Assessing the effectiveness of data mining tools in classifying and predicting road traffic congestion

Areen Arabiat, Muneera Altayeb
Department of Communications and Computer Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan

Article Info

Article history:

Received Jan 29, 2024
Revised Feb 9, 2024
Accepted Feb 16, 2024

Keywords:

K-nearest neighbor support
Logistic regression
Machine learning
Orange data mining tool
Random forest
Vector machine
WEKA data mining tool

ABSTRACT

Traffic congestion is a significant issue in cities, impacting the environment, commuters, and the economy. Predicting congestion is crucial for efficient network operation, but high-quality data and computational techniques are challenging for scientists and engineers. The revolution of data mining and machine learning has enabled the development of effective prediction methods. Machine learning (ML) approaches have shown potential in predicting traffic congestion, with classification being a key area of study. Open-source software tools WEKA and Orange are used to predict and classify traffic congestion. However, there is no single best strategy for every situation. This study compared the effectiveness of both data mining tools for predicting congestion in one of the areas of the capital of the Hashemite Kingdom of Jordan, Amman, by testing several classifiers including support vector machine (SVM), K-nearest neighbors (KNN), logistic regression (LR), and random forest (RF) classifications. The results showed that the Orange mining tool was superior in predicting traffic congestion, with a prediction accuracy of 100% for Random forest, logistic regression, and 99.8% for KNN. On the other hand, results were better in WEKA for the SVM classifier with an accuracy of 99.7%.

Corresponding Author:

Areen Arabiat
Department of Communications and Computer Engineering, Faculty of Engineering
Al-Ahliyya Amman University
Al-Saro, Al-Salt, Amman, Jordan
Email: a.arabiat@ammanu.edu.jo

1. INTRODUCTION

Traffic congestion is one of the most important problems that residents of capital cities around the world suffer from. It can lead to increased stress, delayed delivery, fuel waste, and financial losses. From this standpoint, studies that contribute to reducing this traffic phenomenon are extremely important [1]. Most modern studies of congestion forecasting are based on analyzing peak traffic periods, where forecasting is classified into three types according to traffic flow: short-term, medium-term, and long-term forecasting. Short-term forecasts which last between five and fifteen minutes on average have a lot of random volatility, high complexity, and poor data stability. Given the complexity of the traffic condition, it is imperative to provide reliable short-term forecasting for real-time information determination, On the other hand, medium- and long-term forecast units often extend to days, weeks, months, and years, and because of the huge time lag, the stability of the data is very high; Therefore, this type of forecasting is often used to estimate long-term traffic flow with high accuracy through time series that rely on past data and expected future data [2]. Recent advances in traffic congestion prediction have given rise to an important topic of study, particularly in AI and ML. The vast availability of data aided by navigation systems and fixed sensors has contributed to a significant expansion of this subject of study over the past several decades as traffic data can be analyzed, pattern recognized, and traffic flow insights using ML techniques [3], [4]. Accurate forecasting contributes to traffic flow volume control, traffic management, and optimization, as predicting traffic congestion using ML is considered more accurate than traditional methods, which contributes significantly to improving traffic flow, especially at peak times. However, to fully utilize the potential of machine learning in traffic management, issues such as data reliability and model interpretability must be resolved [5], [6].

The researchers in [7], used monitoring-based data from IoT sensors embedded in smart cities to develop traffic control systems that operate autonomously and reduce traffic jams, to obtain great accuracy and low error rates, a neuro-fuzzy algorithm was used. In validation testing, the model outperformed previous techniques with an accuracy rate of 98%. To create a traffic prediction model, this study uses radio frequencies. The RF algorithm features great accessibility, high stability, and outstanding reliability. Liu and Wu in [8] presented an automated system to predict expected traffic using an RF classifier. The RF algorithm features great accessibility, high stability, and outstanding reliability. The traffic forecast proposed model is created using input data including weather, time of day, season, unusual road conditions, traffic conditions, and holidays. The results showed that the traffic prediction model created using the RF classification approach can be predicted effectively, has a modest generalization error, and has an accuracy rate of 87.5%. Li *et al.* [9] presented a model using an RF classification technique to predict traffic congestion state in another study. The random forest method has a reputation for being practical, flexible, and highly effective. Weather, time of day, unique road conditions, road quality, and holidays were used as model input factors to build road traffic forecasting models. As a result, the results show that the traffic prediction model developed using the random forest classification method can be predicted successfully with an accuracy of 87.5% with low generalization error. It is also more adept at predicting crowded scenarios due to its fast calculation speed.

