A comparative analysis of classification techniques on predicting flood risk

Nazri Mohd Nawi[1], Mokhairi Makhtar[2], Mohd Zaki Salikon[3], Zehan Afizah Afip[4]
[1,3,4]Soft Computing and Data Mining Centre (SMC), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia
[2]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Malaysia

Article Info

Article history:

Received Oct 3, 2019
Revised Dec 5, 2019
Accepted Dec 19, 2019

Keywords:

Classification
Comparative analysis
Flood risk
Predictive

ABSTRACT

Flood is a temporary overflow of a dry area due to overflow of excess water, runoff surface waters or undermining of shoreline. In Malaysia itself in 2014, the country grieved with the catastrophic flood event in Kuala Krai, Kelantan, which caused of human lives, public assets and money lost. Due to uncertainties in flooding event, it is vital for Malaysia to have pre-warning system that assist related agencies in to categorize land areas that face high risk of flood so preventive actions can be planned in place. This paper conducts a comparative analysis of three classifications in classifying the risk of flood, whether high or low. The classification experiment conducts using three variants of Bayesian approaches, which are Bayesian Networks (BN), Naive Bayes (NB), and Tree Augmented Naive Bayes (TAN). The outcome of this research shows that Tree Augmented Naive Bayes (TAN) has the best algorithms as compared to others algorithms in classifying the risk of flood.

Corresponding Author:

Nazri Mohd Nawi,
Soft Computing and Data Mining Centre (SMC),
Faculty of Computer Science and information Technology,
Universiti Tun Hussein Onn Malaysia (UTHM),
Email: nazri@uthm.edu.my

1. INTRODUCTION

Malaysia is a country comprising Peninsular Malaysia, Sabah, and Sarawak. It covers fourteen states that are Perlis, Kedah, Penang, Perak, Selangor, Negeri Sembilan, Pahang, Melaka, Johor, Kelantan, Terengganu, Sabah, and Sarawak. Asian nation additionally has one central consisting of three Territories that are Federal Territory of Malaysian capital, Federal Territory of Labuan and Federal Territory of Putrajaya. Malaysia has two main areas separated by the South China Sea. The northern border is Thailand and the southern border is Singapore. Meanwhile, the border of Indonesia on the south and Brunei on the north. Malaysia is located near the equatorial line at the Latitude 1˚ and North 7˚ and 100˚ and 100˚ East Malaysia covers 329,960.22 km [1]. Malaysia has hot and humid weather throughout the year. The average daily temperature throughout Malaysia is between 21˚C to 32˚C. Typically, the Malaysian climate is experiencing a strong equator influenced by the north eastern monsoon from November to March and the western monsoon from June to October. The annual rainfall is very high which is 2500mm in Peninsular, Malaysia between 2300mm in Sarawak and 3300 mm in Sabah [2].

Due to the high rainfall and river flow, the risk of flood in Malaysia is very high. Flood can be defined as a situation where water flows exceed the carrying capacity of a river resulting in overflows over the river banks [3]. There are several factors that can cause flood such a sudden rise in water levels such as continuous rainfall, land humidity and non-smooth water drainage. The other one contributing factors is the uncontrollable rapid development. Moreover, widespread land clearing and overcutting trees also can causes water absorption to land to decline and runoff continues to the river more rapidly. For every increase of development rate between 0-40 percent, it will result in a flow rate of 190 percent, hence twice the runoff speed.

In addition, to make it worst, the rate of erosion will increase resulting in increased silt in the river. Shallow river will have a lower capacity, unable to accommodate the increased water and cause the water to flood the cliffs. Not to ignore the river basin, which can also cause flood. The size of a large river basin will have a large run of water when heavy rain. If the river capacity is insufficient, floods will occur. Based on those described factors, Malaysia was shocked by the news of the catastrophic natural disaster that flooded Kelantan especially at Kuala Krai, Kelantan back in 2014 [4].

There is a need to have systematic techniques that can predict flood based on those causes which can avoid from any catastrophic disaster to human population [5-6]. Recent research had shown that data mining techniques had gained its popularity and had been used in predicting flood [7-11]. The main goal of this research is to compare the performance of some data mining techniques in classifying the risk of flood.

The remaining paper is organized as follows: Section 2 gives literature review of some data mining techniques. Section 3 described the research method. Results and dicussions are explained in Section 4. Finally, the paper is concluded in Section 5.

