ISSN: 2508-7894 © 2020 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai.2022.10.1.1

A study on Natural Disaster Prediction Using Multi-Class Decision Forest

Tae-Hyuk Eom[1], Kyung-A Kim[2]

Received: April 25, 2022. Revised: April 26, 2022. Accepted: May 15, 2022.

Abstract

In this paper, a study was conducted to predict natural disasters in Afghanistan based on machine learning. Natural disasters need to be prepared not only in Korea but also in other vulnerable countries. Every year in Afghanistan, natural disasters(snow, earthquake, drought, flood) cause property and casualties. We decided to conduct research on this phenomenon because we thought that the damage would be small if we were to prepare for it. The Azure Machine Learning Studio used in the study has the advantage of being more visible and easier to use than other Machine Learning tools. Decision Forest is a model for classifying into decision tree types. Decision forest enables intuitive analysis as a model that is easy to analyze results and presents key variables and separation criteria. Also, since it is a nonparametric model, it is free to assume (normality, independence, equal dispersion) required by the statistical model. Finally, linear/non-linear relationships can be searched considering interactions between variables. Therefore, the study used decision forest. The study found that overall accuracy was 89 percent and average accuracy was 97 percent. Although the results of the experiment showed a little high accuracy, items with low natural disaster frequency were less accurate due to lack of learning. By learning and complementing more data, overall accuracy can be improved, and damage can be reduced by predicting natural disasters.

keywords : Natural disaster, Afghanistan, Azure Machine Learning, Muliclass Decision Forest

Major classifications : Machine Learning, Supervised Learning, Multiclass decision Forest

1. Introduction

AlphaGo has raised interest in artificial intelligence, an d research on artificial intelligence and machine learning h as become active. Machine learning is a field of artificial in telligence in which a computer learns through data and pro vides results based on the data. (Kim, 2017)

Disasters are damages caused by typhoons, floods, heavy rains, storms, tsunamis, heavy snow, droughts, e arthquakes and other similar natural phenomena. In ge neral, when people's social life, human life, or propert y are damaged by external forces such as abnormal n atural phenomena, they are called disasters and divided into natural and man-made disasters according to the cause of occurrence. Among them, natural disasters a re caused by natural phenomena. (Han, 2007)

According to Wikipedia, Afghanistan is geographically located inland in Asia and is somewhat dry, but in winter, the snow in the Hindu Kush Mountains and the Pamir Plateau melts in spring and supplies water to rivers and rivers. However, climate change has led to severe droughts, killing many people in unexpected summer floods. And every year, the Hindukush Mountains record many deaths

1 First Author. Student, Medical Information Intelligence (Department of Medical IT), Eulji University, Korea.
Email: djaxl96@gmail.com
2 Corresponding Author, Researcher, NEID Inc., Korea.
Email: kyungakim@naver.com

in areas prone to earthquakes due to landslides and avalanches during the winter. According to various articles, many people die from natural disasters in Afghanistan every year. There are many disasters caused by floods, and not only casualties but also property damage is significant. However, if information on natural disasters can be predicted and dealt with in advance, various losses such as deaths and property damage that occur every year can be minimized.

Therefore, we predicted the climate of Afghanistan for climate change, and the machine learning algorith m used among the various machine learning algorithm s is multi-class decision forest (Multi-Class Decision Forest).

2. Literature Review

According to Yoon's paper, "Prediction of Land Slide Probability around Railway Using Decision Tree Model", the study was conducted for the first time because the geology of Korea's torrential rain areas is weak and there is a high possibility of landslides. The prediction of landslides was based on the prediction model of the decision tree model landslide, and the prediction map of landslides was prepared based on these prediction results. The Decision tree model analysis method used chi-square and Gini's coefficient as statistical analysis methods to classify areas where landslides could occur (Yun, 2017). According to Choi's paper, "Natural Disaster Damage Cost Prediction Model based on Neural Network and Generic Algorithm," we developed a neural network model to reduce social and economic losses due to disasters and disasters. A neural network imitates a human's ability to learn and consists of a network of artificial neurons, which are in an input layer, a concealment layer, and an output layer and serve to provide an external result value to a user. We predicted natural disasters according to the role of neural network metastasis functions and learning algorithms (Choi, 2010).

