Journal of Computer Science

Original Research Paper

Comparative Performance of Supervised Learning Algorithms for Flood Prediction in Kemaman, Terengganu

Nur Najihah Shaaban, Norlida Hassan, Aida Mustapha and Salama A. Mostafa

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

Article history
Received: 06-10-2020
Revised: 12-12-2020
Accepted: 02-01-2021

Corresponding Author:
Norlida Hassan
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
Email: norlida@uthm.edu.my

Abstract: Flood is one of the most destructive phenomena all over the world. Because the flooding uncertainties and the urgency to prepare for disaster management, three specific technique approaches are compared in this study to predict the flood occurrence based on historical rainfall data. The study involved the rainfall data in Kemaman, Terengganu between 2017 and 2018 extracted from the official portal of the state of Terengganu. The dataset covers daily rainfall reading between January to December of the particular year in millimeter (mm) per day along with flood risks occurrence. This prediction experiment will be conducted using three variations algorithms, which are Decision Tree, Naive Bayes and Support Vector Machine. The comparison using three different algorithms was used to define the best algorithms that work with historical rainfall datasets to predict flood in terms of accuracy, precision, recall and F1-score. In the future, the prediction results are hoped to alert government authorities to make an early strategy to handle flood problems in Malaysia by analyzing the rainfall pattern.

Keywords: Rainfall Prediction, Flood, Supervised Learning, Machine Learning

Introduction

Malaysia consists of Sabah, Sarawak and Peninsular Malaysia. Throughout the year, Malaysia has rainy and hot temperatures. The maximum daily temperature is between 21 and 32°C in Malaysia. Usually, from November to March the north-east monsoon and from June to October the west monsoon are influenced by a strong equator. The annual precipitation is very high, which is about 2,300 mm in Sarawak and 3,300 mm in Sabah in Peninsular Malaysia (Amin *et al*., 2017). Rainfall is one of several important factors affecting watershed water quality (Liu *et al*., 2018) and other water-related problem (Razali *et al*., 2020; Masngut *et al*., 2020). Flood occurs due to heavy rainfall associated with severe thunderstorm within six hours. The flood after effect has caused significant damage and loss to the people in the affected area. During each event, local residents in the affected area have to the safe area to evacuate the place (Teng *et al*., 2017).

Terengganu is considered as one of the riskier place as the occurrence of flood is twice as much compared to other states in Peninsular Malaysia. This is because the geological location of Terengganu near to the Northeast Sea which is affected by the Monsoon Season every November to March. Monsoon flood is triggered by prolonged heavy widespread rain leads that to landslide (Piper *et al*., 2016). The east coast and southern part of Peninsular Malaysia, over the years, have been facing monsoonal flood and have been subjected to many research in predicting the rainfall to mitigate the flood risks (Ata, 2018; Loh *et al*., 2019).

Rainfall forecasting is a challenging issue in the world because of its practical value in popular science and meteorology. In the last decade, several major efforts have been recorded to solve weather forecasting problems using statistical modelling, including machine learning systems with successful results (Lee and Tuan Resdi, 2016; Kannan and Ghosh, 2011; Martínez, 2018; Sikorska and Seibert, 2018; Razali *et al*., 2020).

For an instance, (Lee and Tuan Resdi, 2016) developed an Artificial Neural Network (ANN) model to predict rainfall by using a Nonlinear Autoregressive Network with Exogenous input (NARX) algorithm. The NARX network was created trained and validated using 13 hydrological input datasets to generate four river stage outputs. The network was then measured by using square error, correlation coefficient, normal approximation error and judgment coefficient and performance criterion. The performance of the model was concluded with satisfactory results for concurrent hydrological simulations at several gauge stations, assuming there was no major land use adjustment in the catchment. The research also reported

that the effect of backwater tidal in the prediction model was found to be inconsistent, suggesting its minimal influence on the water level of the river in Kemaman. This is against the widespread claims that flooding in this region is a combined effect of big rain and high tidal water.

