605	0.605	0.605
Neural Network	0.846	0.627	0.625	0.625	0.630
Naive Bayes	0.876	0.638	0.639	0.639	0.637
CN2 rule inducer	0.794	0.613	0.610	0.610	0.625

Summary of the models' predictive performance

In terms of predictive accuracy, for the internet download traffic, the order of model accuracy is as follows for the KNIME-based model: Tree Ensemble > Random Forest > Naïve Bayes > Decision Tree while for the internet upload traffic the order is Random Forest > Tree Ensemble = Naïve Bayes > Decision Tree. The analysis shows that the Decision Tree predictor had the worst performance in both cases which implies that the Decision Tree Algorithm may not be very optimal for predicting internet data traffic using historical internet traffic data without modifications to the model. For the Orange data mining platform, in terms of the AUC for the download traffic, the order of performance is as follows: Naive Bayes > Neural Network > Random Forest > kNN > CN2 rule inducer while for the upload traffic the order is Naive Bayes > Random Forest > Neural Network > kNN > CN2 rule inducer.

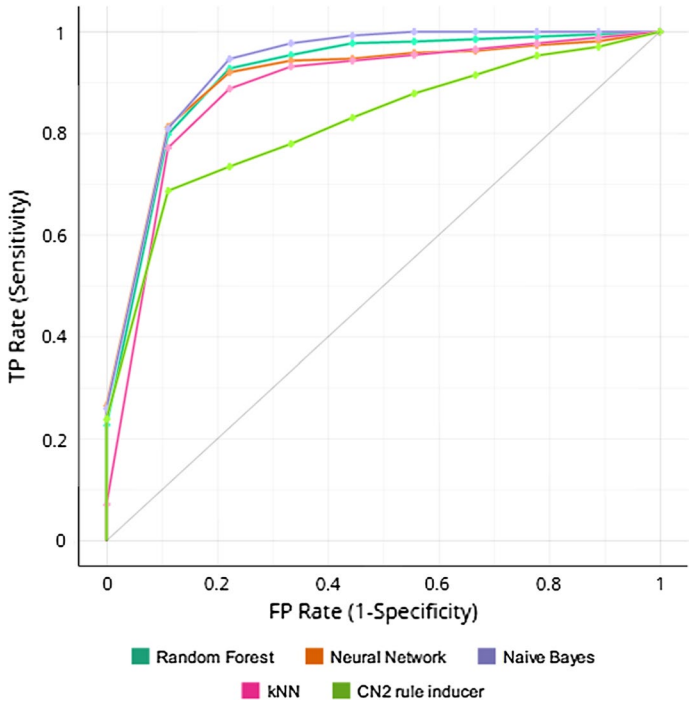

Fig. 14 ROC for the HDT IP upload TSC

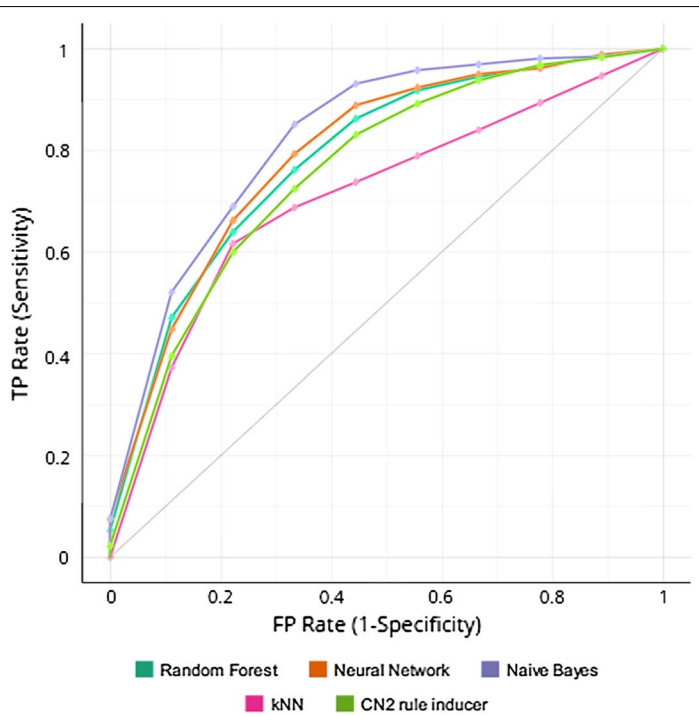