According to An Su Hyun's paper, Machine Learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying instead on patterns and inference. It is seen as a subset of artificial intelligence. It also focuses on representation and generalization. Representation is the evaluation of data, and generalization is the processing of data that is not yet known. The term Machine Learning was first used by Arthur Samuel, an IBM researcher in the field of artificial intelligence. (An, 2021)

Instructional learning is to make machines learn through materials with correct answers and predict results through materials newly learned through learning. Learning involves instructional learning and autonomous learning. Instructional learning is broadly classified and regressed (Choi, 2020). In the paper, "A Study on Methods to Prevent the Spread of COVID-19 Based on Machine Learning", Supervised Learning is a machine learning course that learns the ability to map inputs to outputs based on the example input/output pairs. Inferring labeled functions. In a training data consisting of a series of training examples, each example is a pair consisting of an input object (normal vector) and a desired output value. A supervised learning algorithm analyzes and generates training data. In an estimation function optimal scenario that can be used to map new examples, the class label of an instance where the algorithm is invisible can be correctly determined (Kwak, 2020). In this study, instructional learning was conducted by making natural disasters learn about types and predict them. According to Nam Yu-jin's paper, "A Study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning," Microsoft Azure Machine Learning was used. Two-Class Support Vector Machine, Two-Class Decision Jungle, Multi-Class Decision Jungle algorithm was used to compare the results. As a result, the Two-Class Support Vector Machine, Two-Class Decision Jungle showed high accuracy. Research using classification techniques predicted that the technology for predicting lung cancer would develop. (Nam, 2019)

According to Jung Yong Chan's paper, the Decision Tree, also known as the crystal tree, is one of the instructional learning models that allows both regression and classification analysis. In addition, decision-making trees can be said to be machine learning algorithms similar to those of the 20 top players who continue to ask questions, for example or no. Analysis using this algorithm is currently being used in various fields such as medical research and market research. The structure of the tree consists largely of Root Node, the highest node containing all the data groups, Leaf Node, and Internal Node, which is between Leaf Node and Root Node. The upper node is called Parent Node and the lower node is called Child Node. The algorithm learns huge amounts of data at a relatively faster rate than other data mining techniques. And it can be visualized with the structure of a tree, so it is intuitive and easy to interpret in understanding the model. These models are non-parametric models that do not require mathematical households such as linearity, regularity, and equivariance. On the other hand, there is a disadvantage that the results are highly variable depending on the independent variables used in the prediction. And the more variables are used, the more complex the tree model becomes, the less predictive it is, and the harder it is to interpret the result. Advantages can be overcome by

branching. Furthermore, the overfitting problem of excessive learning of the learning data and increasing differences to the verification data can be solved through the branching method. Random Forest is an Ensemble learning method used in regression and classification analysis and is a representative machine learning technique that consists of many decision trees and outputs predictions. Ensemble learning is the selection of the results predicted by the majority through the majority voting method in some models. Random Forest is also based on the Bagging algorithm. (Jeong, 2021) Multi-class Decision Forest is an ensemble learning method for classification. After creating multiple decision trees, the most popular output class is selected. Selection standardizes the results in the form of aggregations in which each tree in the classification decision forest outputs an unnormalized frequency histogram of labels, resulting in probabilities for each label. If the prediction reliability is high, the final decision has a higher weight. A decision tree can display nonlinear decision boundaries and has the advantage of selecting and classifying integrated functions. Both the calculation during learning and prediction and the memory usage are efficient. (Yun, J.M., 2017) According to Choi's paper, "Applying Artificial Intelligence for Diagnostic Classification", Azure ML using a multiclass decision forest algorithm was applied, and the diagnostic algorithm score value of 1,269 Korean ADI-R test data was used for prediction. In the second experiment, we used 539 Korean ADI-R case data to apply mutual information to rank items used in the ADI diagnostic algorithm (Choi, 2020). The reason for using the Multi-Class Decision Forest algorithm is that it has a higher accuracy in learning and contrasting data than other classification algorithms.