In another catchment, (Kannan and Ghosh, 2011) used different rain predictive precipitation data. They described a model based on *k*-Means clustering technique coupled with a supervised data classification technique, namely Classification And Regression Tree (CART) algorithm to generate rainfall states in a river basin from large atmospheric variables. The findings corroborated the likelihood of a rise in almost dry and high precipitation states and a decline in poor rainfall states. Note that the conversion frequency of medium to moderate precipitation is expected to increase in the future. The research concluded that due to climate change, the amount of regular precipitation in the Mahanadi basin will be severely affected due to upward trend in high rainfall levels that will cause flooding in the basin. This claim was supported by the number of comparisons in terms of occurring days at present time under various precipitation states and the predicted change in the river basin due to global warming.

CART, along with a Decision Tree (DT) algorithm C4.5, has also been used in designing a rule-based hourly rainfall prediction model by (Ji *et al*., 2012). The research calculated the chance of rain in order to correctly estimate the rainfall. Therefore, the hourly forecast of precipitation is only carried out if there is a chance of rain. Olaiya and Adeyemo (2012), the work explored the use of data mining techniques to forecast maximum temperature, rainfall, evaporation and wind speed. This was done using algorithms from the ANN and DT on meteorological data from the city of Ibadan, Nigeria, collected between 2000 and 2009. Sukanya and Prabha (2017) developed Support Vector Machine (SVM) and Backpropagation Neural Networks (BPNN). The BPNN algorithm was found to increase precipitation forecasting performance by evaluating historical and current facts to make reliable predictions.

Sikorska and Seibert (2018), a specific rainfall data were used in the alpine timber to precisely predict the events and the precipitation reliable and representative data required. In the study, the researchers investigated three values of precipitation datasets commonly used in hydrological studies. Datasets included station network precipitation, Interpolated Grid Precipitation (IGP) and Radar-Based Precipitation (RBP) for forecasting flood in an alpine catchment. With an improved explanation of the model structural errors, the researcher carried out a Bayesian study of instability to measure their effect on runoff simulations.

Razali *et al*. (2020) applied Bayesian Networks (BN), DT, KNN and SVM in prediction future flood in Kelantan. The study found out that single BN has the best performance in prediction the future flood for normal dataset. The study also found out SMOTE method are highly useful in combating with imbalance dataset. Naïve Bayes has also been used in (Zainudin *et al*., 2016) for a comparative analysis of various supervised learning techniques including Decision Tree (DT), Neural Network (NN) and Random Forest (RF) on their ability to predict rainfall data for multiple stations in Selangor, Malaysia. The aim of this study is to identify the best technique for predicting rainfall. Therefore, a comparative analysis was performed after the application of the three techniques in order to determine the most appropriate technique.

This project is set to predict flood based on historical rainfall data using a data mining approach. Three prediction algorithms will be used to model the flood risk patterns and compared to explore the best algorithm that works with the rainfall datasets, which are Decision Tree, Naive Bayes and Support Vector Machine. The performance were evaluated in terms of precision, recall and F1-score. The outcomes of this research will show the best algorithm for flood prediction. It is expected the findings in this research will serve as an early warning indication system for decision maker, Terengganu's government, engineer and hydrologist in the state to have better prediction towards the potential flood occurrence and assist in making proper decision regarding flood management.

The remaining of this study is organized as follows. Section 2 presents the research methodology used to perform the data mining task along with the dataset and the evaluation metrics. Section 3 presents the results and section 4 concludes the current paper with some direction for future work.

Materials and Methods

This research is set to study in depth the capability in predicting flood occurrences using three data mining algorithms, which are Decision Tree, Naive Bayes and Support Vector Machine. These techniques are applied to a case study in a catchment at Kemaman, Terengganu. To develop the flood occurrence prediction model, this research adopts the Cross Industry Standard Data Mining Process (CRISP-DM) methodology. This methodology is useful for making large data mining projects to be less costly, more reliable, more repeatable, more manageable and faster (Wirth and Hipp, 2000).

The CRISP-DM methodology consists of six stages as shown in Fig. 1. The arrows represent the most important dependencies between phases. The large outer circle indicates the iterative nature of this framework: Going back and forth between steps is often needed, as findings along the way trigger new questions (Shearer, 2000). This methodology begin with phase of literature review, data understanding, data preparation, modelling, evaluation and deployment.

