A Study on Flood Prediction Model Using Machine Learning : Focused on Busan Metropolitan City

Ha, Ji Hye
Pusan National University[1]
Wlgp5346@pusan.ac.kr

Kang, Jung Eun
Pusan National University[1]
jekang@pusan.ac.kr

Abstract

Recently, flood damage in urban cities have been aggravated due to the increase of abnormal rainfall from climate change and localized heavy rains along with the urbanization from economic development. Therefore, there needs to be a comprehensive urban planning and preventive strategies. To reduce the flood damage in urban cities, there should be spatial structure designs and land use planning to evaluate the flood risk. However, as flood damage is affected be various factors, it is difficult to develop a predictive model. Therefore the purpose of this study is to develop a flood prediction function using machine learning and establish a flood risk map.

This study used the 2014 Busan Metropolitan Flood Information Data, and analyzed after dividing the grid on a 30m×30m scale. The actual analysis used four machine learning techniques such as Decision Tree, Random Forest, Naïve Bayes and Support Vector Machines using R-programming to develop the flood prediction model, and the study developed the flood risk map using the Jenks Natural Breaks Classification of ArcMap and categorizing and visualizing the risks into five levels.

After comparing the four models, it was shown that the Random Forest model was the most appropriate model for flood predictions. Therefore the weight of the variables was deduced by the importance of the contribution to the model, and the values from the results were used to develop the flood risk map. The results showed that the districts with the highest risks were Jeonggwan-eup, Gijang-gun, Geumjeong-gu, Dongnae-gu, and Yeonje-gu, and it was shown that the application level of the flood risk map is high as the results showed overall similar results when comparing with the actual areas with flood damage. The results of this study will lead to avoiding inappropriate developments in areas with flood risks and inducing developments for areas with low risks, and will be applied as important data for guidelines on flood risk evaluations in the future.

Keywords: Urban Flood, Big Data, Machine Learning, Random Forest, Decision Tree

---

[1] 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan, Republic of Korea

Introduction

The degree of natural disaster is increasing globally due to the intensified climate changes. In Korea, storm and flood damage from typhoons and heavy rain consist most of the natural disasters. In addition, due to the topographical trait where more than 70% of the country is mountainous, there is an increase of runoff in the urban areas in case of localized heavy rains, leading to great damage (Lee et al., 2016). There are various causes and types of urban disasters, and as population is focused on urban areas and residential and industrial facilities are highly developed, in case of a disaster, the damage is largely expanded, leading to not only property damage but also loss of lives, and also a total paralysis in the urban functions (Korea Research Institute for Human Settlements, 2008; Song, 2012). The 'flood damage in urban cities' in the urban areas can be defined as the damage in lives, physical body and property due to the external floods which occurs in case of floods, and internal floods which occurs when the sewage and other drainage systems cannot release storm water (Korea Research Institute for Human Settlements, 2008). Recently, flood damage in urban cities have been aggravated due to the increase of abnormal rainfall from climate change and localized heavy rains along with the urbanization from economic development which is more focused in terms of time and place compared to the past (Park et al., 2007; Son et al., 2010; Ministry of Land, Infrastructure and Transport, 2015). In addition, as the high density development of the built-up areas, the increase of impermeable areas due to urban development and artificially created areas such as underground space increase the flood damages, the flood damage in urban cities can be considered man-made.

Therefore, the increase of urban flood damage due to climate change and urbanization requires a comprehensive urban planning and preventive strategies. Currently, Korea does not have detailed policies for maintenance of areas prone to floods. Although there was basis for redevelopment or reconstruction along with building regulations and incentives by regulating the 'Disaster Management Areas' in the Building Act, as part of the rationalization of regulation in December 2005, the 'Disaster Management Areas' from the Building Act disappeared and became unified with 'Disaster Prevention Zones' by the National Land Planning Act. However, as there is still no detailed ordinance on the 'Disaster Prevention Zones', there are problems in starting the maintenance projects (Shin, 2006).

To reduce flood damage in urban cities, there should be spatial structure designs and land use planning to evaluate the flood risk. However, as flood damage is affected be various factors, it is difficult to develop a predictive model. These problems can be solved by using Big Data. Therefore the purpose of this study is to develop a flood prediction function using machine learning and establish a flood risk map.