3. Experiment

3.1. Experimental environment

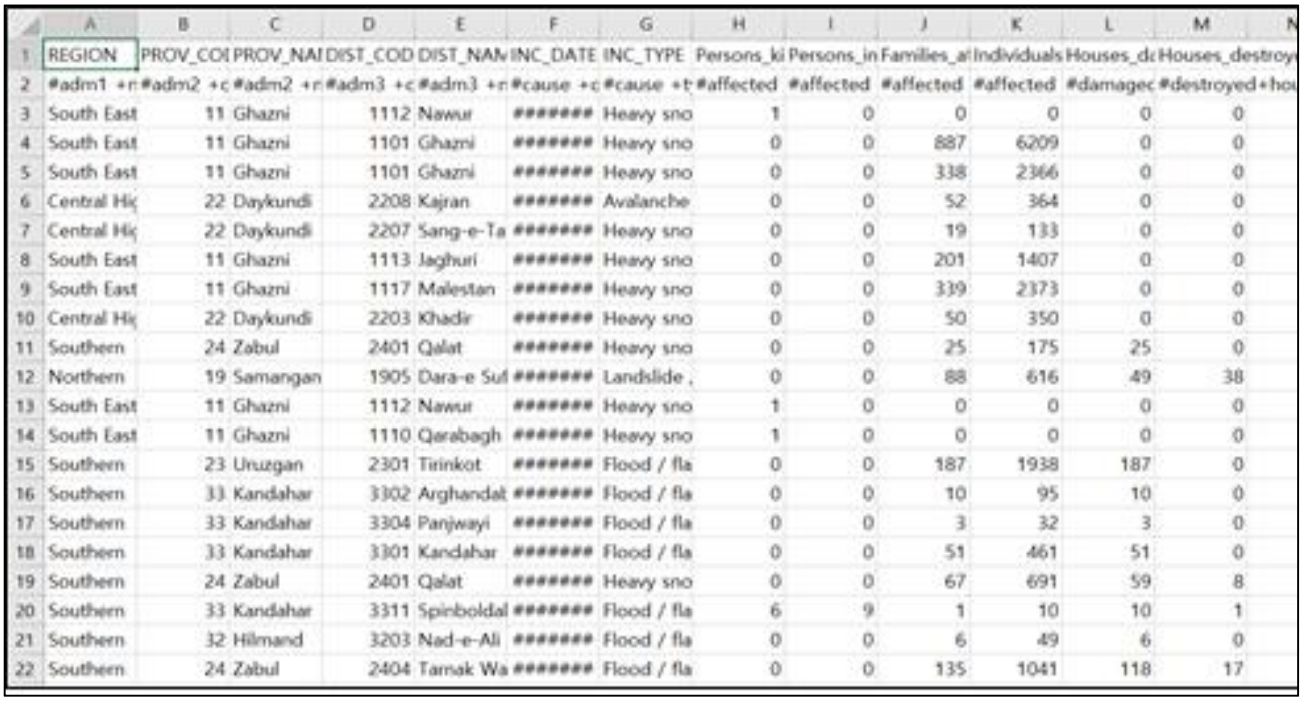
The Azure Machine Learning Studio leverages the Azure platform to collect and manage data in the cloud, create models through the Azure Machine Learning Studio, easily build web services, and then apply them to a variety of devices. In addition, unlike existing cloud platform Machine Learning libraries and tools, it provides an easy-to-access GUI environment for user convenience. Chronic problems with existing Teaching Learning libraries and tools Lack of flexibility in computing resources needed to learn, complexity in configuring GPU-based learning, difficulty in installing tools and environments needed to learn, and difficulty recording and versioning. Unlike traditional machine learning tools, Azure Machine Learning Studio pulls blocks in Drag & Drop format and makes it easy to model. In addition, not only can scripts combined in R and Python languages be inserted in block form and utilized, but the results can be confirmed through visualization. Thanks to the simple structure, anyone can easily create and distribute predictive models if they know how to use them. Azure Machine Learning Studio basically supports data input, output, and visualization, and provides a representative machine learning algorithm that data scientists prefer. These components are utilized to insert experimental data (which can be used in many ways by bringing data to Azure Cloud when learning models), preprocessing experimental data (which needs to be preprocessed in case of missing data to learn models), and saving learned models. Models can be developed and distributed in the same process. (Kang, 2018)

3.2. Data Analysis

Figure 1 shows the data of natural disasters in Afghanistan processed and combined by 2018, 2019, and 2020. This is because they thought that it would not be enough to study data in just one year. Therefore, we combined the data from 2018 to 2020 using Excel. The data consists of 1075 rows and 13 columns. The source of the data is the site that holds the Data world data stack, and the keyword is "Nature Disaster". The data includes avalanches, earthquakes, floods, heavy rain, heavy snow, landslides and mud flows in Afghanistan. It also describes the date, number of victims, number of deaths, and area indicators(Data world, 2020).