Fig. 15 ROC for the MDT IP upload TSC

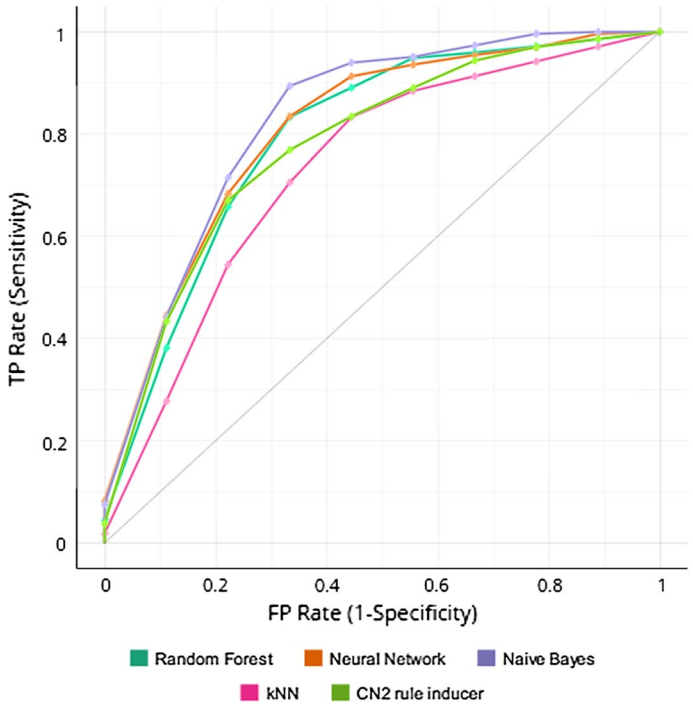

Fig. 16 ROC for the SDT IP upload TSC

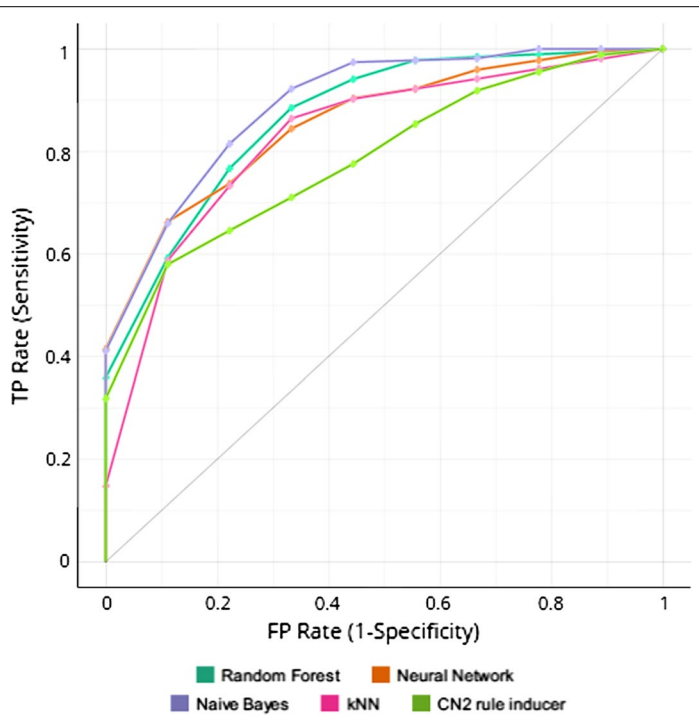

Fig. 17 ROC for the LDT IP upload TSC

Conclusion

Internet data traffic monitoring and measurement is vital to the operations of Internet Service Providers, and this can be achieved using flow-based traffic monitoring approach. The logged internet traffic data acquired through traffic monitoring contains useful information and knowledge which can be accessed via data analysis. In this study, the upload and download internet traffic data generated in Covenant University, in Nigeria for the year 2017 was statistically analysed and predictive KNIME and Orange based models were developed for forecasting internet data traffic on a given day using the traffic data of the previous days. The Tree Ensemble, the Decision Tree, the Random Forest, and the Naïve Bayes data mining algorithms were applied on the KNIME model while the Naive Bayes, Neural Network, Random Forest, kNN and the CN2 rule inducer were applied on the Orange platform as a supervised-learning data mining model for predictive analysis.