Following Mrozek *et al.* (2015), the experiments were carried out using the Azure Machine Learning tool. The data were split into two sections, namely training and testing set, respectively. In this stage, the data were split into various percentages to study its performances. The collection of rows in each category is randomized based stratified sampling is used. In stratified sampling, a single column of data must be chosen for which values from the two test datasets are to be spread equally.

Dataset

In this research, comparative experiments will be evaluated using daily rainfall data in catchment area of Hulu Jabor, Kemaman Terangganu. The research focuses on predicting daily rainfall by using historical data from 2017 and 2018, obtained from Department of Drainage (JPS) Malaysia and the time series data for Hulu Jabor catchment in Kemaman are shown in Fig. 2. Based from this figure, it is observed that rainfall data in Hulu Jabor catchment area has clear trend of up and down indication its seasonal characteristics as heavy rainfall usually occurs in November until March for each year.

In order to enhance the performance of the prediction model, the dataset has to undergo data pre-processing task. In this task, data cleaning was carried out to clean missing values. Usually, missing values in rainfall data occurred for various reasons such as relocation, environmental changes, instruments malfunctioning and etc. (Nor *et al*., 2020). Data cleaning is very important as missing data may cause incorrect matches during prediction. Details of missing values in each feature are shown as the Table 1.

In this phase, the missing values were treated by applying the Cleaning Mode in Azure ML whereby the missing values are replaced with mean of the rainfall reading.

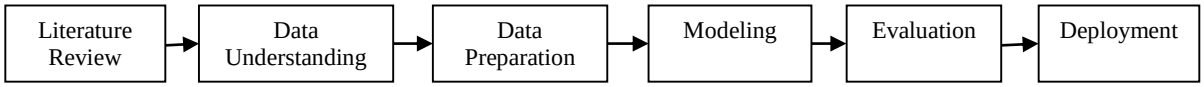


Fig. 1: Stages according to the CRISP-DM methodology

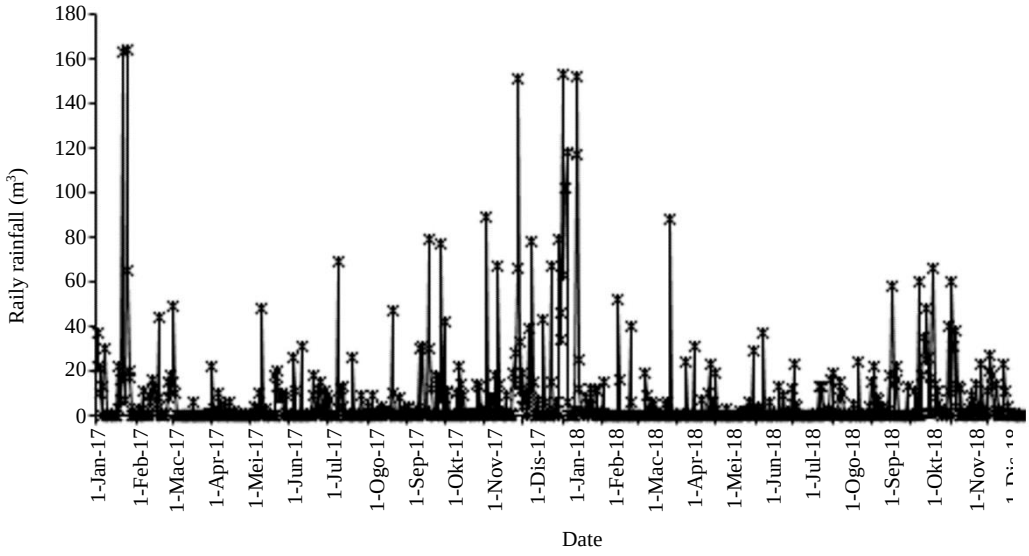


Fig. 2: Daily rainfall data in Hulu Jabor, Kemaman from Jan 1st 2017 to Dec 31st 2018