2. LITERATURE REVIEW

Studies on flood prediction using Bayesian approaches has been very active especially in the recent movement on awareness of climate change as well as its ability to deal with uncertainties. Most recently in 2018, researcher [12] proposed a dynamic flood assessment and discovered that urban underground facilities tend to be prone to flood due to breaking of a dam or a barrier, or a flash flood when exceptional degree of rain occurred. Rapid and dynamic assessment of underground flood evolution method vital for safety evacuation and reduce disaster. Research had shown that Bayesian Networks could improve rapidly and dynamically access the flood evolution method in underground areas. In the networks, 17 nods represented the flood disaster drivers, flood disaster bearers, flood mitigation action, and additionally on the spot feedback data. The results showed that the projected framework significantly helpful by dynamically evaluating underground flood method and to spot the crucial influencing factors.

On the other hand for dealing with dynamic systems Rashid and Othman [13] focused on flood data to test three different dynamic algorithms with different tools, which were Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Dynamic Evolving Spiking Neural Network (deSNN). The proposed algorithm, deSNN, achieved best accuracy rate in predicting flood when fed with Spatio/Spectro Temporal Data Modeling (SSTD). Since SSTD data is not supported with existing data mining tool such as WEKA therefore the analysis of the data were based on the analysis of space and time. As a comparative experiment, conventional machine learning methods such as MLP and SVM are used as a baseline performance and accuracy measures.

In an alpine catchment, Sikorska et. al [14] used different precipitation data for flood prediction to accurately predict such events, accurate and representative precipitation data required. In the study, three value of precipitation datasets commonly used in hydrological studies were investigated. The datasets include station network precipitation (SNP), interpolated grid precipitation (IGP), and radar-based precipitation (RBP). They performed a Bayesian uncertainty analysis with an improved description of model systematic errors to quantify their effects on runoff simulations.

Monthly precipitation forecast by Sharma A, Goyal [15] used Bayesian approach technique for monthly mean precipitation prediction at twenty-one stations in Assam, India. The interstation precipitation dependencies and independencies are delineating mistreatment Bayesian Networks (BN) structure and five atmospherically variables including temperature, relative humidity, wind speed, overcast, and southern oscillation index were used as predictors. The research aimed to match between two different structural learning rules in Bayesian Networks, which were K2 and Markov Chain Monte Carlo (MCMC) algorithm. 13 different models are developed with different combinations from 5 predictors. At the end of this experiments, K2 algorithms outperformed MCMC algorithms for all combinations.

According to Martina [16] rainfall, thresholds are primarily based on flood warning. Therefore, it is important to derive the likelihood of providing flood warnings at given water course sections based on comparison of quantitative precipitation forecast with important precipitation threshold values although this was not necessarily the requirement of real time statement system. The proposed resulted in an especially simplified alert system employed by non-technical stakeholders and may be used additionally to support the normal flood statement system just in case of system failures.

Recently, data mining techniques have been widely used by researchers in many applications in classifying and predicting the effect based on some cause and factors. In addition, it is vital for Malaysia to have such powerful data mining techniques in assisting related agencies in to categorize land areas that face high risk of flood so preventive actions can be planned in place.

Therefore, this research conduct a comparative analysis of some data mining techniques which includes Bayesian approaches in flood risk prediction using the data from Kuala Krai, Kelantan. The dataset contains six attributes, which are water level, rainfall daily, rainfall monthly, wind, humidity, and temperature. The data are classified into two classes, which are low risk and high risk of flood. This research compares the performance between three variations algorithms that are a Bayesian Network, Naive Bayes and Tree Augmented Naive Bayes for flood prediction.

3. RESEARCH METHOD

All data mining techniques selected for this paper used the Cross Industry Standard Process for Data Mining (CRISP-DM) approach as shown in Figure 1. All experiments are simulated using the WEKA tool [17] with 10-fold validation method for training and testing. Furthermore, cross-validation method with k-fold is used because it can reduce computational time while maintaining accurate estimates.

In order to test the flood risk prediction all algorithms were tested using the flood data in Kuala Krai, Kelantan. The trend of the flood dataset obtained from Malaysia Department of Meteorology where it recorded from 1st January to 31th December between 2012 and 2016 extracted from [18-19]. The dataset consists of 1,828 instances and each is described by rainfall monthly (RF Month), rainfall daily (RF Daily), water level (in cm), humidity, wind and temperature. The features correspond to predict a binary class of flood and no flood. Figure 2 shows the excerpt of the dataset.

In this paper, the flood data have imbalance class, which is the amount of training data between the two classes different. One of its classes represents very large amount of data (majority class) while other classes represent very small amount of data (minority class). There are different kind of methods that can be used to treat imbalanced datasets, which include method sampling and oversampling, random oversampling and SMOTE [20].