On the other hand, researchers in [10] developed a long short-term memory (LSTM) network as a means of predicting congestion propagation across road networks, where the model predicts congestion propagation over 5 minutes in Buxton, UK, which is a congested city. The study used both univariate and multivariate LSTM models, with the former relying entirely on the speed recorded over the past five minutes while the latter takes traffic flow rate and vehicle progress into account. The accuracy of the models ranged between 84-95% depending on the route configuration and these results revealed that both models may produce adequate prediction of congestion propagation over short periods, with accuracy mostly determined by the topology of the local road network. The researchers in [11] developed a proposed model for accurate prediction of short-term traffic conditions in smart transportation. Intelligent transportation systems (ITS) systems using machine learning classifiers including LD-SVM, decision forests, MLP, and CN2 rule induction. According to the results, decision forests outperformed other methods with an average improvement of 0.982 and 0.975, respectively. This method solves the problem of overfitting in existing modelling methodologies. Ratra and Gulia [12], presented a comparison of different techniques using empirical and parameter analysis to evaluate two open data mining tools, WEKA and Orange. The results reveal that WEKA outperforms Orange in terms of the qualities required for a fully functional and easy-to-use rating platform. WEKA is suitable for data mining classification challenges. Additionally, the study analysed proactivity and recall across datasets, and found that Orange had 82.4% greater proactivity and 80.6% greater recall. WEKA has greater precision (83.7%) and recall (83.7%). This comparison includes Naïve Bayes, Random Forest, and nearest neighbor classifiers. The precision value of the k-nearest algorithm is larger, with WEKA having a precision of 75.3% and a recall of 75.2%. This work presents a unique comparison between two distinct data mining tools, where a large data set containing approximately 8,671 records was tested and the accuracy of the evaluations was approximately 100%. This unique experiment for this study aims to determine the volume of traffic flow through one of the most traffic-congested districts in the Jordanian capital, Amman.

2. METHOD

By testing the efficiency of several classifiers, the model proposed in this work provides a comparison between the WEKA and Orange data mining tools, to measure the effectiveness of both tools for predicting traffic congestion in the Jordanian capital, Amman, as the Greater Amman Municipality provided the traffic data that was used in the classification process. A system architecture for predicting traffic Congestion is shown in Figure 1, where the basic steps are stated. The first and most important stage in developing a machine learning classifier is pre-processing to improve the quality of the dataset, making it ready to feed a machine learning model. In order to predict traffic congestion in this work, the classifiers RF, SVM, KNN, and LR were employed to find the optimum data mining technique and classifier for traffic congestion prediction. The results will be evaluated using confusion matrices such as accuracy, sensitivity, precision, and F-measure.

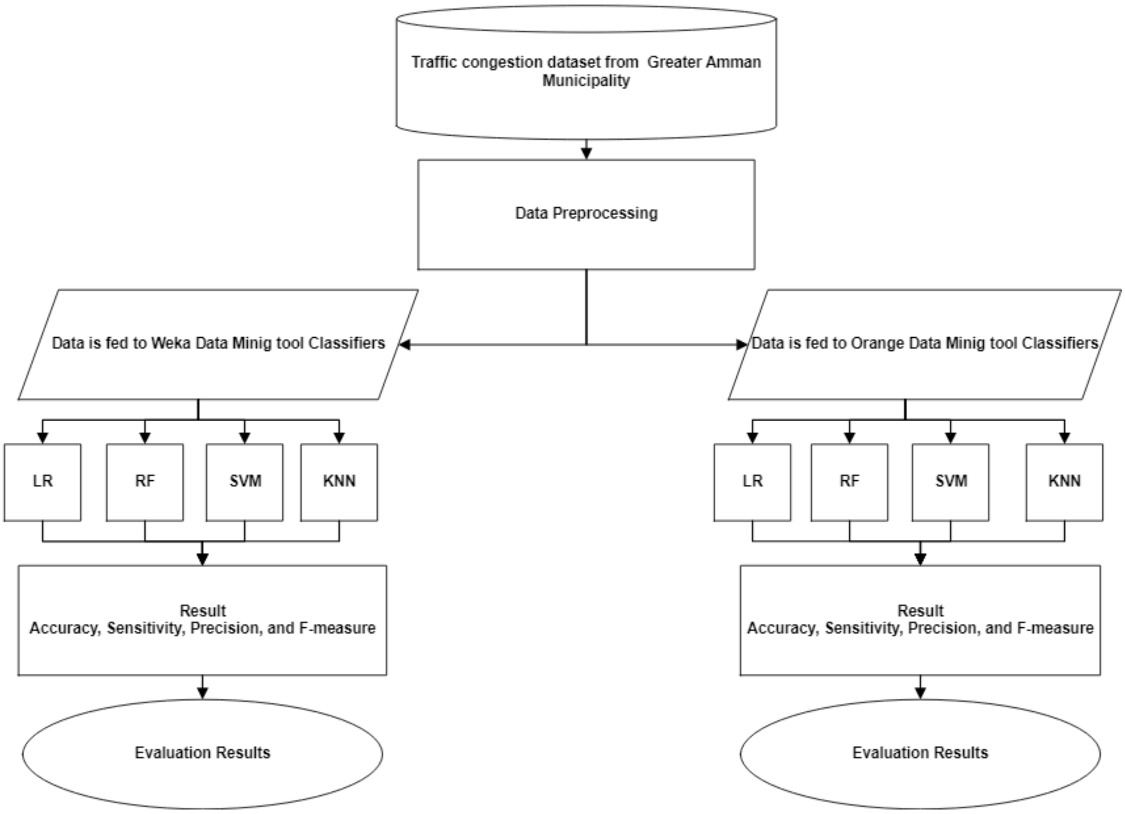

Traffic congestion dataset from Greater Amman Municipality
Data Preprocessing
Data is fed to Weka Data Minig tool Classifiers
Data is fed to Orange Data Minig tool Classifiers
LR
RF
SVM
KNN
LR
RF
SVM
KNN
Result
Accuracy, Sensitivity, Precision, and F-measure
Result
Accuracy, Sensitivity, Precision, and F-measure
Evaluation Results
Evaluation Results

Figure 1. System architecture for predicting traffic congestion

2.1. Dataset

The Greater Amman Municipality provided the data set for King Abdullah Street in Amman, in 2018. This data included time, date, traffic flow, capacity, number of lanes, road width, and traffic volume. This data was collected with high accuracy using detectors and sensors that can count the number of passing vehicles on a lane and calculate the traffic volume for each lane approach every hour of every day, every month, for the whole year [13].