Table 1: Details of missing values

Feature	Missing value	Feature	Missing value
Day	0	July	0
January	0	August	0
February	3	September	0
March	0	October	0
April	1	November	1
May	0	December	0
June	1		

Decision Tree Algorithm

Decision Tree (DT) or also known as Classification and Regression Trees (CART) is one of popular ML techniques. The target class in classification trees is categorical type class while numerical type class for regression trees (Razali *et al*., 2020). It is shows that trees are capable to process both discrete and continuous data. DT is define as Eq. 1:

$$E(S)=\sum_{i=1}^{C}-p_i \log_2 p_i \quad (1)$$

where, P_i is the frequentist probability of an element/class i in the data. Figure 3 shows the graphical model.

Naïve Bayes Algorithm

Naïve Bayes is a basic probabilistic classifier that determines the probability of transmitting the frequency and combos of values from the data set. The algorithm uses the Bayes Theorem and assumes the value of the class variable for all the independent or non-international attributes (Patil, 2013). NB is based on a simplistic assumption that attribute values are free of charge if the value of the input is defined. In other words, the probability of mutual discovery, given the output value, is the sum of the individual probability.

NB also performs much better than predicted in most complex real-world scenarios (Pattekari and Parveen, 2012) as the equation is based on a later expectation, which incorporates previous knowledge and likelihood of an occurrence. The Bayes theorem states that Eq. 2 displays how the posterior probability can be calculated:

$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)} \quad (2)$$

where, P_i represents the set of parents of X_i in the networks. Figure 4 shows a graphical model of Naïve Bayes while its class implementation in Azure is available.

Support Vector Machines

A Support Vector Machine (SVM) is a supervised model of machine learning which uses classification algorithms for problems of classification of two classes. SVM are defined as:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0,1-Y_i\left(\vec{w}\cdot\vec{x_i}-b\right)\right]+\lambda\left\|\vec{w}\right\|2 \quad (3)$$

where the parameter λ defines a trade-off between increasing the margin size and ensuring x is located in the right margin side. Therefore, for relatively small values, of λ, the second term in the loss function will become trivial and it will behave similar to the hard- margin SVM if the input data is linearly classifiable, but will also determine whether or not a classification rule is viable or not. Figure 5 shows the graphical model for SVM.