Theory and Method

Since the mid-2000s, urban flood was focused as a serious urban problem, and there were numerous

studies to develop detailed techniques to analyze floods by the urban characteristics in Korea following the development of meteorological technology, spatial information technology and hydrological technology. Although the studies collected various data and secured technology, there are still problems from the topographical interpretations centered on lowland floods, and difficulties in data collection and modification due to excessive mediating variables (Lee et al., 2016).

Prior studies on the vulnerability and risks of urban floods nationally and internationally normally used statistical approaches or the GIS method. However, there are studies in recent days that apply machine learning techniques, which showed higher prediction levels than the linear regression models, in the field of disaster prevention (Sakr et al., 2010; Asim et al., 2017; Choi et al., 2018).

Studies on urban floods using machine learning has not been considered in Korea. However, there were studies that used the Decision Tree and applied the vulnerability index based on spatial information for all of Korea to set and analyze the preventive measures for extreme floods with the results and analysis of the vulnerability by sub basin areas (Jang et al., 2009), used the Random Forest model and the Boosted Tree model to analyze the vulnerability of floods and avalanches in Seoul (Lee, 2017) and used the Random Forest model and the Support Vector Machine model to develop predictive functions of heavy rain damage of the metropolitan areas (Choi et al., 2018).

1) Scopes of the Study

This study analyzed the flood prediction function by using the 2014 Busan Metropolitan Flood Information Data. The spatial range of the study was on 881,350 grids with a 30m×30m scale of the entire Busan Metropolitan area for analysis, and the rater data included in the grid were deduced by the means. The temporal scope was set at 2014, when there was a great damage due to the localized heavy rains. The actual analysis used four machine learning techniques such as Decision Tree, Random Forest, Naïve Bayes and Support Vector Machines using R-programming to develop the flood prediction model, and the study developed the flood risk map using the Jenks Natural Breaks Classification of ArcMap and categorizing and visualizing the risks into five levels.

2) Variables

The variables used in this study are shown from <Table 1>. Based on the information on the floods of Busan Metropolitan, the indexes such as traces of floods, number of days with heavy rain of at least 80mm/day, maximum rainfall in an hour, altitude, slope, distance from rivers, impermeability, and disaster prevention facilities were divided into disaster experience, climate factors, geographical/topographical factors, development factors and facility factors. The indexes were processed into the geographical spatial information for the analysis.

The climate factors that are expected to directly affect floods were identified as the number of days with heavy rain of at least 80mm/day and the maximum rainfall in an hour that can reflect the extreme

weather, and the geographical/topographical factors that are expected to affect the flow of water and the decision of joining points were identified as the altitude, slope and distance from rivers. In addition, the impermeability was identified to show the development level of the urban areas to consider the surface runoff and drainage abilities of the soil, and disaster prevention facilities were identified as the facility factors.

Table 1. Variables used for Analysis

Index		Types
Disaster Experience	Traces of Floods	Discrete
Climate Factors	Number of days with heavy rain of at least 80mm/day	Continuous
	Maximum rainfall in an hour	Continuous
Geographical/ Topographical Factors	Altitude	Continuous
	Slope	Continuous
	Distance from rivers	Continuous
Development Factors	Impermeability	Discrete
Facility Factors	Disaster Prevention Facilities	Discrete

3) Analysis

Data Preprocessing

As machine learning techniques need to classify the data for the model training and verification of the data, after dividing in 7:3, there were 616,945 areas of training datasets and 264,405 areas of test datasets. The sample() function that randomly extracts data was used in this process, and replace=FALSE was set to avoid repetition of the same area.

This study developed the flood prediction function with the training dataset (70% of the total data) and applied the developed function to the test dataset to compare the predicted values and the actual values to evaluate the predictive abilities.

Decision Tree Model

The Decision Tree classifies the decision making rules of a particular item in a structure of a tree, and has the advantage of being more objective and easy to interpret than the other methods (Breiman et al., 1984). Therefore, it can provide the classification standards of each influence variables used as independent variables for flood prediction, and can calculate the weights of each influence variables.