Three algorithms were used for flood risk prediction: (a) Bayesian Networks [21], (b) Naive Bayes [22], and (c) Tree Augmented Naive Bayes [23] with oversampling technique called SMOTE. Oversampling data were needed by considering the imbalanced nature of flood risk classes between flood and no flood. Bayesian Networks is a probabilistic-based data modelling method that represents variable and conditional interdependencies through a DAG (Directed Acyclic Graph). By applying Markov Chain-Rule, the joint probability distribution of the nodes in Bayesian Network can be decomposed as shown in (1).

$$P_B\left(X_{1,\ \ldots,}X_n\right) = \prod_{i-1}^{n} P(X_i\,|Pa_i) \qquad (1)$$

where Pa_i represents the set of parents of X_i in the networks. Figure 3 shows a graphical model of Bayesian Networks.

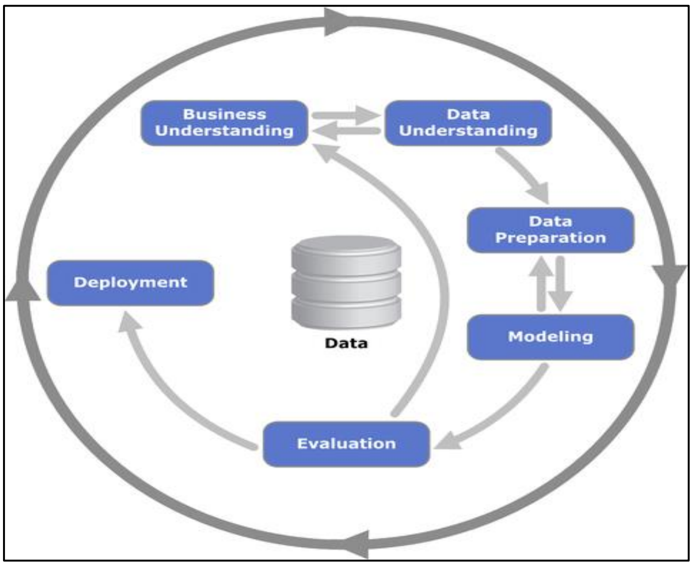
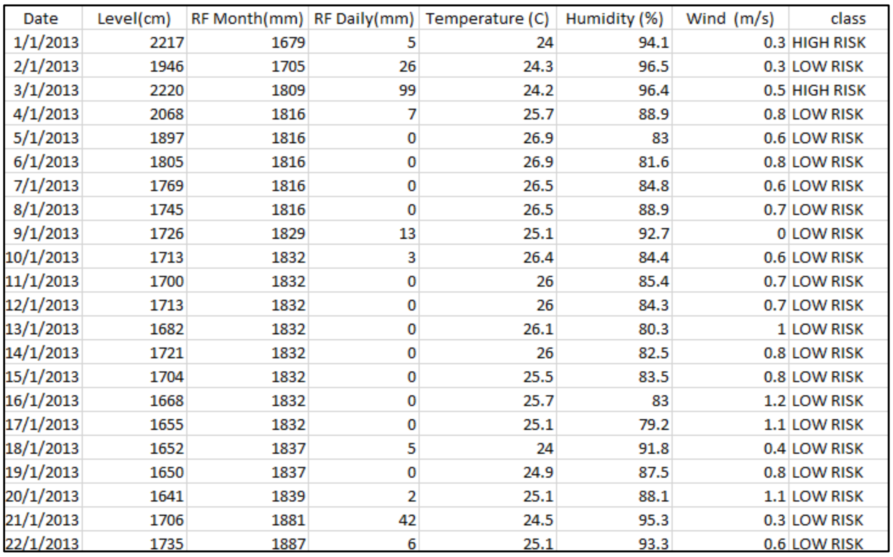


Figure 1. CRISP-DM process model for data mining



Date	Level(cm)	RF Month(mm)	RF Daily(mm)	Temperature (C)	Humidity (%)	Wind (m/s)	class
1/1/2013	2217	1679	5	24	94.1	0.3	HIGH RISK
2/1/2013	1946	1705	26	24.3	96.5	0.3	LOW RISK
3/1/2013	2220	1809	99	24.2	96.4	0.5	HIGH RISK
4/1/2013	2068	1816	7	25.7	88.9	0.8	LOW RISK
5/1/2013	1897	1816	0	26.9	83	0.6	LOW RISK
6/1/2013	1805	1816	0	26.9	81.6	0.8	LOW RISK
7/1/2013	1769	1816	0	26.5	84.8	0.6	LOW RISK
8/1/2013	1745	1816	0	26.5	88.9	0.7	LOW RISK
9/1/2013	1726	1829	13	25.1	92.7	0	LOW RISK
10/1/2013	1713	1832	3	26.4	84.4	0.6	LOW RISK
11/1/2013	1700	1832	0	26	85.4	0.7	LOW RISK
12/1/2013	1713	1832	0	26	84.3	0.7	LOW RISK
13/1/2013	1682	1832	0	26.1	80.3	1	LOW RISK
14/1/2013	1721	1832	0	26	82.5	0.8	LOW RISK
15/1/2013	1704	1832	0	25.5	83.5	0.8	LOW RISK
16/1/2013	1668	1832	0	25.7	83	1.2	LOW RISK
17/1/2013	1655	1832	0	25.1	79.2	1.1	LOW RISK
18/1/2013	1652	1837	5	24	91.8	0.4	LOW RISK
19/1/2013	1650	1837	0	24.9	87.5	0.8	LOW RISK
20/1/2013	1641	1839	2	25.1	88.1	1.1	LOW RISK
21/1/2013	1706	1881	42	24.5	95.3	0.3	LOW RISK
22/1/2013	1735	1887	6	25.1	93.3	0.6	LOW RISK