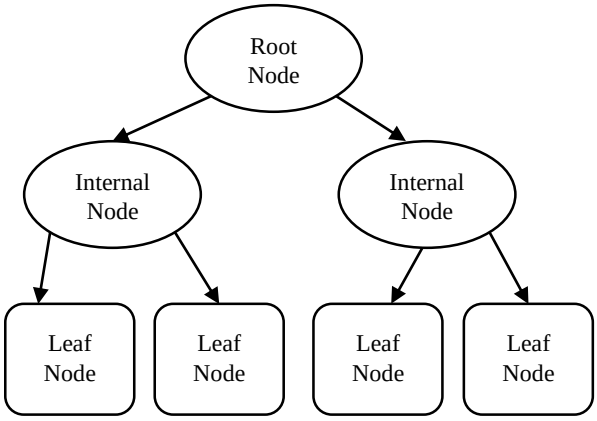


Fig. 3: Graphical model of decision tree

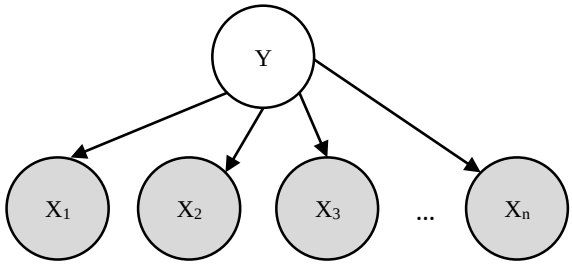


Fig. 4: Graphical Model of Naïve Bayes

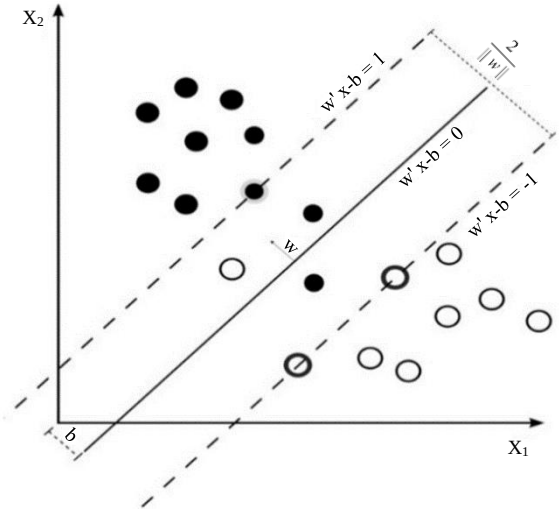


Fig. 5: Graphical model of support vector machine

Note that all the algorithms used are based on the modules in the Azure ML tool, which are the two-class boosted decision tree, two-class Naive Bayes and two-class support vector machine with all standard parameter.

Evaluation Metrics

The evaluation metrics used in the experiments are accuracy, precision, recall and F1-score. Note that TP is True Positive, FP is False Positive, TN is True Negative and FN is False Negative. Accuracy is the total number of samples properly classified to the total number of classified samples. Eq. 4 shows the formula for calculating accuracy:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

Precision is positively classified number of samples correctly divided by total samples as positive samples. Equation shows the the formula for calculating precision:

$$Precision = \frac{TP}{(TP + FN)} \quad (5)$$

Recall is the number of samples marked as positive divided by the maximum sample in the positive classification set. The formula is shown in Eq. 6:

$$Recall = \frac{TP}{(TP + FN)} \quad (6)$$

F1-score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. The formula for calculating F1-score is shown in Eq. 7:

$$F1 - Socre = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \quad (7)$$

Results

This experiment has its objective which is to compare the performance of Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms in predicting the flood occurrences based on rainfall data.

Table 2: Experimental results for Decision Tree (DT), Naive Bayes (NB) and Support Vector Machine (SVM) algorithms based on various data splitting

Algorithm	Score bin	Accuracy	F1-Score	Precision	Recall
Decision trees	(90,10)	0.737	0.737	0.778	0.700
	(80,20)	0.579	0.636	0.583	0.700
	(70,30)	0.579	0.667	0.571	0.800
	(60,40)	0.579	0.667	0.571	0.800
	(50,50)	0.632	0.720	0.600	0.900
	(40,60)	0.579	0.692	0.563	0.900
	(30,70)	0.526	0.667	0.529	0.900
	(20,80)	0.474	0.643	0.500	0.900
	(10,90)	0.526	0.690	0.526	1.000
Naïve Bayes	(90,10)	0.526	0.182	1.000	0.100
	(80,20)	0.632	0.462	1.000	0.300
	(70,30)	0.474	0.500	0.500	0.500
	(60,40)	0.474	0.583	0.500	0.700
	(50,50)	0.474	0.643	0.526	0.900
	(40,60)	0.526	0.690	0.526	1.000
	(30,70)	0.526	0.690	0.526	1.000
	(20,80)	0.526	0.690	0.526	1.000
	(10,90)	0.526	0.690	0.526	1.000
Support vector machines	(90,10)	0.526	0.182	1.000	0.100
	(80,20)	0.579	0.500	0.667	0.400
	(70,30)	0.579	0.636	0.583	0.700
	(60,40)	0.526	0.690	0.526	1.000
	(50,50)	0.526	0.690	0.526	1.000
	(40,60)	0.526	0.690	0.526	1.000
	(30,70)	0.526	0.690	0.526	1.000
	(20,80)	0.526	0.690	0.526	1.000
	(10,90)	0.526	0.690	0.526	1.000

Table 3: Performance comparison between Decision Tree (DT), Naive Bayes (NB) and Support Vector Machine (SVM)

Algorithm	Accuracy	F1-score	Precision	Recall
DT	0.737	0.737	0.778	0.700
NB	0.632	0.462	1.000	0.300
SVM	0.579	0.636	0.583	0.700

The tests were carried out using Azure Machine Learning, a suite of Machine Learning applications that involves various techniques. Note that training and training adopted the Sampling method in the Azure ML tool. Table 2 shows the experimental results for all algorithms using different split.