This study used the ctree() function in the party package of R-programming. The ctree() function uses the unbiased recursive partitioning based on permutation test and selects the variables to be cut based on the significance of the p-test, therefore there are no risks of over-fitting or to be biased and no need for puming. In addition, it applies the processes that consider the problem of multiple testing from repeatedly dividing the notes in the Decision Tree to stop dividing the nodes at an appropriate point in time.

Random Forest Model

The Random Forest model assigns the maximum randomness to solve the problem of over-fitting in the Decision Tree model and increases the predictability of the ensemble model (Yoo, 2015).

This study used the randomforest() function of the R-programming to develop the model, and used the importance() function to deduce the importance of each variables for weights. In addition, as Random Forest model uses the bootstrap from probability random extraction, there are no fixed values when repeated, thus the study set the seed values through the set.seed() function for identical values.

Although the Random Forest model can go through OOB (out-of-bag) analysis without dividing the model into training datasets and test datasets unlike the Decision Tree model, this study used the training datasets and test datasets at a 7:3 ratio for an accurate comparison with other models.

Naïve Bayes Model

The Naïve Bayes model is based on the Bayes theory, and simplified the calculation of the posteriori probability by assuming the conditional independence. The Naïve Bayes model is generally appropriate for problems that need to consider various attribute data to measure the total probability of a result. If all situations are independent, it is impossible to predict another situation by observing another situation. This study used the naiveBayes() function of the e1071 package of R-programming.

Support Vector Machine

The Support Vector Machine (SVM hereon) model is a model that finds the line or plane with the maximum distance between data of different classifications and classifies the data based on the line or plane (Lee, Chung, et al., 2016). SVM is generally used as the optimal method out of the classification methods that operate well in various data distributions, as it has greater accuracy and lower possibility of over-fitting compared to other classification methods (Choi et al., 2013). When SVM predicts a new data, it measures the distance between the data and each support vector, and the decision on the classification is based on the distance to the support vector and the importance of the support vector is learned during the training process.

This study used the svm() function of the e1071 package of R-programming for the analysis. Although there are packages such as e1071, klaR, kernlab, shogun, and svmpath for the SVM model, the e1071 package realized the OpenSource SVM program library by C++ in R, was first introduced in R and is the most objective.

Results

1) Results of Machine Learning Analysis

Results of the Decision Tree Model

The results of the Decision Tree Model were classified into 15 ranges. ①NODE 4(n=240722): When

the maximum rainfall in an hour is 72.607mm and under, there are no floods when the maximum rainfall in an hour is 70.747mm in non-impermeable areas. ②NODE 6(n=862): When the maximum rainfall in an hour is 72.607mm and under, there may be rare cases of floods when the maximum rainfall in an hour exceeds 70.747mm in non-impermeable areas and the distance from rivers is 90m and under. ③ NODE 7(n=7616): When the maximum rainfall in an hour is 72.607mm and under, there are no floods when the maximum rainfall in an hour exceeds 70.747mm in non-impermeable areas and the distance from fivers exceeds 90m. ④NODE 10(n=105616): When the maximum rainfall in an hour is 72.607mm and under, there are no floods when the maximum rainfall in an hour is 66.664mm and under in impermeable areas. ⑤NODE 11(n= 4948): When the maximum rainfall in an hour is 72.607mm and under, there may be rare cases of floods when the maximum rainfall in an hour exceeds 66.664mm and is equal to or less than 69.45mm in impermeable areas. ⑥NODE 13(n=306): When the maximum rainfall in an hour is 72.607mm and under, there is approximately 25% of flood when the maximum rainfall in an hour exceeds 69.45mm and the slope is equal to or less than 0.135 degrees in impermeable areas. ⑦NODE 14(n=7421): When the maximum rainfall in an hour is 72.607mm and under, there is approximately 2.5% of flood when the maximum rainfall in an hour exceeds 69.45mm and the slope exceeds 0.135 degrees in impermeable areas. ⑧NODE 18(n=4712): When the maximum rainfall in an hour exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 32% of flood when the slope is 0.579 degrees and under in impermeable areas. ⑨NODE 19(n=8895): When the maximum rainfall in an hour exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 10% of flood when the slope exceeds 0.579 degrees in impermeable areas. ⑩NODE 21(n=7103): When the maximum rainfall in an hour exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 2.5% of flood when the number of days with heavy rain of at least 80mm/day is 4.042 days and under in impermeable areas. ⑪ NODE 22(n=11685): When the maximum rainfall in an hour exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 10% of flood when the number of days with heavy rain of at least 80mm/day exceeds 4.042 days in impermeable areas. ⑫NODE 25(n=8127): When the maximum rainfall in an hour exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 10% of flood when the distance from rivers is 90m and under, and the number of days with heavy rain of at least 80mm/day is 4.772 days and under. ⑬NODE 26(n=878): When the maximum rainfall in an hour exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 30% of flood when the distance from rivers is 90m and under and the number of days with heavy rain of at least 80mm/day exceeds 4.772 days. ⑭NODE 28(n=34082): When the maximum rainfall in an hour exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 3% of flood when the distance from rivers exceeds 90m and the slope is 4.941 degrees and under. ⑮NODE 29(n=173972): When the maximum rainfall in an hour