Figure 2. Data flood in microsoft excel

Naive Bayes could be a straightforward probabilistic classifier that calculates a collection of chances by forward the frequency and combos of values from the given datasets. The algorithm uses the Bayes theorem and assumes all the independent or non-interdependent attributes given by the value of the class variable [24]. Naive Bayes is based on a simplified assumption that attribute values are conditional on each other free of charge if given output value. In other words, given the output value, the probability of collectively observing is the product of the individual probability [24]. Naive Bayes often works much better in most complex real-world situations than expected [24] because the algorithm is based on posterior probability that combines previous experience and likelihood of event. According to the Bayes theorem, as shown in (2) shows on how to calculate posterior probability,

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)} \tag{2}$$

where $P(c|x)$ is the posterior probability of class (target) given predict (attribute), $P(c)$ is the prior probability of class, $P(x|c)$ is the likelihood which is probability of predictor given class and $P(x)$ is the prior probability of predictor. Figure 4 shows a graphical model of Naïve Bayes.

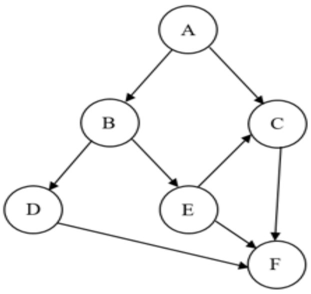


Figure 3. Graphical model of bayesian networks

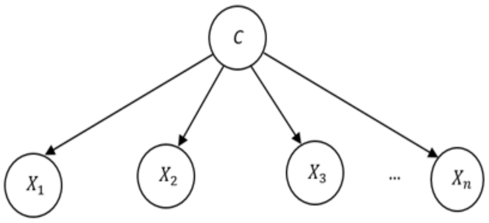


Figure 4. Graphical Model of Naive Bayes

Tree Augmented Naive Bayes (TAN) is related to Naive Bayes classifier because it is a continuation of the Naive Bayes classifier. Naive Bayes classifier is obtained by learning D training data by determining the probability of each attribute X_i when given the class C variable. This is because Naive Bayes does not realistic to be applied to real data, so there is a Naive Bayes fix called Augmented Naive Bayes. In developing Augmented Naive Bayes classifier is equivalents as finding a good Bayesian Network with class C variable as root [24]. Because of intensive computing, an efficient solution to finding the Bayesian

Network is the ability to influence each other between variables. Figure 5 shows a graphical model of Tree Augmented Naive Bayes.

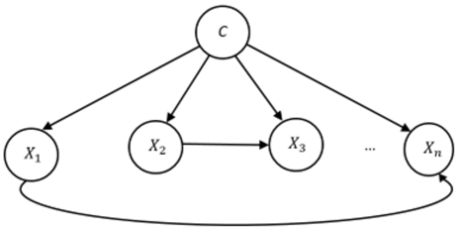


Figure 5. Graphical model of tree augmented naive bayes

The evaluation metrics used in this paper are accuracy, precision, recall, and *f*-measure.

a) Accuracy. Accuracy is total number of samples correctly classified to the total number of samples classified. The formula for calculating accuracy is shown in (3), where TP is True Positive, TN is True Negative and FN is False Negative.

$$\text{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+TN+FP+FN})} \tag{3}$$

b) Precision. Precision the number of samples is categorized positively classed correctly divided by total samples are classified as positive samples. The formula for calculating precision is shown in (4), where TP is True Positive and FP is False Positive.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \tag{4}$$

c) Recall. Recall is the number of samples is classified as positive divided by the total sample in the testing set positive category. The formula for calculating recall is shown in (5), where TP is True Positive and FN is False Negative.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})} \tag{5}$$

d) *f*-Measure. F-Measure 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 (6).

$$f - \text{Measure} = \frac{2*(\text{Recall}*\text{Precision})}{(\text{Recall+Precision})} \tag{6}$$

4. RESULTS AND ANALYSIS

The purpose of this section is to demonstrate the performance of Naive Bayes (NB), Tree Augmented Naïve Bayes (TAN) and Bayesian Networks (BN) algorithms with oversampling technique (SMOTE) and without the oversampling technique (Normal). In these experiments, the WEKA tools have been used to get the results.