The algorithms were effectively trained with 70% of the dataset samples while the remaining 30% was applied for model evaluation. The model performance evaluation result shows that the Tree Ensemble predictor had the best accuracy while the Decision Tree predictor had the least accuracy for the internet download prediction on KNIME. The Naïve Bayes and the Tree Ensemble predictors had the same accuracy for the internet upload traffic, and the Decision Tree predictor once again had the least accuracy for the upload traffic analysis on KNIME. The least accuracy recorded for all the cases considered is 55.66% while the maximum accuracy is 63.208%. This shows that data mining approach using interactive, visual data pipeline workflows is reasonably accurate for predicting internet traffic trends in a smart university but further studies will be required in order to improve the performance of the models.

Abbreviations

TSC: traffic status classification; HDT: heavy data traffic; MDT: moderate data traffic; SDT: slight data traffic; LDT: low data traffic; KNIME: Konstanz information miner; kNN: K-nearest neighbour; WEKA: Waikato environment for knowledge analysis; ROC: receiver operating characteristic; TP: true positive; FP: false positive; TN: true negative; FN: false negative.

Authors' contributions

AIA conceptualized the methodology and prepared the dataset for analysis. All authors contributed to the analysis, result interpretation and manuscript development. All authors read and approved the final manuscript.

Acknowledgements

The Authors appreciate Covenant University Centre for Research, Innovation and Discovery (CUCRID) for creating a productive research environment and for supporting the publication of this research.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset analysed in this study is available via the SmartCU Research Cluster of Covenant University [18] at (https://ars.els-cdn.com/content/image/1-s2.0-S2352340918308126-mmc1.xlsx).

Funding

Not applicable.

Appendix

The KNIME workflow in Fig. 18 shows the data pipeline from the first stage (input) where the data is imported into the model as an excel file, the data is pre-processed and then supplied to the data mining algorithms for knowledge acquisition. The output stages consist of excel writers, scorers, PMML writer and scatter plot nodes. Figure 19 shows the data workflow on the Orange platform.