exceeds 72.607mm, out of the areas with a slope of less than 2.051 degrees, there is approximately 10% of flood, there are no floods when the distance from rivers exceeds 90m and the slope exceeds 4.941 degrees.

Out of the seven variables in the Decision Tree model for flood prediction, the tree model nodes used 5 variables of maximum rainfall in an hour(preciptm), impermeability (permeable), slope (slope), distance from rivers (river), and the number of days with heavy rain of at least 80mm/day(precipi80) for flood prediction. The tree model primarily considered the maximum rainfall in an hour when deciding the floods. To evaluate the accuracy of the model, after comparing the test datasets with the results of the predicted functions, the accuracy was 98.9%.

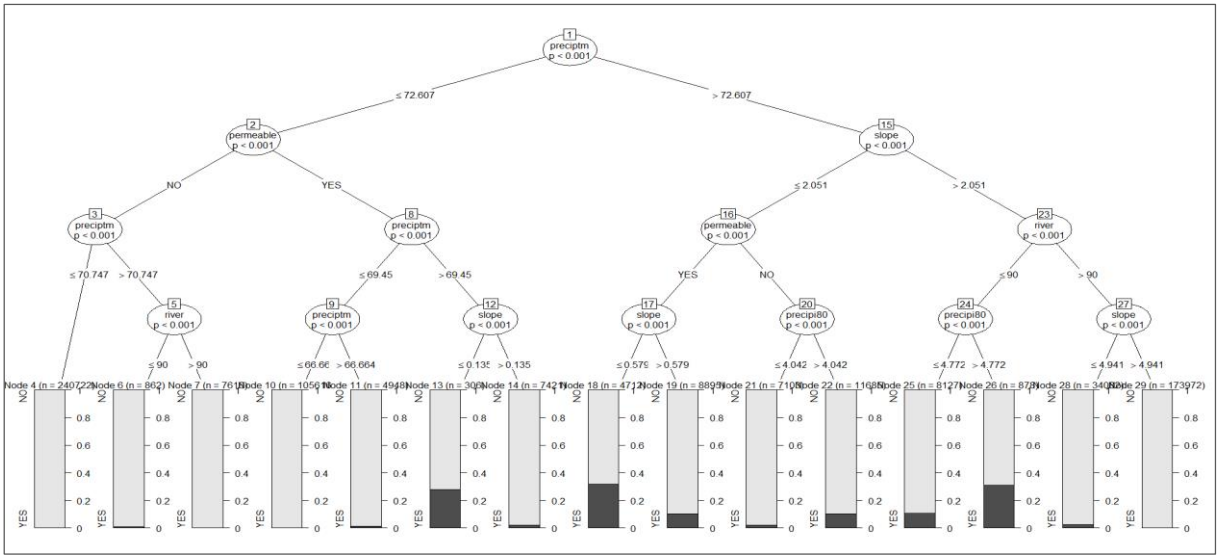

Figure 1. Results of the Decision Tree Model

Results of the Random Forest Model

When showing the Random Forest model with default values without tuning the mediating variables, the number of trees (ntrees) was 500, and the number of possible explanation variables for each node (mtry) was 2, and the OOB error was 0.36%. When comparing with the previously classified test dataset, the explanation power had an accuracy of 99.65%, the sensitivity was 99.9% and the specificity was 77.12%. Although the ntree and mtry automatically sets the default values, it is necessary to modify the parameter values to improve the functions of the model and the accuracy. Therefore there was a cross-evaluation to find the appropriate parameter values.