In an oversampling process such as using the SMOTE technique, the first step is to determine the number of its nearest neighbors which is five. This is based on the consideration that the value of the attribute on synthetic data formed from the nearest neighbor is five. The nearest number set to five neighbors is also frequently used in experimental methods that apply SMOTE such as by [25]. As a comparison in the performing tests, sampling methods used will include random oversampling in WEKA, known as resample. This experiment evaluated training models by 10-fold cross validation technique. That means, applying the algorithm 10 times, each time 9 of the folds are used for training and 1 fold is used for testing. Table 1 shows the results in terms of accuracy with oversampling technique (SMOTE) and without oversampling (Normal).

Based on Table 1, TAN algorithm is the most efficient classifier with accuracy 100% for classifying flood risk datasets. It shows that TAN algorithm is more stable and robust when dealing with over sampling and noise datasets. Where SMOTE really represents the real cases of real datasets with noise and un-clean data. However, without oversampling, BN algorithm has been the best accuracy with 100%. Figure 6 shows the bar graph comparison between three Bayesian variants in terms of accuracy.

In Table 2, once again, TAN algorithm performs better for oversampling SMOTE. It shows that TAN algorithm achieves the higher precision with 1.0% and clearly indicates that the algorithm is more robust to noise and it more stable. However, without oversampling technique, BN algorithm demonstrates the higher precision with 1.0%. The comparison between three Bayesian variants in terms of precision is shown in Figure 7.

Table 1. The Comparison Results in Terms of Accuracy

Algorithm	Accuracy %	
	SMOTE	Normal
Naive Bayes (NB)	98.290	97.920
Tree Augmented Naive Bayes (TAN)	100.000	99.450
Bayesian Networks (BN)	99.880	100.000

Table 2. The Performance Results in Terms of Precision

Algorithm	Precision %	
	SMOTE	Normal
Naive Bayes (NB)	0.984	0.990
Tree Augmented Naive Bayes (TAN)	1.000	0.999
Bayesian Networks (BN)	0.999	1.000

