Comparative Analysis for Flood Prediction: A Data Mining Approach

Muhamad Dinnie Ismail, Mohd Zaki Mohd Salikon,
Aida Mustapha, Salama A Mostafa

Introduction

This presents the comparative analysis for flood prediction by using Data Mining approach; Neural Networks (NN), Random Forest (RF), and Decision Trees (DT).

The methodology using in this research is the Knowledge Discovery in Database KDD methodology.

The results showed that Neural Network (NN) has the best results in accuracy with average 98.9%, followed by Decision Tree (DT) with 97.8%, and Random Forest at 95.6% average accuracy

Background

- Flood is a temporary dry land flood due to excess water runoff, ripple surface waters or shoreline undermining.
- Malaysia protested in 2014 about the public assets and a total devastating flood occurrence in Kuala Krai, Kelantan, which lost human livesloss of RM 2 billion.
- Physical factors such as geological setting and topography are crucial in the analysis of the cause and effect of severe flood happened in the central Kelantan Basin on December 2014.

Background

- It was found that prolonged heavy daily rainfalls on the upstream catchments caused the failure of storage systems to cope with the exceptional rainfall event.
- There are several factors more that can cause flood such a sudden rise in water levels due to continuous rainfall, land humidity and non-smooth water drainage.
- This can be attributed to the uncontrollable rapid development that involves widespread land clearing and overcutting trees, hence declining land water absorption and the runoff continues to the river.
- The literature has shown extensive studies on flood phenomenon especially using the Kuala Krai incident as the focal study point.
- The cumulative rainfall amount and the location of the heavy rainfall center are sensitive to the choices of Cumulus Parameterisation Scheme (CPS) and suggested that the CPS gave larger impact to the forecast quality compared to boundary conditions.

Related Works

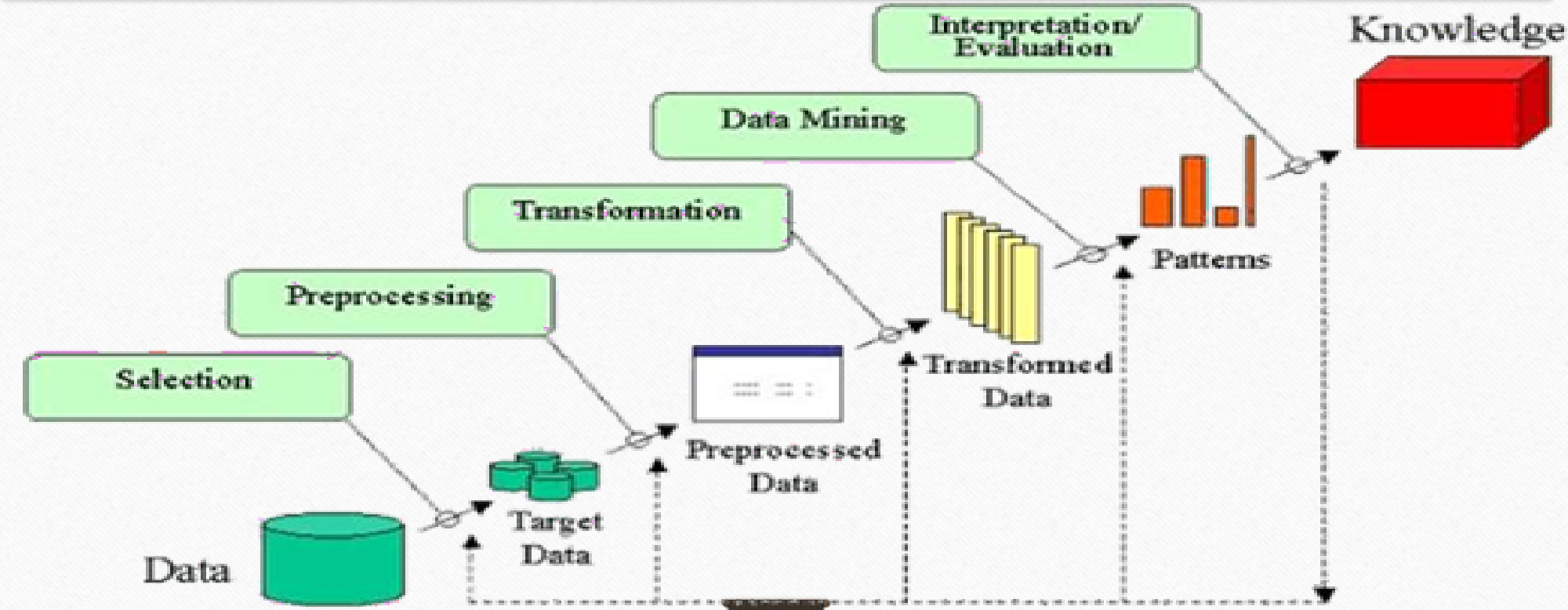
- The potential evapo-transpiration estimation methods for water balance analysis Using SWAT (M.K. Husain et al., 2019)
- B. Winarta et al. (2019) investigates a calibration and confirmation method of hydrologic model using HEC-HMS (Hydrologic Engineering Center-Hydrologic Modeling System) applied in Lebir River.
- The relationship between climate change and flooding in Kelantan River, Lebir River and Galas River has been explored. The research determined which period of months has the highest frequency of peak rainfall volume, water level and stream flow. It was observed that the rainfall volume, water level and stream flow are interrelated. When rainfall volume increases, the water level and stream flow will also increase and vice versa. (Khai et al., 2019)