A total of 9 combinations were compared and analyzed with ntree of 400, 500, and 600 and mtry of 2, 3, and 4. The combination of the lowest OOB error was (500,3), (600,3), and (600,4), with a low error of 0.33% whereas the error of the original model was 0.36%.

There was a re-analysis with a combination of ntree=600 and mtry=3 with low errors, and the importance of each variables of the model is shown from <Table 2>. The results showed that the variables affect flood in the order of the maximum rainfall in an hour (preciptm), number of days with

heavy rain of at least 80mm/day (precipi80), altitude(dem), slope (slope), distance from rivers (river), impermeability (permeable), and disaster prevention facilities (defence).

After comparing the flood prediction of the Random Forest model with the test dataset, out of the 264,405 test datasets, the model accurately predicted no floods for 261,148 datasets and accurately predicted floods for 2,358 datasets, showing an accuracy of 99.66%. The sensitivity was 99.87% and the specificity was 80.78%.

Table 2. Value of Importance of Variables

	NO	YES	Mean Decrease Accuracy	Mean Decrease Gini
dem	34.366	249.071	48.733	2847.628
slope	61.338	173.087	93.152	2061.701
permeable	33.766	69.585	67.065	277.139
precipi80	60.629	186.688	85.581	3010.869
preciptm	64.047	540.652	80.552	3258.2
river	92.909	260.025	179.77	1940.045
defence	9.3911	4.178	10.174	2.905

Results of the Naïve Bayes Model

The results of the Naïve Bayes model for flood prediction showed an accurate classification of 98.89% and inaccurate classification of 0.011%.

After comparing the flood predictions of the Naïve Bayes model with the test datasets, out of the 264,405 test datasets, the model accurately predicted no floods for 244,280 datasets and accurately predicted floods for 1,768, resulting in a prediction accuracy of 93.06%. The sensitivity was 93.42% and the specificity was 60.57%, and the dual results table of the Naïve Bayes model is shown from <Figure 2>.

	nbpred		
test.data$flooding	NO	YES	Row Total
NO	244280	17206	261486
	10.007	129.447	
	0.934	0.066	0.989
	0.995	0.907	
	0.924	0.065	
YES	1151	1768	2919
	896.471	11595.957	
	0.394	0.606	0.011
	0.005	0.093	
	0.004	0.007	
Column Total	245431	18974	264405
	0.928	0.072	

Figure 2. The Dual Results Table of the Naïve Bayes Model

Results of the Support Vector Machine Model

Gamma and cost are the parameter values of the SVM model. Gamma is the parameter that is required in all kernels aside from linear kernels, and cost refers to the cost of violating the restraints. The gamma in this study was 0.125 from 1/(data dimension), and the cost was set at the default level of 1.

The results of the SVM model for flood prediction showed a relatively high accuracy with 98.896%, but after verifying the predicted data, it was only possible to predict when there are no floods. The distribution of the variables by the results of the classification from the plot() function is shown as boxes in <Figure 3>. The boxplot is a graph that summarizes the descriptive statistics such as the median, the first quartile and the third quartile in boxes. Therefore it was proven to be inappropriate for flood prediction and evaluate the flood risks. In addition, after comparing with the test dataset, out of the 264,405 test datasets, the model accurately predicted no floods for 261,486 datasets. The model was shown to be highly extreme with a sensibility of 100% and specificity of 0%; as it predicts that all data will not result to floods, it has a clear error of classification, which can lead to a possible modification of the parameters in the future.