Based on the experimental results, it can be noted that the best results for DT was obtained from data splitting of 90-10, which the accuracy and F1-scores of 73.7%, precision score of 77.8% while recall of 70%. Furthermore, the table also shows the performance comparisons for NB algorithm. The best results were obtained when the data were split into 80% training and 20% testing. At this point, the accuracy obtained is 63.2% and precision of 1.0.

SVM produced the best outcomes when the training of the experiment was split at 70% of instruction and 30% of research. The result of the accuracy obtained at this point is 57.9%, which is the same accuracy with the experiment split at 80% of instruction and 20% of research, but the result of F1 score, precision and recall at 70-30% is higher compared to 80-20% split during this experiment.

Discussion

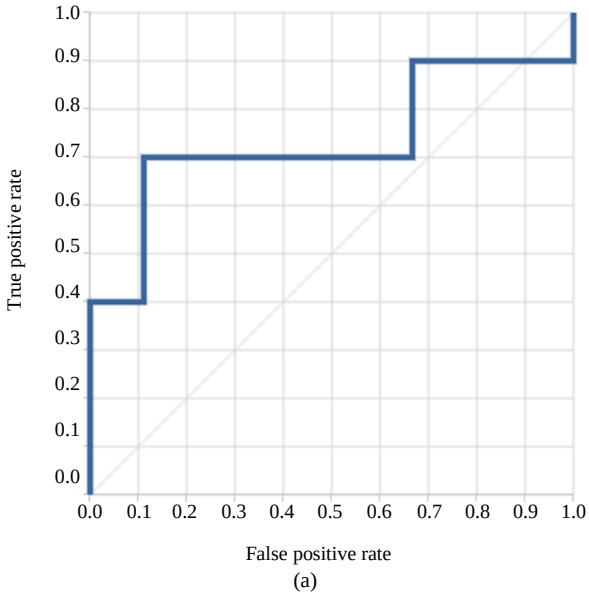
Table 3 shows the performance comparison between decision Tree (DT), Naive Bayes (NB) and Support Vector Machine (SVM). From the results, it can concluded that DT has the most accurate classifier for predict the rainfall with the accuracy achieved is 73.7%. DT is one of the simplest ways of defining the most significant variables and the relationship between two or more variables. It can build new variables with the aid of decision trees that have improved power to predict target variable. That is why DT got the highest accuracy compared to NB and SVM.

For the second classifier, which is NB classifier, the results suggested that attributes have separate distributions, so that they are not prone to sensitive characteristics. NBs models still use the maximum likelihood approach, which is why it only took a limited amount of predictive training data during the experiment. For this experiment, NB classifier is easy to use because the amount of data rainfall is 744 which is (31 days ×12 months ×2 years).The best accuracy outcomes for NB is 63.2%. SVM classifiers is more suitable on a large amount of data this is why the accuracy for the experiment is the lowest which is 57.9% when the training of the experiment was split at 70% of instruction and 30% of research.

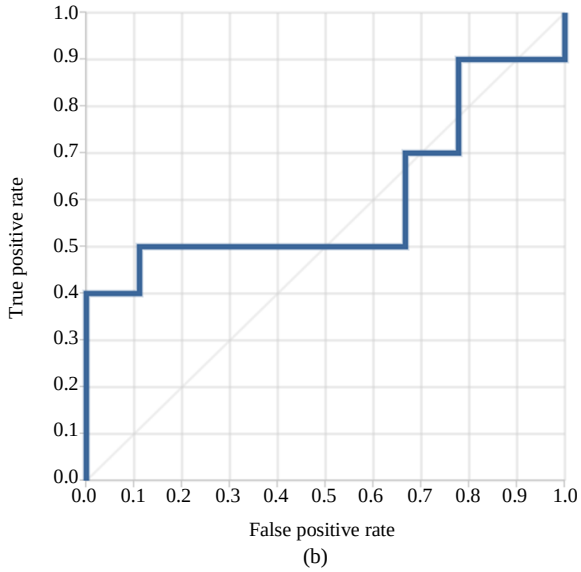
Comparison of these three flood occurrence prediction models, DT are the top performers and continue to NB and the last one is SVM. This is confirmed by the fact that DT and NB are good when trained on a low amount of training data and these two models were able to predict the higher portion of testing data with the top accuracy results as compared with SVM that were supposedly trained on a large amount of training data and predict the higher portion of test data. This is why SVM scored lower accuracy compared to the two other algorithms.