Objectives

- To comparative the three technique on flood risk dataset.
- To apply three algorithm data mining using Neural Networks, Random Forest, and Decision Tree.
- To evaluate the performance comparative between the three techniques on the flood risk dataset

Methodology

KDD Model



Flood risk data of Kuala Krai, Kelantan Dataset

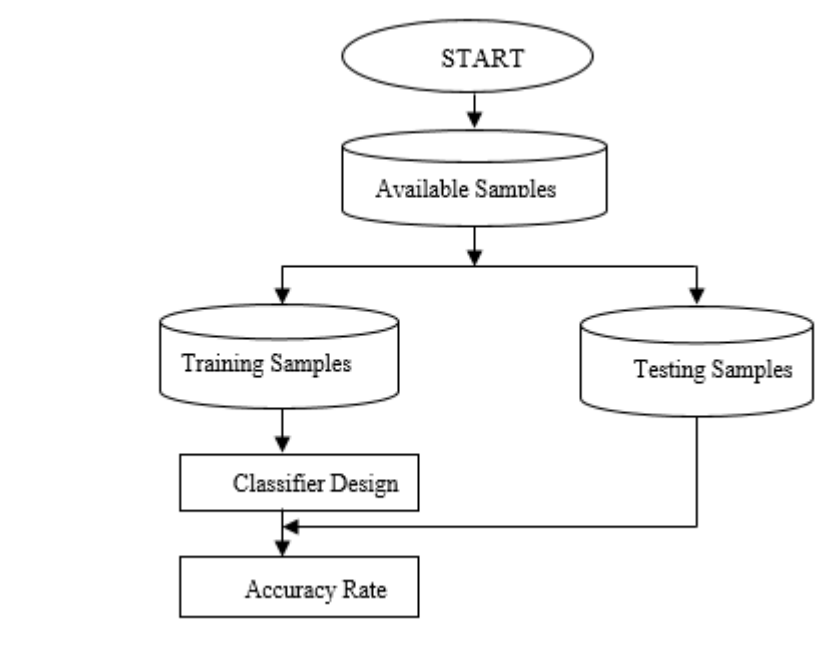
Date	Level(cm)	RF Month(mm)	RF Daily(mm)	Temperature (?C)	Humidity (%)	Wind (m/s)	class
01-01-14	1683	2794	0	24.8	89.3	0.4	NOFLOOD
02-01-14	1684	2797	3	24.9	88.5	0.4	NOFLOOD
03-01-14	1753	2846	49	25.2	89.4	0.5	NOFLOOD
04-01-14	1726	2847	1	25.6	87.5	0.4	NOFLOOD
05-01-14	1691	2849	2	25	88.8	0.8	NOFLOOD
06-01-14	1680	2854	5	25.9	85.3	1	NOFLOOD
07-01-14	1697	2854	0	26.2	80.9	1	NOFLOOD
08-01-14	1663	2854	0	25	87.4	0.5	NOFLOOD
09-01-14	1650	2859	5	24.8	90.1	0.6	NOFLOOD
10-01-14	1904	2881	19	24.5	90.7	1.3	NOFLOOD
11-01-14	2249	2968	87	23.3	93.7	1	FLOOD
12-01-14	2218	3006	32	24.9	88.8	1	FLOOD
13-01-14	1897	3008	2	25	86.3	0.7	NOFLOOD
14-01-14	1808	3008	0	24.6	85.6	0.5	NOFLOOD
15-01-14	1764	3008	0	24.3	83	0.9	NOFLOOD
16-01-14	1747	3008	0	24.1	82.4	0.8	NOFLOOD
17-01-14	1719	3008	0	24.6	81.3	0.8	NOFLOOD
18-01-14	1702	3008	0	24.3	80.9	1	NOFLOOD
19-01-14	1691	3008	0	23.3	81.7	0.9	NOFLOOD
20-01-14	1672	3008	0	24.2	80.3	0.9	NOFLOOD
21-01-14	1674	3008	0	23.2	79.9	0.7	NOFLOOD
22-01-14	1671	3008	0	22.9	80.8	0.7	NOFLOOD