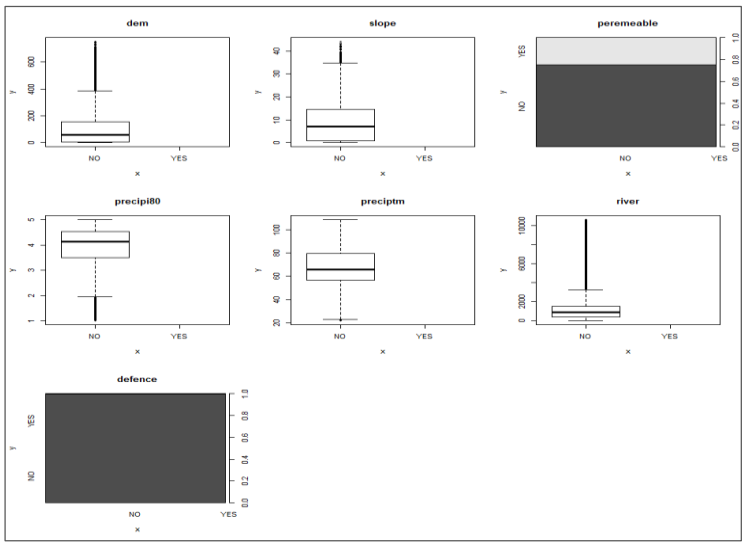

Figure 3. Distribution boxplot of the Variables (SVM)

2) Comparison of the Machine Learning Analysis

The Confusion Matrix is generally used to evaluate the classification, how practically the model classified and how accurately and thoroughly the model classified. Accuracy, out of the indexes that can be identified from the Confusion Matrix is the most prominently used index, which shows how the model accurately classified the data by calculating the percentage of the accurately classified data out of the entire data.

After comparing the accuracies of the four models in this study (Decision Tree, Random Forest, Naïve Bayes and Support Vector Machines), the accuracy of the Decision Tree was 98.9%, Random Forest was 99.66%, Naïve Bayes was 93.06% and the Support Vector Machine was 98.9%. The Random Forest

model was shown to be the most appropriate model for flood prediction <Table 3>.

Table 3. Comparison of Accuracy of the Models

Decision Tree	Random Forest	Naive Bayes	Support Vector Machine
98.9%	99.66%	93.06%	98.9%

3) Evaluation of Flood RiskPrediction

Deducing Weights by Using Random Forest

The Random Forest model, which was selected in the end, had the least errors by categorizing the flood damage of numerous variables with the greatest accuracy. The weights of the variables were calculated by the importance in the contribution to the model, and the results were used to develop the flood risk map.

The weights were deduced as shown from <Figure 4> by using the importance() function of the randomForest package according to the importance of the influenced variables that contributed to the Random Forest model. The values of importance of the influenced variables are showed in positive (+) values, but for the variables such as the distance from the streams, it was shown that they have a negative (-) relationship with the floods in the Decision Tree, thus were given negative weights when calculating the flood risks.

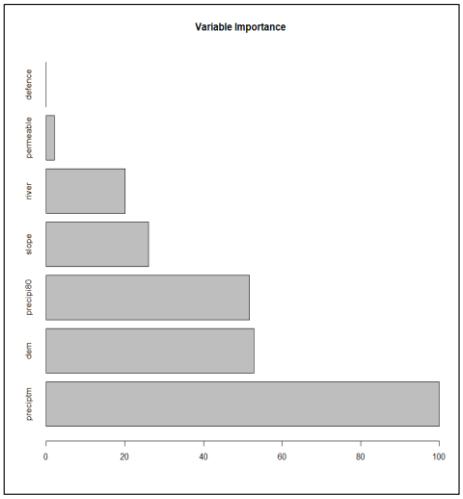

Figure 4. Variable Importance

Flood Risk Map

After evaluating the flood risk predictions, the calculated risks were shown with a map of approximately 900 thousand areas on a scale of 30m×30m in Busan Metropolitan <Figure 5>.

The study used the Jenks Natural Breaks Classification from the ArcMap classification to categorize and visualize the risks into five levels. Jenks Natural Breaks Classification categorizes the arrangement of the data values by optimizing them into natural levels, and minimizes the average deviation and

maximizes the dispersion based on the mean of the total values within a level. In other words, it minimizes the dispersion within a level and maximizes the dispersion among the levels, and is generally used in cases of large differences of data such as this model (Jenks, George F. 1967).