Next, the flood prediction results were further analyzed using the Receiver Operating Characteristics (ROC) Curve. A ROC curve sums up the balance between the TP and FP ratios of a prediction model with various probability thereshold. It has two dimensions, the x-axis indicating a FAP rate and the y-axis refers to the True Positive (TP) rate. The x-axis is a two-dimensional one. Figure 6 shows the Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM) ROC curve. The ROC for DT is 0.600 showing 7 out of 10 result shows a positive result in accuracy.

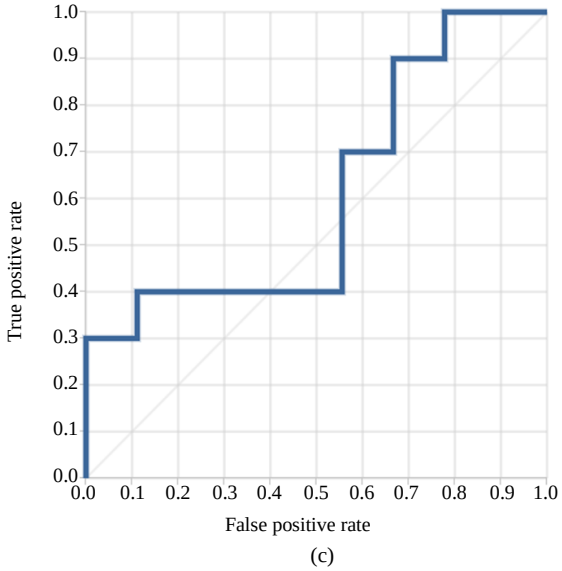
Furthermore, the ROC AUC for NB is 0.611. 5 out of 10 results indicate a good outcome for this algorithm by using the stratified sampling. Meanwhile, the ROC curve for SVM shows that 10 out of 10 result are positive but have the lowest percent of accuracy compare to others. The ROC AUC is 0.611.



(a)



(b)



(c)

Fig. 6: (a-c) ROC curve for Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM)

Conclusion

This research proposed an investigation of predicting flood occurrences based on rainfall data specific in Hulu Jabor, Kemaman, Terengganu. The experimental findings revealed that DT work well for prediction of flood because of their ability to train on small data and predict the larger portion of data with higher accuracy. One major challenge with rainfall data is the nature of imbalanced dataset whereby flood risks based on heavy rainfall reading only occurs in certain months in a particular country. The next step is to perform the experiment and compare the prediction results using these three algorithms approach against other prediction algorithms from the literature. The prediction model from the case study is hoped to be generalized into a more comprehensive model to cover different rainfall data across the world.

Acknowledgement

This research is supported by the Research Management Center, Universiti Tun Hussein Onn Malaysia under Vot E15501.

Author's Contributions

Nur Najihah Shaaban: Shaaban performed the study, experiments and drafted the manuscript under the planning and supervision.

Norlida Hassan and Aida Mustapha: Prepared the manuscript.

Salama A. Mostafa: Helped the experiments and interpreted the results.