2.2. Data preprocessing

Data preprocessing is the process of removing or correcting erroneous, incomplete, or incorrect data from a dataset. Use excel's remove duplicates tool to get rid of unnecessary data, then use conditional formatting to fix any structural issues. To prepare the traffic dataset for use in building an ML classifier using WEKA and Orange data mining tools, it is first saved as a CSV file.

2.3. Classification

Data classification is done in two steps: i) training, sometimes called learning and ii) testing, or evaluation, when an instance's predicted class is compared to its actual class. If the hit rate is considered acceptable by the analyst, then the classifier is considered capable of classifying future occurrences of unknown classes. Normal and congested are the two categories into which the data in this investigation should be divided. The model will be validated using 10-fold cross-validation with 30% of the data used for testing and 70% of the data used for training through RF, SVM, KNN, and LR classifiers.

2.3.1. Random forest

High-dimensional datasets cannot be effectively used with the RF-supervised learning algorithm. An infinite forest is created by combining M numbers of different decision trees [14]. Random forests use decision trees arranged randomly on information units, creating forecasts and determining the best solution through voting. This method provides a useful insight into trait significance [15]. A composite classifier generates multiple decision trees and integrates them for efficient outcomes. The random forest model uses decision trees trained on random characteristics but typically ignores the diverse contributions of trees in different test instances. The aggregate that each tree provides individually is averaged to make predictions. Also, incredibly adept at adjusting to sacrifice is random forest [16], [17]. In classification issues, the towing rule, deviation, and Gini index are the primary rules used to binary divide data of these guidelines, the most often applied is the Gini index in (1), which quantifies the node impurity:

$$\mu = \sum_{a=1}^{A} \mathrm{P}_a(1 - \mathrm{P}_a) \tag{1}$$

The target class is A, and the sample fraction of class an is P_a. A node with a modest value of μ is considered to be pure, meaning that it has good class separation and mostly comprises observations from a single class [18]. Figure 2 depicts the RF structure, Where X = X_1, X_2, X_3,..., X_N, and n is the number of data dimensions or predictive variables, which is an example of an input data set, While T expresses trees T_1(X), T_2(X), T_3(X), T_n(X) that form the RF model [18].

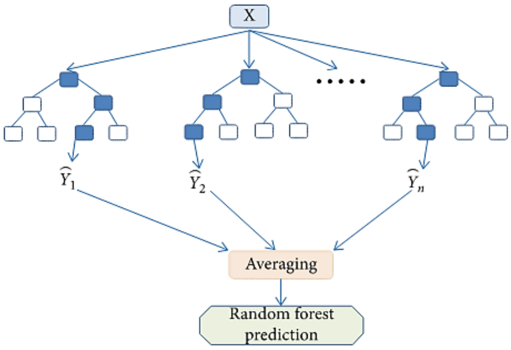


Figure 2. Random forest (RF) structure [19]

2.3.2. Logistic regression

Gaussian-form numerical input variables are used in the binary classification process when an outcome in regression modeling is binary or dichotomous (yes/no). This is a specific case that is known as logistic regression, where each input value has a coefficient that is then transformed via a logistic function. This fast method works well for a variety of classification problems [20]. Assuming a straight-line relationship between independent variables, linear regression is a frequently used kind of regression analysis for continuous outcomes. It is useful for determining how one independent variable affects a continuous result. Nevertheless, it is preferable to use multivariate linear regression to find distinct contributions while concurrently assessing the effects of several components [21]. The logistic regression model has a particular form that is described in (2):

$$\frac{Prob(Yi=1)}{Prob(Yi=0)} = \frac{Pi}{1-Pi} = e^{(\beta_0 + \beta_1 X_{1+} \cdots\cdots + \beta_k X_{ki})} \tag{2}$$

where (1-Pi) is the chance that Y takes a value of 0, Pi is the probability that Y takes a value of 1, and e is the exponential constant [22].

2.3.3. Support vector machine

The machine learning technique known as SVM is grounded in statistical learning theory, as its algorithm can determine the best classification hyperplane by maximizing the interval, as described in Figure 3 where a dataset with two features (x1 and x2) and two classes (0 and 1) [23], [24]. With the use of support vector machine technology, data points may be classified by finding a hyperplane in an N-dimensional space. There are several different hyperplanes for the separation of any two classes of data points. Our goal is to find a plane that has the most margin. Future data points can be classified more easily by maximizing the margin distance, which offers some reinforcement. The main flaw with support vector machines is that they are limited to binary problem classification [25].