Table 1: Data Dictionary

COLUMN NAME	DATA DICTIONARY
REGION	Region of Afghanistan
PROV_CODE	State classification code
PROV_NAME	State Name
DIST_CODE	Region classification code
DIST_NAME	Region Name
INC_DATE	Date
INC_TYPE	Type of natural disaster
PERSON_KILLED	The number of deaths
PERSION_INJURED	The number of wounded
INDIVIDUALS_AFFECTED	Personal influence
FIAMILES_AFFECTED	Family influence
HOUSE_DAMAGED	House damaged
HOUSE_DESTROYED	Destroyed house



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	REGION	PROV_COI	PROV_NAI	DIST_COD	DIST_NAN	INC_DATE	INC_TYPE	Persons_ki	Persons_in	Families_a	Individuals	Houses_dc	Houses_destroy	
2	#adm1 +r	#adm2 +c	#adm2 +r	#adm3 +c	#adm3 +r	#cause +c	#cause +t	#affected	#affected	#affected	#affected	#damagec	#destroyed+ho	
3	South East	11	Ghazni	1112	Nawur	#######	Heavy sno	1	0	0	0	0	0	
4	South East	11	Ghazni	1101	Ghazni	#######	Heavy sno	0	0	887	6209	0	0	
5	South East	11	Ghazni	1101	Ghazni	#######	Heavy sno	0	0	338	2366	0	0	
6	Central Hi	22	Daykundi	2208	Kajran	#######	Avalanche	0	0	52	364	0	0	
7	Central Hi	22	Daykundi	2207	Sang-e-Ta	#######	Heavy sno	0	0	19	133	0	0	
8	South East	11	Ghazni	1113	Jaghuri	#######	Heavy sno	0	0	201	1407	0	0	
9	South East	11	Ghazni	1117	Malestan	#######	Heavy sno	0	0	339	2373	0	0	
10	Central Hi	22	Daykundi	2203	Khadir	#######	Heavy sno	0	0	50	350	0	0	
11	Southern	24	Zabul	2401	Qalat	#######	Heavy sno	0	0	25	175	25	0	
12	Northern	19	Samangan	1905	Dara-e Suf	#######	Landslide ,	0	0	88	616	49	38	
13	South East	11	Ghazni	1112	Nawur	#######	Heavy sno	1	0	0	0	0	0	
14	South East	11	Ghazni	1110	Qarabagh	#######	Heavy sno	1	0	0	0	0	0	
15	Southern	23	Uruzgan	2301	Tirinkot	#######	Flood / fla	0	0	187	1938	187	0	
16	Southern	33	Kandahar	3302	Arghandat	#######	Flood / fla	0	0	10	95	10	0	
17	Southern	33	Kandahar	3304	Panjwayi	#######	Flood / fla	0	0	3	32	3	0	
18	Southern	33	Kandahar	3301	Kandahar	#######	Flood / fla	0	0	51	461	51	0	
19	Southern	24	Zabul	2401	Qalat	#######	Heavy sno	0	0	67	691	59	8	
20	Southern	33	Kandahar	3311	Spinboldal	#######	Flood / fla	6	9	1	10	10	1	
21	Southern	32	Hilmand	3203	Nad-e-Ali	#######	Flood / fla	0	0	6	49	6	0	
22	Southern	24	Zabul	2404	Tamak Wa	#######	Flood / fla	0	0	135	1041	118	17	

Figure 1: Data Set

In order to predict natural disasters in Afghanistan, the seventh type of disaster in the data dictionary is learned without including data such as attention classification codes and regional names that are not necessary for learning data.

3.2. Experimental Process

Register data sets of types and impacts of natural disasters in Afghanistan from 2018 to 2020 in Azure Machine Learning Studio. Data lines with empty or incorrect data values are deleted from the registered data set items so that the experiment can be performed with complete data.