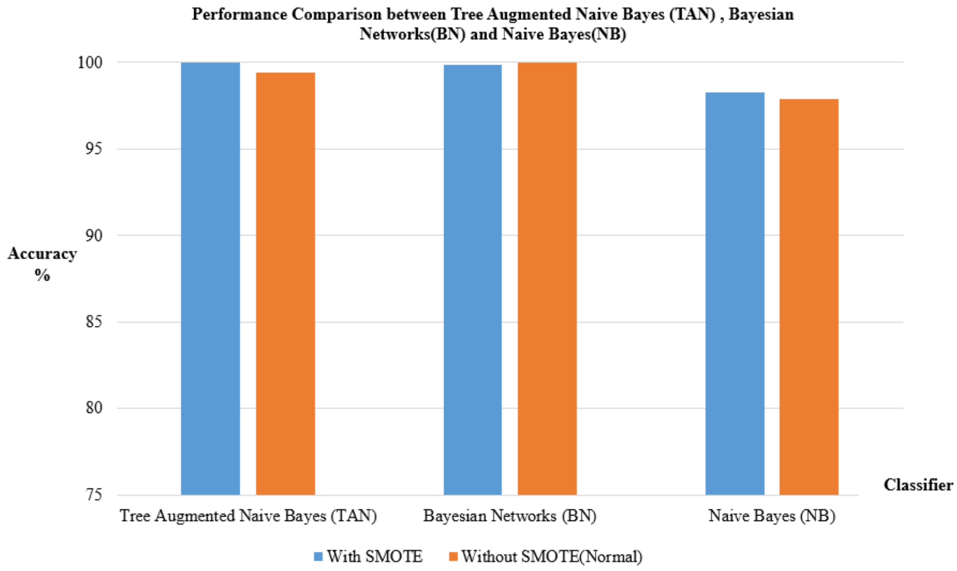


Figure 6. The comparison between three Bayesian variants in terms of accuracy

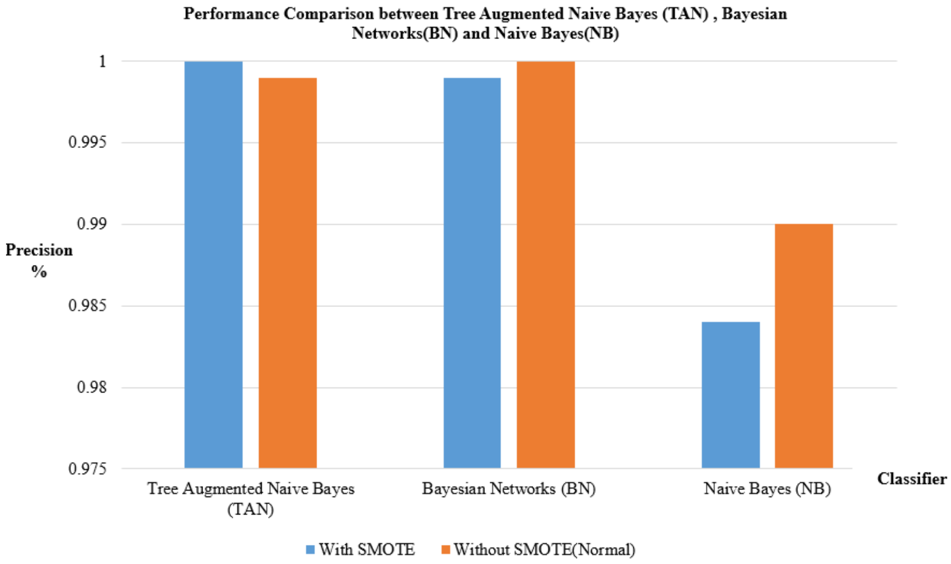


Figure 7. The comparison between three Bayesian variants in terms of precision

Based on Table 3, it is no doubt that for SMOTE oversampling, TAN algorithm still produces higher recall of 1.0%. Whereas, without oversampling, BN algorithm still maintaian the higher recall of 1.0%. Figure 8 shows the bar graph comparison between three Bayesian variants in terms of recall. Table 4 shows that SMOTE oversampling with TAN algorithm has the best f-measure of 1.0%. Meanwhile, without oversampling, BN algorithm has the best f-measure with 1.0%. Figure 9 shows the bar graph comparison between three Bayesian variants in terms of *f*-measure.

Table 3. The Performance Results in Terms of Recall

Algorithm	Recall %	
	SMOTE	Normal
Naive Bayes (NB)	0.983	0.979
Tree Augmented Naive Bayes (TAN)	1.000	0.999
Bayesian Networks (BN)	0.999	1.000

Table 4. The Performance Results in Terms of *f*-measure

Algorithm	*f*-measure %	
	SMOTE	Normal
Naive Bayes (NB)	0.983	0.983
Tree Augmented Naive Bayes (TAN)	1.000	0.999
Bayesian Networks (BN)	0.999	1.000

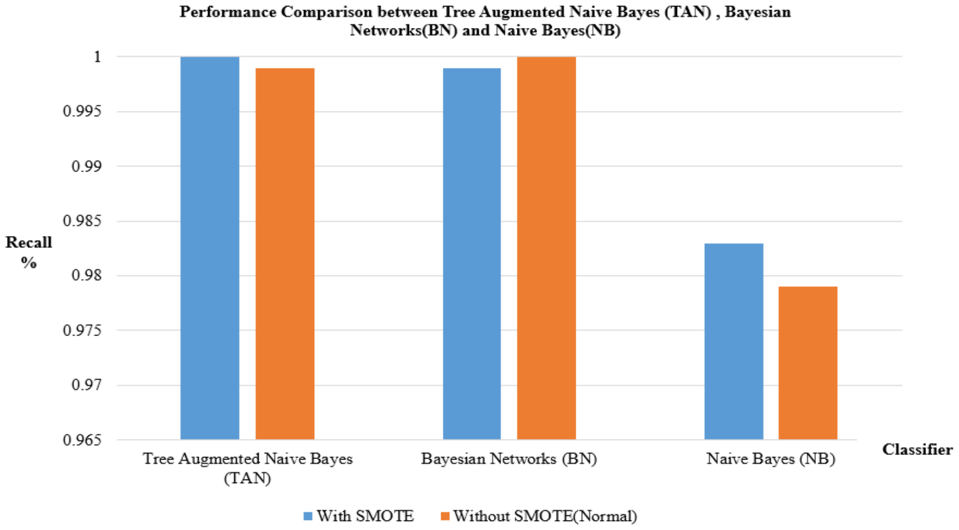


Figure 8. Bar graph comparison between three Bayesian variants in terms of recall