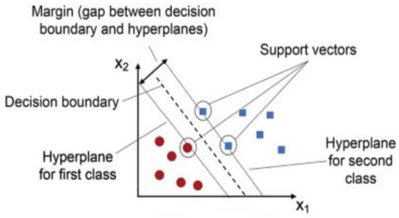

Margin (gap between decision boundary and hyperplanes)
Support vectors
x_2
Decision boundary
Hyperplane for first class
Hyperplane for second class
x_1

Figure 3. Description of SVM [26]

2.3.4. K-nearest neighbor

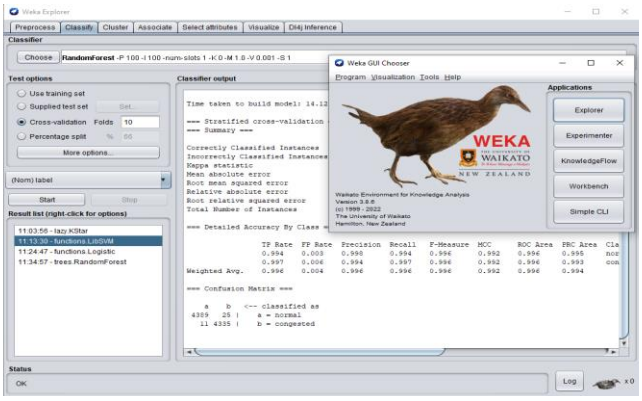
KNN classification is a simple data mining technique that forecasts a set's state based on its K closest neighbors in the training set. It does not require any training but it faced challenges in selecting K values and performing neighbor searches and distance calculations. KNN is a supervised machine learning method based on neighbor similarity, classified based on the distance between data points. The primary obstacles facing KNN are as follows: i) selecting K values and ii) neighbor search and neighbor selection, encompassing neighbor search and distance calculation [27], [28].

2.4. Data mining tools

Data mining software tools are necessary for both the development and implementation of data mining techniques. The process of selecting the best tool gets easier as there are more and more options accessible [29]. The technique of finding important information in large amounts of data is known as data mining or knowledge mining. It entails several techniques to guarantee that a huge amount of data is converted into meaningful information, including data translation, cleansing, integration, pattern analysis, and display [30].

2.4.1. WEKA tool

WEKA, is referred to Waikato Environment for Knowledge Analysis (WEKA), is a machine learning program developed by Waikato University in New Zealand. It is a Java-based tool that provides visualization tools and algorithms for predictive modeling and data analysis. It operates on all computing platforms and includes tasks like data mining, clustering, classification, association, visualization, and feature selection. The program's user-friendly interface and straightforward settings make it accessible to inexperienced users [31], [32]. Precision, recall, accuracy, F-measure, MCC, confusion matrix, and other data may be derived using the WEKA machine learning model to evaluate the result [33]. Figure 4 depicts the WEKA data mining tool's model.



RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) label

11:03:56 - lazy.KStar
11:13:30 - functions.LibSVM
11:24:47 - functions.Logistic
11:34:57 - trees.RandomForest

Status: OK

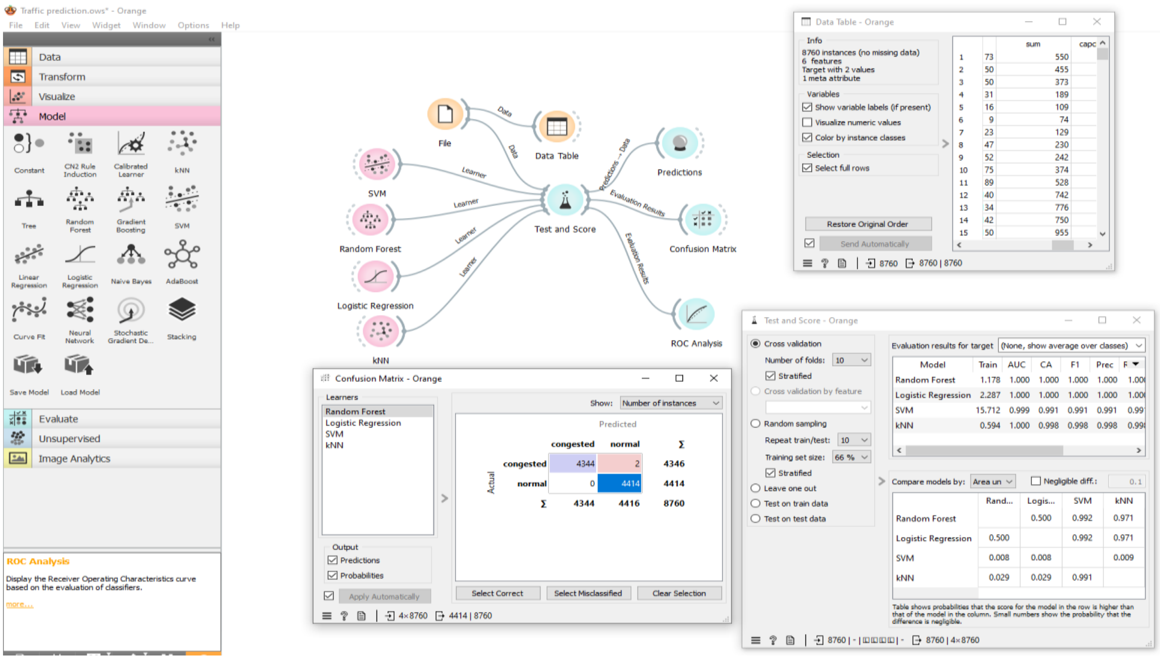
Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

```
Time taken to build model: 14.12

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

Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class =

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.994    0.003    0.998      0.994    0.996      0.992    0.996     0.995     nor
                 0.997    0.006    0.994      0.997    0.996      0.992    0.996     0.993     con
Weighted Avg.    0.996    0.004    0.996      0.996    0.996      0.992    0.996     0.994

=== Confusion Matrix ===

    a    b   <-- classified as
 4389   25 |    a = normal
   11 4335 |    b = congested
```