Preprocessing

- Since each feature of the dataset is interconnected to each other, an instance with a missing value will affect the whole record. Thus, if there is any missing value was found, the entire row will be removed.

Transformation

The flood risk dataset is divided into training and testing sets based on validation method where nine parts is for training the algorithm and the last one part for assessing the algorithm.



Classification Algorithms

- A Neural Networks processes information using a connectionist approach to computation, where it uses an interconnected group of artificial neurons that changes its structure during a learning phase and produce a classification or prediction.
- Random Forests consist of many individual decision trees that operate as an ensemble classifier. Classifier assembles results from several classical decision trees on various sub training samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- Decision Trees refer to a hierarchical model of variables and their relationships in a form of tree and uses traversal strategies to achieve to the predicted class. Classification trees are used to classify an object or an instance such as insurant to predefined set of classes based on their attribute or feature values

Evaluation Metrics

- **Accuracy.** Accuracy is total number of samples correctly classified to the total number of samples classified. The formula for calculating accuracy is shown in Equation 1.

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

- **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 Equation 2.

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{2}$$

- **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 Equation 3.

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{3}$$

- **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 Equation 4.

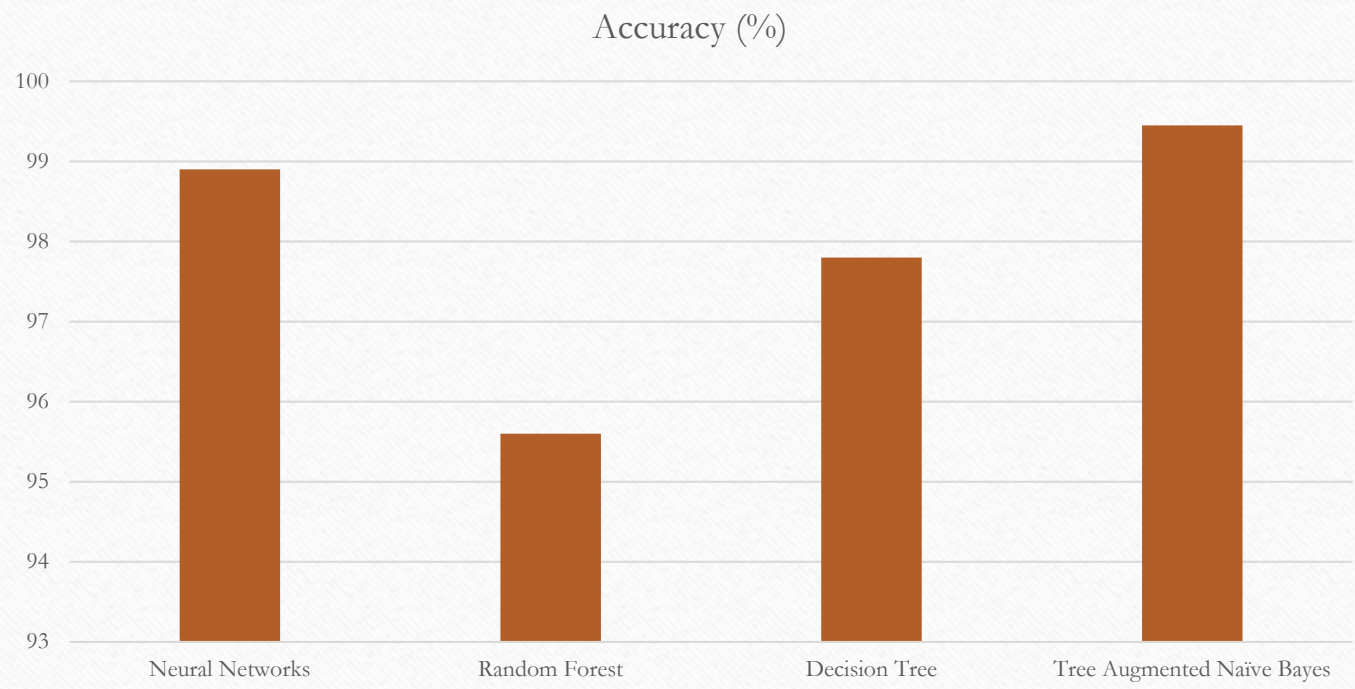
$$F - \text{Measure} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \tag{4}$$

RESULTS AND DISCUSSION

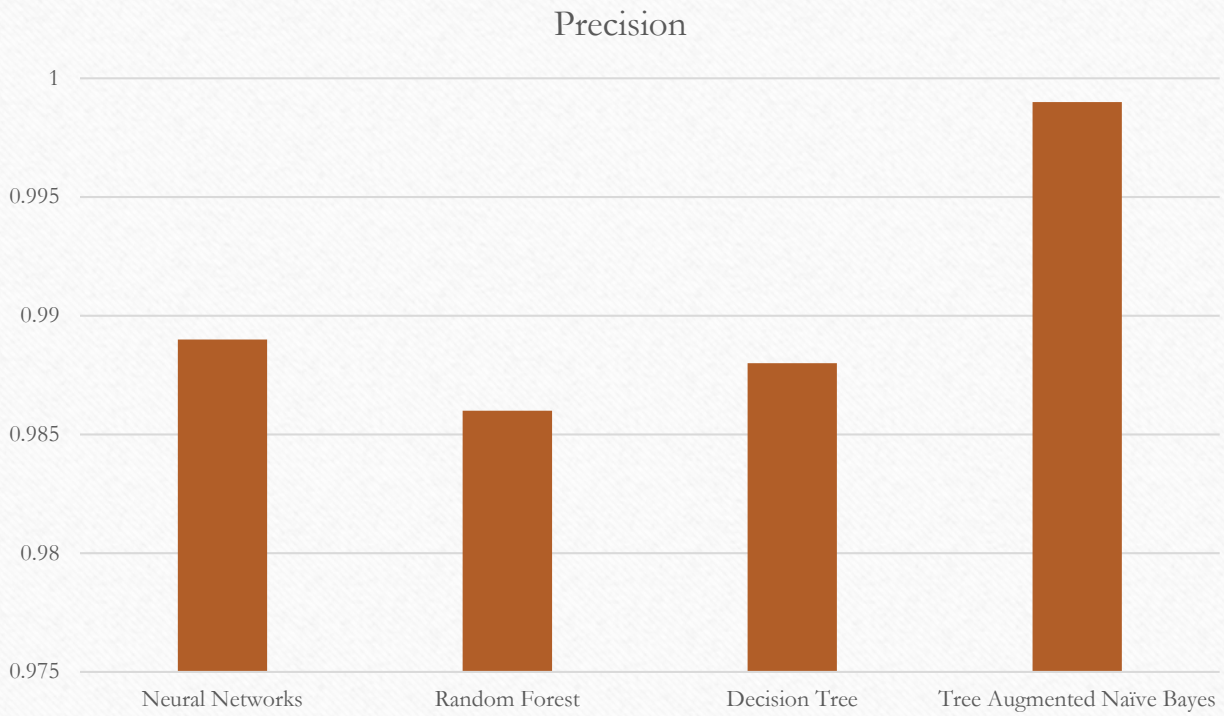
Comparison Result Experiment Performance between Methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Neural Networks	98.900	0.989	1.000	0.994
Random Forest	95.600	0.986	1.000	0.977
Decision Trees	97.800	0.988	1.000	0.989
Tree Augmented Naïve Bayes	99.450	0.999	0.999	1.000