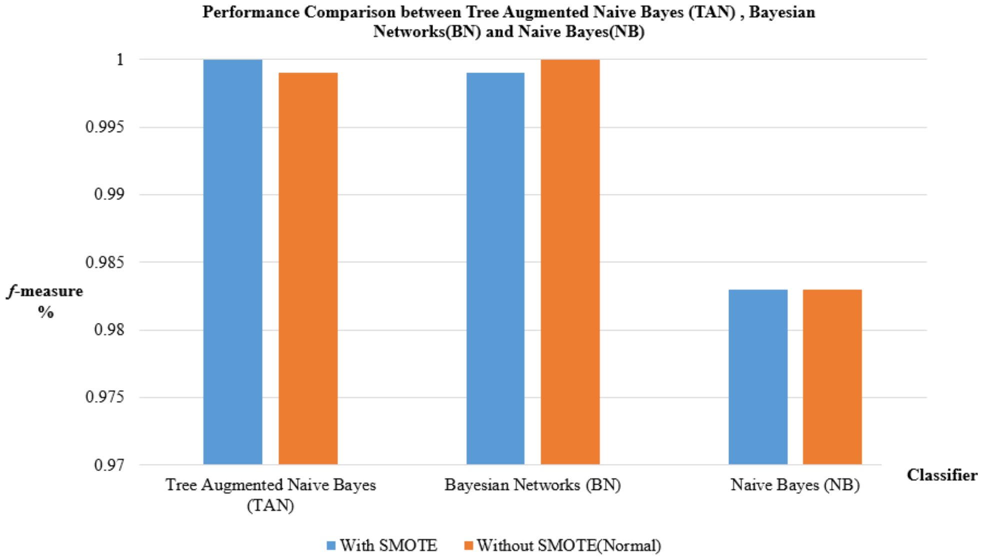


Figure 9. The comparison between three Bayesian variants in terms of *f*-measure

Overall, all prediction model of flood risks perform better with oversampling such as using he SMOTE algorithm in exception of BNs because BN algorithm has better generalization capabilities even when dealing with imbalanced classes as compared to variations of naïve Bayes algorithm such as the NB and TAN.

5. CONCLUSION

In conclusion, this paper presented a comparative analysis between Bayesian approaches to predict flood based on Kuala Krai, Kelantan. Moreover, this paper also explored the used of Synthetic Minority Oversampling (SMOTE) to treat the imbalanced nature of the flood dataset. The used of SMOTE can enable researchers to handle the problem imbalance of the flood dataset with its performance value results. The overall simulation results by treating imbalanced using Synthetic Minority Oversampling (SMOTE) has shown that Tree Augmented Naive Bayes (TAN) performed the best as compared to others algorithms. This is because to the fact that, combining all the datasets resulted in larger training set as a result that the model may well be trained well. This research paper currently only focused on imbalanced dataset. Therefore, for future work, this research proposes to use dynamic Bayesian network to treat the flood dataset as time series data, which can further explain the results in better way.

ACKNOWLEDGEMENTS

The authors would like to express the deepest appreciation to the Ministry of Education Malaysia for financially supporting this Research under Trans-displinary Research Grant Scheme (TRGS-Vote number T003). This work is partially supported by UniSZA (Grant No. R0008). Additional supports in terms of facilities were also provided by Centre Soft Computing and Data Mining Centre (SMC), FSKTM, Universiti Tun Hussein Onn Malaysia (UTHM).

REFERENCES

[1] Jabatan Penerangan Malaysia, Geografi, 2018. Available online: http://pmr.penerangan.gov.my/index.php/profil-malaysia/4-geograf-i.html

[2] Tan BC. , Seratus Negara Asia Tenggara 1, Prisma Sdn. Bhd. 1995.

[3] Goh KC , Geografi Fizikal. Longman, Kuala Lumpur. 1981

[4] Hussin WNTW, Zakaria NH, Ahmad MA,” Knowledge Sharing and Lesson Learned from Flood Disaster: A case in Kelantan,” *Journal of Information System Research and Innovation*. 2015.

[5] C. Li, Y. Liu, J. Yang and Z. Gao , "Prediction of Flooding Velocity in Packed Towers Using Least Squares Support Vector Machine,". 2012.

[6] M. D. Mauro and K. d. Bruijn, "Application and validation of mortality functions to assess the consequences of flooding to people," *Journal of Flood Risk Management*, vol. 5, no. 2, pp. 92-110. 2012.

[7] Banik, S., Anwer, M., Khodadad Khan, A.F.M., Rouf, R.A., Chanchary, F.H. ”Forecasting Bangladeshi monsoon rainfall using neural network and genetic algorithm approaches”. *International Technology Management Review*, 2(1): 2009.

[8] Kannan, M., Prabhakaran, S., Ramachandran, P, “Rainfall Forecasting Using Data Mining Technique”. *International Journal of Engineering and Technology*, 2(6), 397–401: 2010.