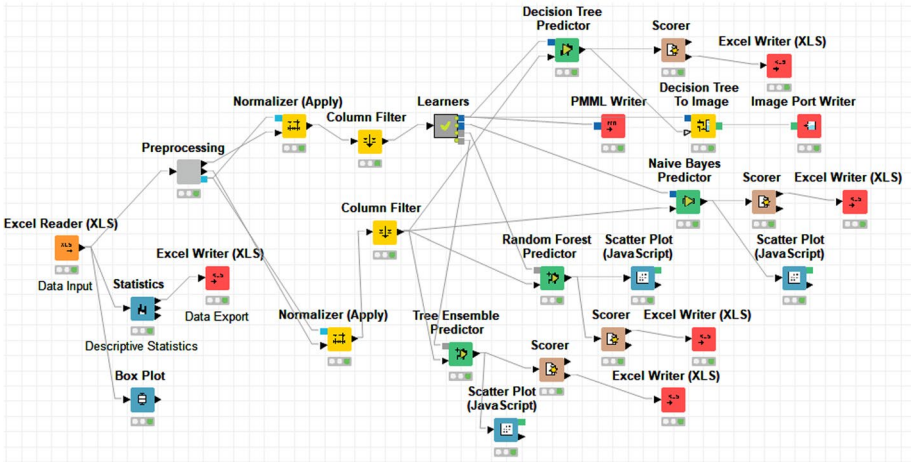

Fig. 18 The KNIME model

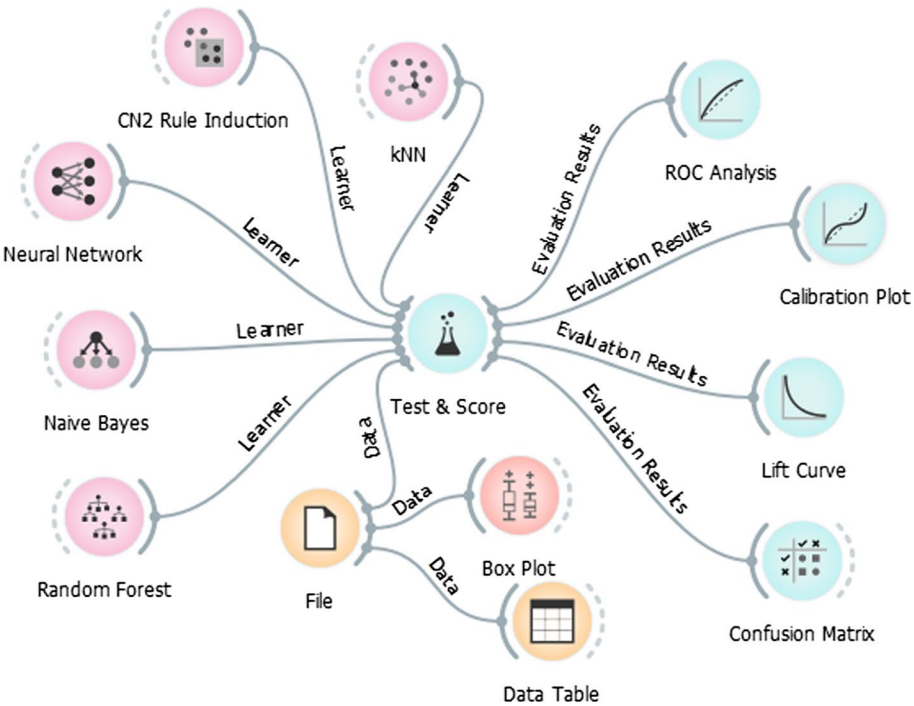

Fig. 19 The Orange model

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 13 December 2018 Accepted: 22 January 2019
Published online: 04 February 2019