Figure 4. WEKA data mining tool's model

2.4.2. Orange tool

Orange is a set of machine learning, data mining, and Python scripting tools developed for interactive data analysis and component-based construction of data mining methods [34]. The bioinformatics laboratory at the University of Ljubljana has developed the visual data mining program Orange, available for free and non-commercial download, although primarily designed for instructional purposes, Orange can be beneficial for data processing and experimental data analysis, offering a platform for experiment selection [12], [35]. Figure 5 depicts the orange data mining tool's model.



ROC Analysis

Display the Receiver Operating Characteristics curve based on the evaluation of classifiers.

more...

Data Table - Orange

Info
8760 instances (no missing data)
6 features
Target with 2 values
1 meta attribute

Variables
☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

Restore Original Order
☑ Send Automatically

		sum	capc
1	73	550	
2	50	455	
3	50	373	
4	31	189	
5	16	109	
6	9	74	
7	23	129	
8	47	230	
9	52	242	
10	75	374	
11	89	528	
12	40	742	
13	34	776	
14	42	750	
15	50	955	

8760 | 8760 | 8760

Confusion Matrix - Orange

Learners
Random Forest
Logistic Regression
SVM
kNN

Show: Number of instances

Predicted

Actual	congested	normal	Σ
congested	4344	2	4346
normal	0	4414	4414
Σ	4344	4416	8760

Output
☑ Predictions
☑ Probabilities

☑ Apply Automatically

Select Correct | Select Misclassified | Clear Selection

4×8760 | 4414 | 8760

Test and Score - Orange

● Cross validation
Number of folds: 10
☑ Stratified
○ Cross validation by feature
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Evaluation results for target (None, show average over classes)

Model	Train	AUC	CA	F1	Prec	R
Random Forest	1.178	1.000	1.000	1.000	1.000	1.00
Logistic Regression	2.287	1.000	1.000	1.000	1.000	1.00
SVM	15.712	0.999	0.991	0.991	0.991	0.99
kNN	0.594	1.000	0.998	0.998	0.998	0.99

Compare models by: Area un ☐ Negligible diff.: 0.1

	Rand...	Logis...	SVM	kNN
Random Forest		0.500	0.992	0.971
Logistic Regression	0.500		0.992	0.971
SVM	0.008	0.008		0.009
kNN	0.029	0.029	0.991	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

8760 | - | 8760 | 4×8760

Figure 5. Orange data mining tool's model

3. PERFORMANCE EVALUATION AND CLASSIFICATION

For machine learning tasks like regression and classification, evaluation metrics are essential and helpful for a variety of tasks. While assuring accurate assessment, proper model evaluation using various measures can increase predictive parentage and power in addition to avoiding bad predictions when applied to unknown data. The objective of this procedure is to do a comparative analysis of each classifier and choose the most accurate one according to the obtained results [36], [37].

3.1. Cross-validation

Since it's easy to use and has a broad range of applications, cross-validation is a common model and tuning parameter selection technique in statistics and machine learning. Fitting and assessing any potential model on different data sets is necessary for ensuring accurate assessment. Models are often overfitted by using conventional techniques like V-fold and leave-one-out. According to preliminary theoretical research, cross-validation requires a training-testing split ratio of zero to reliably choose the right model under low-dimensional linear models. Most statistical software programs employ standard split ratios, such as four-to-one or nine-to-one, because smaller split ratios require larger training samples and lead to less precise model fitting [38]–[40]. In this research, 10-fold cross-validation is used.

3.2. Confusion matrix

The confusion matrix for the binary classification is written as a 2∗2 matrix. Four measurements have been published for a confusion matrix: "true positive" (TP), "true negative" (TN), "false positive" (FP), and "false negative" (FN). The confusion matrix is used to assess classifier performance on datasets in the multiclass problem. Matrixes to differentiate between the actual and expected values of the model's constituent parts in Java applications were classified into faulty and non-faulty classes using four confusion matrix measures: TP, FP, TN, and FN. The figure shows the confusion matrix of binary classification [41]–[44]. Figure 6 shows the confusion matrix of binary classification [45].

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

Figure 6. Confusion matrix of binary classification [45]

3.3. Classifiers performance

Datasets and classification algorithms are used to compare the WEKA and Orange data mining tools in this study. Evaluation criteria include accuracy, sensitivity, precision, and F-measure. By dividing the total number of properly categorized instances by the total value of instances, the accuracy measure is used to evaluate performance. Results are assessed utilizing datasets, tools, methods, separation, algorithms, and an overall total. All tests yielded an 100% categorization accuracy for the research [46]. Table 1 depicts the classifier's performance.