[9] Nayak, D., Mahapatra, A., Mishra, P, “A Survey on Rainfall Prediction using Artificial Neural Network”. *International Journal of Computer Applications*, 72; (16): 2013.

[10] Mehdi Ramezanifard, B. S. Mousavi ,” *Digital image classification by optimised fuzzy system*”, *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, Vol. 14, No. 3, p.p: 1196-1202: 2019.

[11] Jesmeen M. Z. H, J. Hossen, S. Sayeed, CK Ho, Tawsif K, Armanur Rahman, E.M.H. Arif, “A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics*”, Indonesian journal of Electrical Engineering and Computer Science (IJEECS),*Vol. 10, No. 3, pp1234-1243: 2019.

[12] Wu J, Fang W, Hu, Z, Hong B, “Application of Bayesian Approach to Dynamic Assessment of Flood in Urban Underground Spaces. Water,” 10(9), 1112: 2018.

[13] Rashid NAMA, Othman M, “Predicting Flood Risk Using Spiking Neural Network: A Framework. Dissertation,” Faculty Computer Science and Information Technology, University Tun Hussein Onn Malaysia. 2017.

[14] Sikorska AE, Seibert J, “Value of different precipitation data for flood prediction in an alpine catchment: A Bayesian approach”. *Journal of Hydrology*. 2016.

[15] Sharma A, Goyal MK, “Bayesian network for monthly rainfall forecast: a comparison of K2 and MCMC algorithm,” *International Journal of Computers and Applications*, 38(4), 199-206: 2016.

[16] Martina MLV, Todini E, Libralon A, “A Bayesian decision approach to rainfall thresholds based flood warning,” *Hydrology and Earth System Sciences Discussions*, 2(6), 2663-2706: 2005.

[17] Singhal S, Jena M, “A study on WEKA tool for data preprocessing, classification and clustering,” *International Journal of Innovative Technology and Exploring Engineering*, 2(6), 250-253: 2013.

[18] Info Banjir Portal, available online: http://infobanjirwater gov.my, 2016. 22/2/2027.

[19] "Meteorologi Portal," available online: http://www.met.gov.my/, 2016. 26/11/2014.
[20] Chawla, Bowyer, Hall, and Kegelmeyer. "SMOTE: Synthetic Minority". 2002.
[21] Chawla N. V, Bowyer K. W, Hall L. O, Kegelmeyer W. P, "SMOTE: Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research*, 16: 321-357. 2002.
[22] Duc Truong Pham, Gonzalo A. Ruz "*Networks for Data Clustering*." In Proceedings of the Royal Society A-Mathematical Physical and Engineering Sciences, 465, 2927-2948: 2009.
[23] Patil TR, Sherekar MS, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *International Journal of Computer Science and Applications*, 6(2), 256-261. 2013.
[24] Friedman N, "Bayesian Network Classifier". Machine Learning, 29, 131–161: 1997.
[25] Machado EL, Ladeira, "Dealing with Rare Cases and Avoiding Overfitting: Combining Cluster Based Oversampling and SMOTE," Department of Computer Science. Brazil. 2007

BIOGRAPHIES OF AUTHORS

Nazri Mohd Nawi received his B.S. degree in Computer Science from University of Science Malaysia (USM), Penang, Malaysia. His M.Sc. degree in computer science was received from University of Technology Malaysia (UTM), Skudai, Johor, Malaysia. He received his Ph.D. degree in Mechanical Engineering department, Swansea University, Wales Swansea. He is currently a lecturer in Software Engineering Department at Universiti Tun Hussein Onn Malaysia (UTHM). His research interests are in optimization, data mining techniques and neural networks.

Mokhairi Makhtar is an Associate Professor of Computing from Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, He received his Ph.D. in Computer Science from the University of Bradford in 2012. His research interests include Machine Learning, Data Mining and big data analytics for toxicology, education, health and business applications.

Mohd Zaki Salikon currently is currently a lecturer in Software Engineering Department at Universiti Tun Hussein Onn Malaysia (UTHM). He received his bachelor degree in computer science from Universiti Teknlogi Malaysia (UTM) Malaysia. He did his master degree in in computer science from Universiti Utara Malaysia (UUM). His research areas are soft computing, scheduling, and data base system

Zehan Afizah Afip currently is currently a lecturer in Software Engineering Department at Universiti Tun Hussein Onn Malaysia (UTHM). He received his bachelor degree in computer science from Universiti Teknlogi Malaysia (UTM) Malaysia. He did his master degree in in computer science from Universiti Utara Malaysia (UUM). His research areas are soft computing, scheduling, and data base system