The developed flood risk map of Busan Metropolitan categorizes the flood risks into five levels; the areas of high risks were Jeonggwan-eup, Gijang-gun, Geumjeong-gu, Dongnae-gu, and Yeonje-gu, and Gangseo-gu was shown to be a low-risk area, safe from floods.

After comparing the results with the actual areas of flood damage, the actual flooded areas of 2014 were distributed in the flood risk areas. During the localized heavy rain in Busan Metropolitan in 2014 greater than 100mm per hour, Gijang-gun had an especially serious damage. The Jwagwangcheon and Deokseoncheon in the neighboring areas were flooded and the Naedeok reservoir was collapsed, leading to a property loss of approximately 68.5 billion KRW. As the results of comparing the actual cases with the analysis results, the two were shown to be similar, it can be determined that the application level of the flood risk map is high.

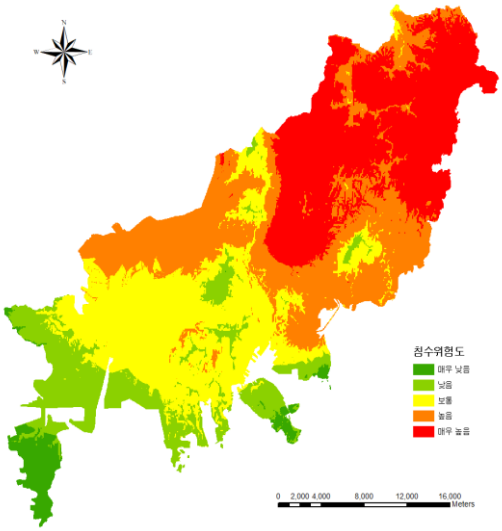

Figure 5. Flood Risk Map

Discussion

This study developed a flood prediction function by using machine learning (Decision Tree, Random Forest, Naïve Bayes and Support Vector Machines) to predict in advance the flood damage of Busan Metropolitan, and developed a flood risk map by using the Jenks Natural Breaks Classification of ArcMap and categorizing the flood risks into 5 levels. After evaluating the predictions of each functions by applying them to the evaluated data, it was shown that the Random Forest model was the most appropriate. Therefore the weights of the variables were deduced by the importance of the variables on the contribution to the Random Forest model, and the results were used to develop the flood risk map.

The areas at risk of floods in Busan Metropolitan were Jeonggwan-eup, Gijang-gun, Geumjeong-gu,

Dongnae-gu, and Yeonje-gu, and after comparing the results with actual areas of flood damage, the actual areas of flood damage were distributed in the high-risk areas of floods. Therefore it can be seen that the application level of the flood risk map is high. However, to verify the results of the analysis on a more quantitative viewpoint, there should be improvements in the reliability verification process such as the comparison between the actual level of damage by the damaged areas and the predicted levels of damage.

Lastly, the flood risk map will lead to avoiding inappropriate developments in areas with flood risks and inducing developments for areas with low risks, and will be applied as important data for guidelines on flood risk evaluations in the future.

Acknowledgment

“This is financially supported by Korea Ministry of Land, Infrastructure and Transport(MOLIT) as 「Innovative Talent Education Program for Smart City」.”

References

Choi, H.S., Park, H.W., and Park, C.Y. (2013). Support vector machines for big data analysis. Journal of the Korean Data & Information Science Society, Vol. 24, No.5, pp.989-998.

Lee, H.H., Chung, S.H., and Choi, E.J. (2016). A case study on machine learning applications and performance improvement in learning algorithm. Journal of Digital Convergence, Vol.14, No. 2, pp. 245-258.

S Lee, JC Kim, HS Jung, MJ Lee, S Lee. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea, Geomatics, Natural Hazards and Risk, Vol. 8, No. 2, pp.1185-1203.

Stumpp, C., Żurek, A.J., Wachniew, P. et al. (2016) A decision tree tool supporting the assessment of groundwater vulnerability. Environ Earth Sci, 75: 1057.

XUNLAI CHEN, HUI LI, SHUTING ZHANG, YUANZHAO CHEN, AND QI FAN. (2019). High Spatial Resolution PM 2.5 Retrieval Using MODIS and Ground Observation Station Data Based on Ensemble Random Forest. IEEE Access, Vol. 7, pp.44416-44430.