Table 1. Classifier’s performance [47]

Performance matrices	Equation
Accuracy	TP + TN/TP + FP + TN + FN
Sensitivity	TP/TP + FN
Precision	TP/TP + FP
F-measure	2(Sen ∗ Pre)/ (Sen + Pre)

4. RESULTS AND DISCUSSION

The dataset was classified using a variety of techniques, including SVM, KNN, LR, and RF. Based on this analysis, the orange tool provided superior results for accuracy (100%) for LR and RF; for KNN and SVM, the tool achieved CA with values of 99.8% and 99.1%, respectively. On the other hand, the results using the WEKA tool were also satisfactory. On the other hand, the results using the WEKA tool were also satisfactory, as SVM obtained a classification accuracy of 99.7%, while KNN, LR, and RF obtained (CA) of 98.7%, 97.6%, and 96.2%, respectively. According to these results, we can notice that the orange data mining tool is the most effective method for this data, according to the predictions made about traffic congestion. Table 2 depicts the classifier's performance using the WEKA vs. orange tool, while Figure 7 shows a comparative analysis of different classifiers' performance. On the other hand, it can also be said that the Orange3 model proposed in this work for predicting traffic congestion has outperformed previous studies reported in the literature, such as the study presented by Liu and Wu [8], where the accuracy reached 87.5% and it was 87.5% in Li *et al.* [9]. While the accuracy was in the work presented by Majumdar *et al.* [10], 84–95%.

Table 2. Classifier's performance using WEKA vs. orange tool

	RF		LR		SVM		KNN	
	WEKA	Orange	WEKA	Orange	WEKA	Orange	WEKA	Orange
Accuracy	0.962	1.000	0.976	1.000	0.997	0.991	0.987	0.998
Sensitivity	0.954	1.000	0.975	1.000	0.995	0.991	0.985	0.998
Precision	0.970	1.000	0.977	1.000	0.998	0.991	0.990	0.998
F-measure	0.962	1.000	0.976	1.000	0.997	0.991	0.988	0.998

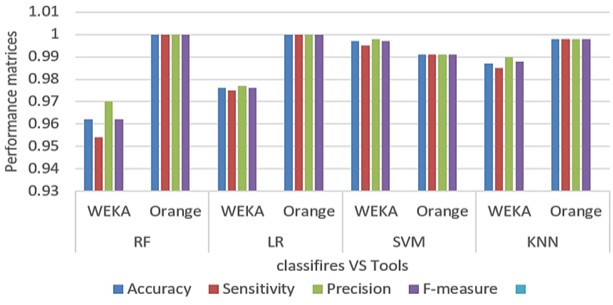


Figure 7. Comparative analysis of different classifiers' performances

5. CONCLUSION

The purpose of this study was to determine which data mining tool could provide the best prediction of the accuracy of traffic data. Comparative studies of the tools were conducted to see how successful different data mining techniques are and how different features affect traffic congestion prediction. The data was obtained from the Greater Amman Municipality, which mainly contained the traffic volume for the study area in the capital of the Kingdom of Jordan, Amman, for the year 2018. Various classification methods were used on the dataset, including RF, LR, KNN, and SVM. Cross-validation is used by 10-fold to improve the performance of the algorithms. Using the orange tool, RF, and LR have a higher grade of 100% based on this investigation, while KNN and SVM have scores of 99.8% and 99.1%, respectively. In contrast, using the WEKA Tool, SVM had a higher grade of 99.7% based on this investigation, while KNN, LR, and RF had scores of 98.7%, 97.6%, and 96.2, respectively. In the end, we can point out that the results of this comparison that we presented in our research paper indicate that the model built using the Orange3 data mining tool outperformed the model built using WEKA, as the accuracy in the first reached 100%.

REFERENCES

[1] T. S. Tamir *et al.*, "Traffic Congestion Prediction using Decision Tree, Logistic Regression and Neural Networks," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 512–517, 2020, doi: 10.1016/j.ifacol.2021.04.138.

[2] W. Zhuang and Y. Cao, "Short-Term Traffic Flow Prediction Based on a K-Nearest Neighbor and Bidirectional Long Short-Term Memory Model," *Applied Sciences (Switzerland)*, vol. 13, no. 4, 2023, doi: 10.3390/app13042681.

[3] Y. Xing, X. Ban, X. Liu, and Q. Shen, "Large-Scale Traffic Congestion Prediction Based on the Symmetric Extreme Learning Machine Cluster Fast Learning Method," *Symmetry*, vol. 11, no. 6, p. 730, May 2019, doi: 10.3390/sym11060730.

[4] T. Adetiloye and A. Awasthi, "Multimodal Big Data Fusion for Traffic Congestion Prediction," in *Multimodal Analytics for Next-Generation Big Data Technologies and Applications*, Cham: Springer International Publishing, 2019, pp. 319–335.

[5] M. W. Ei Leen, N. H. A. Jafry, N. M. Salleh, H. J. Hwang, and N. A. Jalil, "Mitigating Traffic Congestion in Smart and Sustainable Cities Using Machine Learning: A Review," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13957 LNCS, pp. 321–331, 2023, doi: 10.1007/978-3-031-36808-0_21.

[6] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," *Procedia Computer Science*, vol. 171, pp. 1251–1260, 2020, doi: 10.1016/j.procs.2020.04.133.