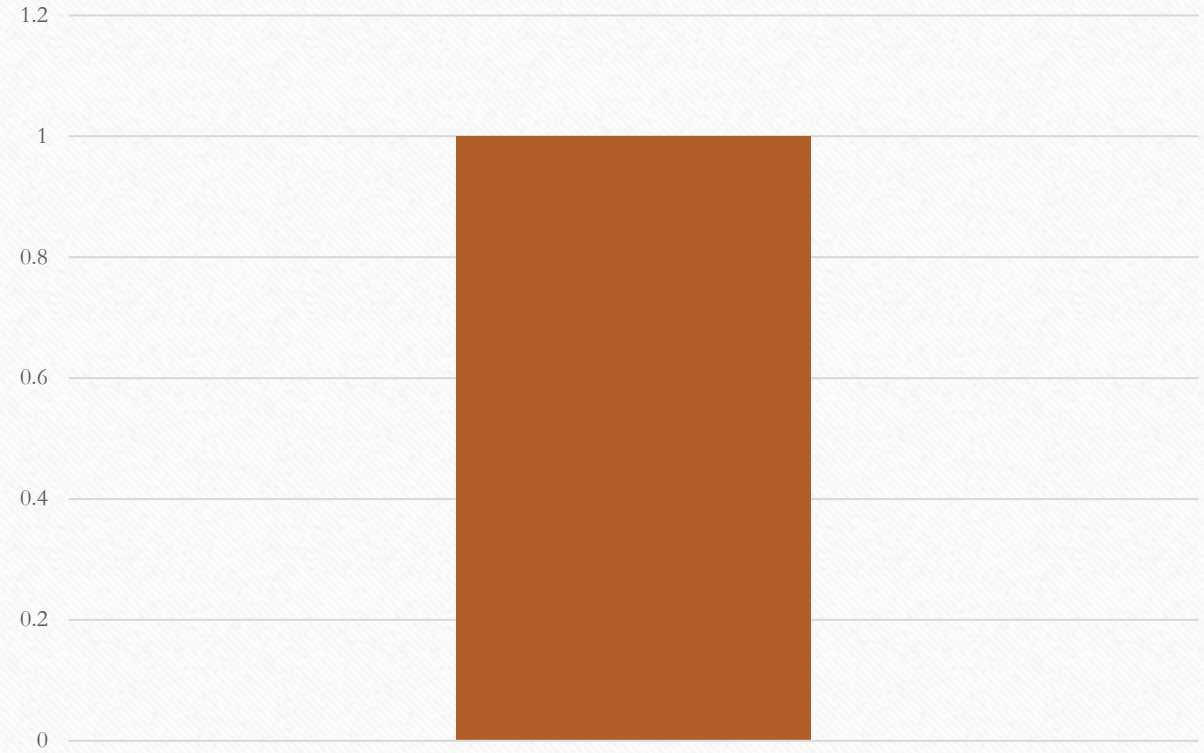
Comparison of Accuracy



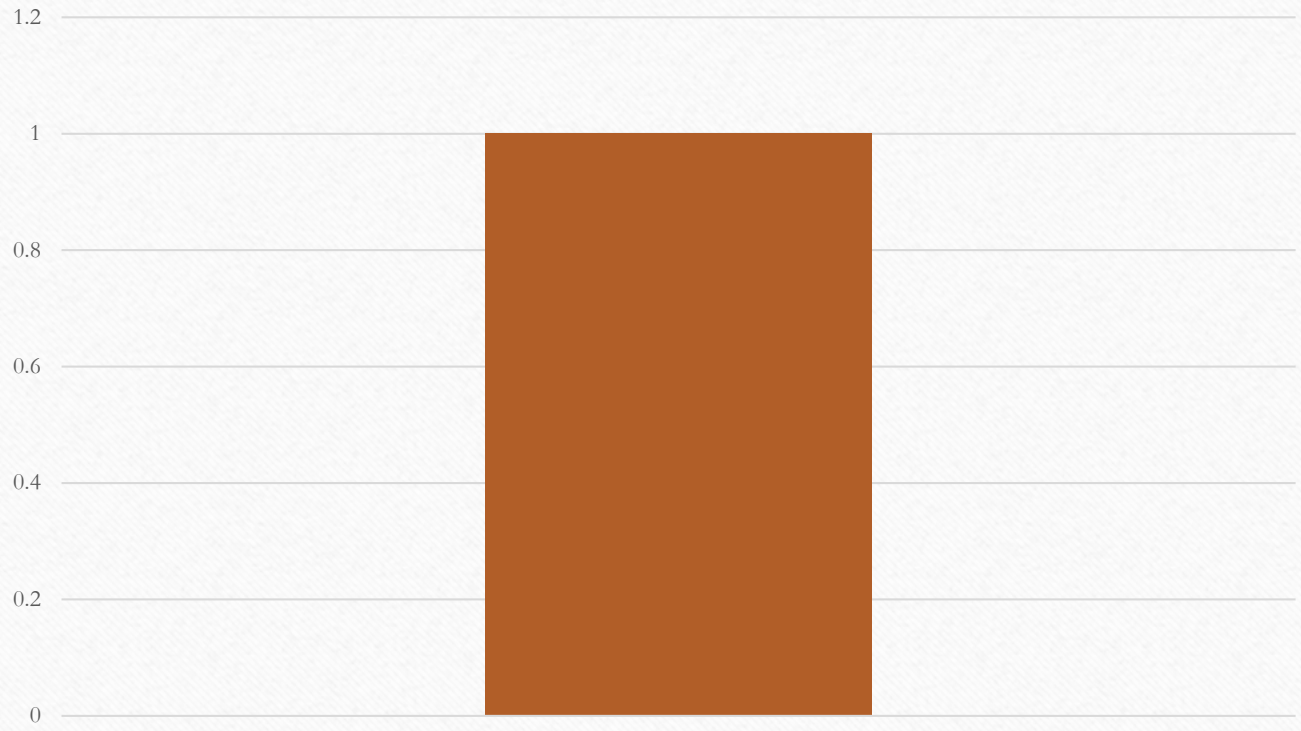
Comparison of Precision



Comparison of Recall



Comparison of F-Measure



Conclusions

- This research proposed an investigation of three machine learning algorithms; Neural Networks, Random Forests, and Decision Tree for predicting flood risks.
- The challenge was to choose the prediction algorithms that were suitable for huge numerical data of people that involve in the flood.
- The prediction model from the case study is expected to be generalized into a more comprehensive model to cover different flood risk data.
- Neural Networks is the most efficient algorithm in this investigation since the accuracy are higher than other type of approaches.

Future Work

High dimensionality, scalability and accuracy are focus to consideration for further research which different algorithms can be tested.

Thank You