Ethics

All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

References

Amin, M. Z. M., Shaaban, A. J., Ercan, A., Ishida, K., Kavvas, M. L., Chen, Z. Q., & Jang, S. (2017). Future climate change impact assessment of watershed scale hydrologic processes in Peninsular Malaysia by a regional climate model coupled with a physically-based hydrology modelo. Science of the Total Environment, 575, 12-22. https://doi.org/10.1016/j.scitotenv.2016.10.009

Ata, F. M., (2018). Impact of Hydrological Study to Water Quality Status in Kuantan River, Pahang, Malaysia. International Journal of Engineering & Technology, pp. 35-43. https://doi.org/10.14419/ijet.v7i3.14.16859

Ji, S. Y., Sharma, S., Yu, B., & Jeong, D. H. (2012, August). Designing a rule-based hourly rainfall prediction model. In 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI) (pp. 303-308). IEEE. https://doi.org/10.1109/IRI.2012.6303024

Kannan, S., & Ghosh, S. (2011). Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. Stochastic Environmental Research and Risk Assessment, 25(4), 457-474. https://doi.org/10.1007/s00477-010-0415-y

Lee, W. K., & Tuan Resdi, T. A. (2016). Simultaneous hydrological prediction at multiple gauging stations using the NARX network for Kemaman catchment, Terengganu, Malaysia. Hydrological Sciences Journal, 61(16), 2930-2945. https://doi.org/10.1080/02626667.2016.1174333

Liu, J., Shen, Z., & Chen, L. (2018). Assessing how spatial variations of land use pattern affect water quality across a typical urbanized watershed in Beijing, China. Landscape and Urban Planning, 176, 51-63. https://doi.org/10.1016/j.landurbplan.2018.04.006

Loh, E. C., Ismail, S. B., & Khamis, A. (2019). Empirical Mode Decomposition Couple with Artificial Neural Network for Water Level Prediction. https://doi.org/10.13189/cea.2019.071403

Martínez, V., (2018). State of the Art of Hydrological Models and Suitability in Basins with Limited Information. 22nd International Congress on Project Management and Engineering, pp. 4-013.

Masngut, M. A., Ismail, S., Mustapha, A., & Yasin, S. M. (2020). Comparison of daily rainfall forecasting using multilayer perceptron neural network model. IAES International Journal of Artificial Intelligence, 9(3), 456. https://doi.org/10.11591/ijai.v9.i3.pp456-463

Mrozek, D., Gosk, P., & Małysiak-Mrozek, B. (2015). Scaling Ab into predictions of 3D protein structures in Microsoft Azure cloud. Journal of Grid Computing, 13(4), 561-585. https://doi.org/10.1007/s10723-015-9353-8

Nor, S. M. C. M., Shaharudin, S. M., Ismail, S., Zainuddin, N. H., & Tan, M. L. (2020). A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia. Bulletin of Electrical Engineering and Informatics, 9(2), 635-643. https://doi.org/10.11591/eei.v9i2.2090

Olaiya, F., & Adeyemo, A. B. (2012). Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1), 51. https://doi.org/10.5815/ijieeb.2012.01.07

Patil, T. R. (2013). MSSS Performance analysis of naive bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications, 6(2).

Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), 290-294.

Piper, D., Kunz, M., Ehmele, F., Mohr, S., Mühr, B., Kron, A., & Daniell, J. (2016). Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany-Part 1: Meteorological background. Natural Hazards and Earth System Sciences, 16(12), 2835. https://doi.org/10.5194/nhess-16-2835-2016

Razali, N., Ismail, S., & Mustapha, A. (2020). Machine learning approach for flood risks prediction. IAES International Journal of Artificial Intelligence, 9(1), 73. https://doi.org/10.11591/ijai.v9.i1.pp73-80

Sikorska, A. E., & Seibert, J. (2018). Value of different precipitation data for flood prediction in an alpine catchment: A Bayesian approach. Journal of Hydrology, 556, 961-971. https://doi.org/10.1016/j.jhydrol.2016.06.031

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22.

Sukanya, R., & Prabha, K. (2017). Comparative Analysis for Prediction of Rainfall using Data Mining Techniques with Artificial Neural Network. Volume-5, Issue-6, Page, (288-292).

Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F., Dutta, D., & Kim, S. (2017). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. Environmental Modelling & Software, 90, 201-216. https://doi.org/10.1016/j.envsoft.2017.01.006

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.

Zainudin, S., Jasim, D. S., & Bakar, A. A. (2016). Comparative analysis of data mining techniques for Malaysian rainfall prediction. International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1148-1153. https://doi.org/10.18517/ijaseit.6.6.1487