[7] S. M. Abdullah et al., " Optimizing Traffic Flow in Smart Cities: Soft GRU-Based Recurrent Neural Networks for Enhanced Congestion Prediction Using Deep Learning," *Sustainability*, vol. 15, no. 7, p. 5949, doi: 10.3390/su15075949

[8] Y. Liu and H. Wu, "Prediction of road traffic congestion based on random forest," in *Proceedings - 2017 10th International Symposium on Computational Intelligence and Design, ISCID 2017*, 2018, vol. 2, pp. 361–364, doi: 10.1109/ISCID.2017.216.

[9] Z. Li, P. Liu, C. Xu, H. Duan, and W. Wang, "Reinforcement Learning-Based Variable Speed Limit Control Strategy to Reduce Traffic Congestion at Freeway Recurrent Bottlenecks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3204–3217, 2017, doi: 10.1109/TITS.2017.2687620.

[10] S. Majumdar, M. M. Subhani, B. Roullier, A. Anjum, and R. Zhu, "Congestion prediction for smart sustainable cities using IoT and machine learning approaches," *Sustainable Cities and Society*, vol. 64, 2021, doi: 10.1016/j.scs.2020.102500.

[11] M. Zahid, Y. Chen, A. Jamal, and M. Q. Memon, "Short term traffic state prediction via hyperparameter optimization based classifiers," *Sensors (Switzerland)*, vol. 20, no. 3, 2020, doi: 10.3390/s20030685.

[12] R. Ratra and P. Gulia, "Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange)," *International Journal of Engineering Trends and Technology*, vol. 68, no. 8, pp. 30–35, Aug. 2020, doi: 10.14445/22315381/IJETT-V68I8P206S.

[13] "Greater Amman Municipality." Jordan, [Online]. Available: http://www.ammancity.gov.jo/en/gam/index.asp (Accessed Sep. 6, 2022).

[14] E. Scornet, G. Biau, and J. P. Vert, "Consistency of random forests," *Annals of Statistics*, vol. 43, no. 4, pp. 1716–1741, 2015, doi: 10.1214/15-AOS1321.

[15] N. Absar *et al.*, "The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction," *Healthcare (Switzerland)*, vol. 10, no. 6, 2022, doi: 10.3390/healthcare10061137.

[16] M. Z. Islam, J. Liu, J. Li, L. Liu, and W. Kang, "A semantics aware random forest for text classification," in *International Conference on Information and Knowledge Management, Proceedings*, 2019, pp. 1061–1070, doi: 10.1145/3357384.3357891.

[17] A. Chahal, P. Gulia, N. S. Gill, and J. M. Chatterjee, "Performance Analysis of an Optimized ANN Model to Predict the Stability of Smart Grid," *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/7319010.

[18] F. B. de Santana, W. Borges Neto, and R. J. Poppi, "Random forest as one-class classifier and infrared spectroscopy for food adulteration detection," *Food Chemistry*, vol. 293, pp. 323–332, 2019, doi: 10.1016/j.foodchem.2019.04.073.

[19] H. B. Ly, T. A. Nguyen, and B. T. Pham, "Estimation of Soil Cohesion Using Machine Learning Method: A Random Forest Approach," *Advances in Civil Engineering*, vol. 2021, 2021, doi: 10.1155/2021/8873993.

[20] A. Das, "Logistic Regression," in *Encyclopedia of Quality of Life and Well-Being Research*, Cham: Springer International Publishing, 2021, pp. 1–2.

[21] J. C. Stoltzfus, "Logistic regression: A brief primer," *Academic Emergency Medicine*, vol. 18, no. 10, pp. 1099–1104, 2011, doi: 10.1111/j.1553-2712.2011.01185.x.

[22] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," *Journal of Data Analysis and Information Processing*, vol. 07, no. 04, pp. 190–207, 2019, doi: 10.4236/jdaip.2019.74012.

[23] X. Zhang, C. Li, X. Wang, and H. Wu, "A novel fault diagnosis procedure based on improved symplectic geometry mode decomposition and optimized SVM," *Measurement: Journal of the International Measurement Confederation*, vol. 173, 2021, doi: 10.1016/j.measurement.2020.108644.

[24] B. M. Asl, S. K. Setarehdan, and M. Mohebbi, "Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal," *Artificial Intelligence in Medicine*, vol. 44, no. 1, pp. 51–64, 2008, doi: 10.1016/j.artmed.2008.04.007.

[25] J. Zhou, M. Xiao, Y. Niu, and G. Ji, "Rolling Bearing Fault Diagnosis Based on WGWOA-VMD-SVM," *Sensors*, vol. 22, no. 16, 2022, doi: 10.3390/s22166281.

[26] A. Rani, N. Kumar, J. Kumar, and N. K. Sinha, "Machine learning for soil moisture assessment," *Deep Learning for Sustainable Agriculture*, pp. 143–168, 2022, doi: 10.1016/B978-0-323-85214-2.00001-X.

[27] S. Zhang and J. Li, "KNN Classification With One-Step Computation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2711–2723, 2023, doi: 10.1109/TKDE.2021.3119140.

[28] J. Li, J. Zhang, J. Zhang, and S. Zhang, "Quantum KNN Classification With K Value Selection and Neighbor Selection," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023, doi: 10.1109/TCAD.2023.3345251.

[29] R. Mikut and M. Reischl, “Data mining tools,” *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 5, pp. 431–443, Sep. 2011, doi: 10.1002/widm.24.