References

1. Coffman KG, Odlyzko AM. Internet growth: Is there a "Moore's Law" for data traffic? Handbook of massive data sets. Berlin: Springer; 2002. p. 47–93.
2. Thompson K, Miller GJ, Wilder R. Wide-area Internet traffic patterns and characteristics. IEEE Network. 1997;11:10–23.
3. Odlyzko AM. Internet traffic growth: sources and implications. Optical Trans Syst Equip WDM Netw. 2003;2:1–16.
4. Ram P, Murali Krishna S, Siva Kumar AP. Privacy preservation techniques in big data analytics: a survey. J Big Data. 2018;5:33.
5. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. Journal of Big Data. 2018;5:1.
6. Auld T, Moore AW, Gull SF. Bayesian neural networks for internet traffic classification. IEEE Trans Neural Networks. 2007;18:223–39.
7. Lee Y, Kang W, Son H. An internet traffic analysis method with map reduce. In: Network operations and management symposium workshops (NOMS Wksps), 2010 IEEE/IFIP. 2010, p. 357–361.
8. Brandauer C, Iannaccone G, Diot C, Ziegler T, Fdida S, May M. Comparison of tail drop and active queue management performance for bulk-data and web-like internet traffic. In: Proceedings sixth IEEE symposium on computers and communications. 2001, p. 122–9.
9. Claffy KC, Polyzos GC, Braun HW. Traffic characteristics of the T1 NSFNET backbone. In: IEEE INFOCOM'93 proceedings twelfth annual joint conference of the ieee computer and communications societies. networking: foundation for the future. 1993, p. 885–92.
10. Coffman KG, Odlyzko AM. The size and growth rate of the Internet. First Monday. 1998;3:l–25.
11. Glommen C, Barrelet B. Internet website traffic flow analysis using timestamp data. Google Patents, 2004.
12. Kim H, Claffy KC, Fomenkov M, Barman D, Faloutsos M, Lee K. Internet traffic classification demystified: myths, caveats, and the best practices. In: Proceedings of the 2008 ACM CoNEXT conference, 2008, p. 11.
13. Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions. In: ACM SIGCOMM computer communication review. 2005, p. 217–28.
14. Othman SM, Ba-Alwi FM, Alsohybe NT, Al-Hashida AY. Intrusion detection model using machine learning algorithm on Big Data environment. J Big Data. 2018;5:34.
15. Mohammadkhani S, Esmaeilpour M. A new method for behavioural-based malware detection using reinforcement learning. Int J Data Mining Model Manag. 2018;10:314–30.
16. Chowdhury S, Khanzadeh M, Akula R, Zhang F, Zhang S, Medal H, et al. Botnet detection using graph-based feature clustering. J Big Data. 2017;4:14.
17. Claffy K, Monk T. What's next for Internet data analysis? Status and challenges facing the community. Proc IEEE. 1997;85:1563–71.
18. Adeyemi OJ, Popoola SI, Atayero AA, Afolayan DG, Ariyo M, Adetiba E. Exploration of daily internet data traffic generated in a smart university campus. Data Brief. 2018;20:30–52.
19. Markelov O, Duc VN, Bogachev M. Statistical modeling of the Internet traffic dynamics: to which extent do we need long-term correlations? Physica A. 2017;485:48–60.
20. Al-Turjman F. Information-centric framework for the Internet of Things (IoT): traffic modeling and optimization. Future Gener Comput Syst. 2018;80:63–75.
21. Lakshman TV, Madhow U. The performance of TCP/IP for networks with high bandwidth-delay products and random loss. IEEE/ACM Trans Netw. 1997;5:336–50.
22. S. S. Lor, R. Landa, M. Rio. Packet re-cycling: eliminating packet losses due to network failures. In: Proceedings of the 9th ACM SIGCOMM workshop on hot topics in networks, Monterey, California, 2010.
23. Caballero-Águila R, Hermoso-Carazo A, Linares-Pérez J. Networked distributed fusion estimation under uncertain outputs with random transmission delays, packet losses and multi-packet processing. Signal Process. 2019;156:71–83.
24. Alotaibi SS. Enhanced packet loss calculation in wireless sensor networks. Berlin: Springer; 2019. p. 73–81.
25. Okokpujie K, Emmanuel C, Noma-Osaghae E, Odusanmi M, Okokpujie IP. A unique mathematical queuing model for wired and wireless networks. Int J Civil Eng Technol. 2018;9:810–31.
26. Tokuyama Y, Fukushima Y, Yokohira T. The effect of using attribute information in network traffic prediction with deep learning. In: 2018 international conference on information and communication technology convergence (ICTC). 2018, p. 521–5.
27. Narejo S, Pasero E. An application of internet traffic prediction with deep neural network. Multidisciplinary approaches to neural computing. Berlin: Springer; 2018. p. 139–49.
28. M. Hasegawa, G. Wu, M. Mizuni. Applications of nonlinear prediction methods to the internet traffic. In: The 2001 IEEE international symposium on circuits and systems, 2001. ISCAS 2001. 2001, p. 169–72.
29. Abdalla BMA, Hamdan M, Mohammed MS, Bassi JS, Ismail I, Marsono MN. Impact of packet inter-arrival time features for online peer-to-peer (P2P) classification. Int J Electric Comput Eng. 2018;8:2521–30.
30. Xu F, Lin Y, Huang J, Wu D, Shi H, Song J, et al. Big data driven mobile traffic understanding and forecasting: a time series approach. IEEE Trans Serv Comput. 2016;9:796–805.
31. Kong F, Li J, Jiang B, Song H. Short-term traffic flow prediction in smart multimedia system for Internet of Vehicles based on deep belief network. Future Gener Comput Syst. 2018;93:460–72.
32. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME-the Konstanz information miner: version 2.0 and beyond. ACM SIGKDD Expl Newsl. 2009;11:26–31.
33. KNIME. *KNIME Analytics Platform*. 2018. https://www.knime.com/knime-software/knime-analytics-platform. Accessed 27 Dec 2018.
34. Çakır A, Çalış H, Küçüksille EU. Data mining approach for supply unbalance detection in induction motor. Exp Syst Appl. 2009;36:11808–13.
35. Azevedo A. Data mining and knowledge discovery in databases. Encyclopedia of information science and technology. 4th ed. Pennsylvania: IGI Global; 2018. p. 1907–18.

36. Ait-Mlouk A, Agouti T, Gharnati F. Mining and prioritization of association rules for big data: multi-criteria decision analysis approach. J Big Data. 2017;4:42.
37. Moore AW, Zuev D. Internet traffic classification using bayesian analysis techniques. ACM SIGMETRICS Perf Eval Rev. 2005;33:50–60.
38. A. McGregor, M. Hall, P. Lorier, J. Brunskill. Flow clustering using machine learning techniques. In International workshop on passive and active network measurement. 2004, p. 205–14.
39. Mehrotra S, Kohli S, Sharan A. To identify the usage of clustering techniques for improving search result of a website. Int J Data Mining Model Manag. 2018;10:229–49.
40. Soule A, Salamatia K, Taft N, Emilion R, Papagiannaki K. Flow classification by histograms: or how to go on safari in the internet. ACM SIGMETRICS Perf Eval Rev. 2004;32:49–60.
41. Al-Sheikh ES, Hasanat MH. Social media mining for assessing brand popularity. IJDWM. 2018;14(1):40–59.
42. D. M. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.
43. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27:861–74.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com