[30] S. Verma and P. Rattan, “Introduction To Data Mining Tools and Techniques & Applications: a Review,” *in Business*, no. July, 2021, [Online]. Available: https://www.researchgate.net/profile/Rashmi-Gujrati-2/publication/355170587_Role_of_Technology_in_New_Decades/links/6163de531eb5da761e794894/Role-of-Technology-in-New-Decades.pdf#page=57.

[31] Sunita B Aher and LOBO L.M.R.J., “Data Mining in Educational System using WEKA,” 2011, [Online]. Available: http://www.ijcaonline.org/icett/number3/icett021.pdf.

[32] E. Kulkarni G. and R. Kulkarni B., “WEKA Powerful Tool in Data Mining,” *International Journal of Computer Applications National Seminar on Recent Trends in Data Mining*, vol. 5, no. Rtdm, pp. 975–8887, 2016, [Online]. Available: http://research.ijcaonline.org/rtdm2016/number2/rtdm2575.pdf.

[33] P. Debnath *et al.*, “Analysis of Earthquake Forecasting in India Using Supervised Machine Learning Classifiers,” *Sustainability*, vol. 13, no. 2, p. 971, Jan. 2021, doi: 10.3390/su13020971.

[34] J. Demšar *et al.*, “Orange: Data mining toolbox in python,” *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.

[35] Z. Dobesova, “Experiment in Finding Look-Alike European Cities Using Urban Atlas Data,” *ISPRS International Journal of Geo-Information*, vol. 9, no. 6, p. 406, Jun. 2020, doi: 10.3390/ijgi9060406.

[36] Ž. Vujović, “Classification Model Evaluation Metrics,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.

[37] M. R. Mahmood, M. B. Abdulrazzaq, S. R. M. Zeebaree, A. K. Ibrahim, R. R. Zebari, and H. I. Dino, “Classification techniques’ performance evaluation for facial expression recognition,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 1176–1184, 2020, doi: 10.11591/ijeecs.v21.i2.pp1176-1184.

[38] P. Zhang, “Model Selection Via Multifold Cross Validation,” *The Annals of Statistics*, vol. 21, no. 1, 2007, doi: 10.1214/aos/1176349027.

[39] J. Lei, “Cross-Validation With Confidence,” *Journal of the American Statistical Association*, vol. 115, no. 532, pp. 1978–1997, 2020, doi: 10.1080/01621459.2019.1672556.

[40] J. Shao, “Linear Model Selection by Cross-validation,” *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, Jun. 1993, doi: 10.1080/01621459.1993.10476299.

[41] R. Rajalakshmi and C. Aravindan, “A Naive Bayes approach for URL classification with supervised feature selection and rejection framework,” *Computational Intelligence*, vol. 34, no. 1, pp. 363–396, 2018, doi: 10.1111/coin.12158.

[42] J. Font, L. Arcega, Ø. Haugen, and C. Cetina, “Achieving feature location in families of models through the use of search-based software engineering,” *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 3, pp. 363–377, 2018, doi: 10.1109/TEVC.2017.2751100.

[43] Y. S. Taspinar, M. Koklu, and M. Altin, “Classification of flame extinction based on acoustic oscillations using artificial intelligence methods,” *Case Studies in Thermal Engineering*, vol. 28, 2021, doi: 10.1016/j.csite.2021.101561.

[44] A. Feyzioglu and Y. S. Taspinar, “Beef Quality Classification with Reduced E-Nose Data Features According to Beef Cut Types,” *Sensors*, vol. 23, no. 4, 2023, doi: 10.3390/s23042222.

[45] I. Markoulidakis, G. Kopsiaftis, I. Rallis, and I. Georgoulas, “Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem,” *ACM International Conference Proceeding Series*, pp. 412–419, 2021, doi: 10.1145/3453892.3461323.

[46] R. Panigrahi *et al.*, “Performance assessment of supervised classifiers for designing intrusion detection systems: A comprehensive review and recommendations for future research,” *Mathematics*, vol. 9, no. 6, 2021, doi: 10.3390/math9060690.

[47] F. Sajid *et al.*, “Secure and Efficient Data Storage Operations by Using Intelligent Classification Technique and RSA Algorithm in IoT-Based Cloud Computing,” *Scientific Programming*, vol. 2022, 2022, doi: 10.1155/2022/2195646.

BIOGRAPHIES OF AUTHORS

Areen Arabiat obtained B.Sc in computer engineering in 2005 from Al Balqa Applied University (BAU), and her MSc in Intelligent Transportation Systems (ITS) from Al Ahliyya Amman University (AAU) in 2022. She is currently a computer lab supervisor at the Faculty of Engineering/Al-Ahliyya Amman University (AAU) since 2013. Her research interests are focused on the following areas: machine learning, data mining, artificial intelligence, and image processing. She can be contacted email: a.arabiat@ammanu.edu.jo.

Muneera Altayeb obtained a bachelor’s degree in computer engineering in 2007, and a master’s degree in communications engineering from the University of Jordan in 2010. She has been working as a lecturer in the Department of Communications and Computer Engineering at Al-Ahliyya Amman University since 2015, in addition to her administrative experience as assistant dean of the Faculty of Engineering during the period (2020-2023). Her research interests focus on the following areas: digital signals and image processing, machine learning, robotics, and artificial intelligence. She can be contacted at email: m.altayeb@ammanu.edu.jo.