Then, the data used for the experiment and the data column not used were separated and classified to proceed with the experiment using the data column to be used. In this experiment, we proceeded with the learning by removing PROV_CODE (Afghan State Classification Code), PROV_NAME (Afghan State Name), DIST_CODE (Afghan Region Classification Code), and DIST_NAME (Afghan Region Name), which are not related to the experimental results. We propose to split the data set to prevent the model from overfitting the data. When learning a model, it is very important to separate the dataset from the training data and the test data to avoid overfitting. Usually, training data and destination data tend to be divided into 7:3 In the Split property window, the instruction data ratio and Random seed parameter values are determined, but the training data ratio is 0.7 and Random seed value is random (12345). Search for Split in the search palette and place it on canvas and set the training data ratio and Random seed parameters in the Split Properties window. Training data can be entered at 70%, and random randomized can be entered at random. Random seed is a parameter value that is randomly selected because the computer itself is logically assembled. The Multi-class Decision Forest algorithm is an ensemble learning method for classification. It was used in a method of learning by giving a weighted value by providing probability for labels with high prediction reliability for natural disasters. Train Model used INC_TYPE (Afghanistan's natural disaster type) as a step of selecting the variable you want to learn from among various variables. In Split Data, 70% were used for learning data and the remaining 30% were used to predict. Predicted data are presented to show the learning results for each item. In Split Data, 70% were used for learning data and the remaining 30% were used to predict. Predicted data are presented to show the learning results for each item. In the Evaluated Model, accuracy, average accuracy, and average reproducibility can be identified as the result of the evaluation of the prediction model, and visualization can be confirmed by comparing the actual data with the predicted data.

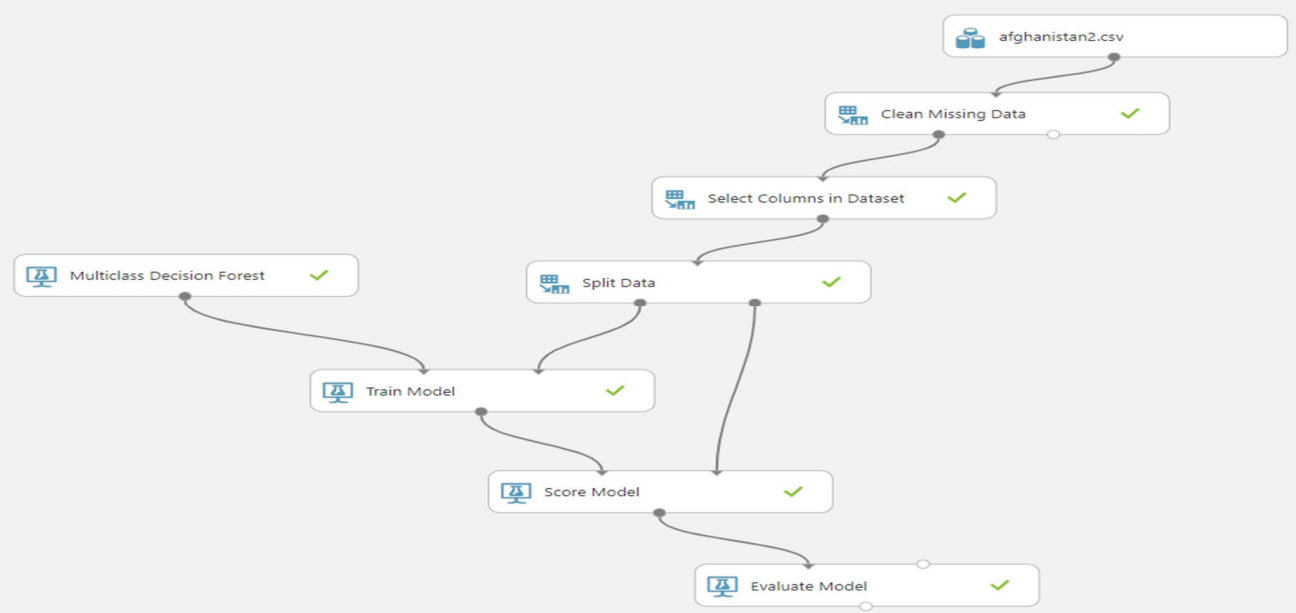


Figure 3: Model Arrangement

4. Results

As a result of advancing the experiment by dividing the learning data and the prediction data in 7:3 in the 775-line, 232 data were predicted. Score Model predicted Avalanche, Drought, Earthquake, Flashflood, Heavy snowfall, and Landslide, respectively. The item with the highest weight among the predicted items is shown as the predicted result and as the result of the Scored Label.

Based on Score Model results, the number of Scored Labels and INC_TYPE predictions are as follows.

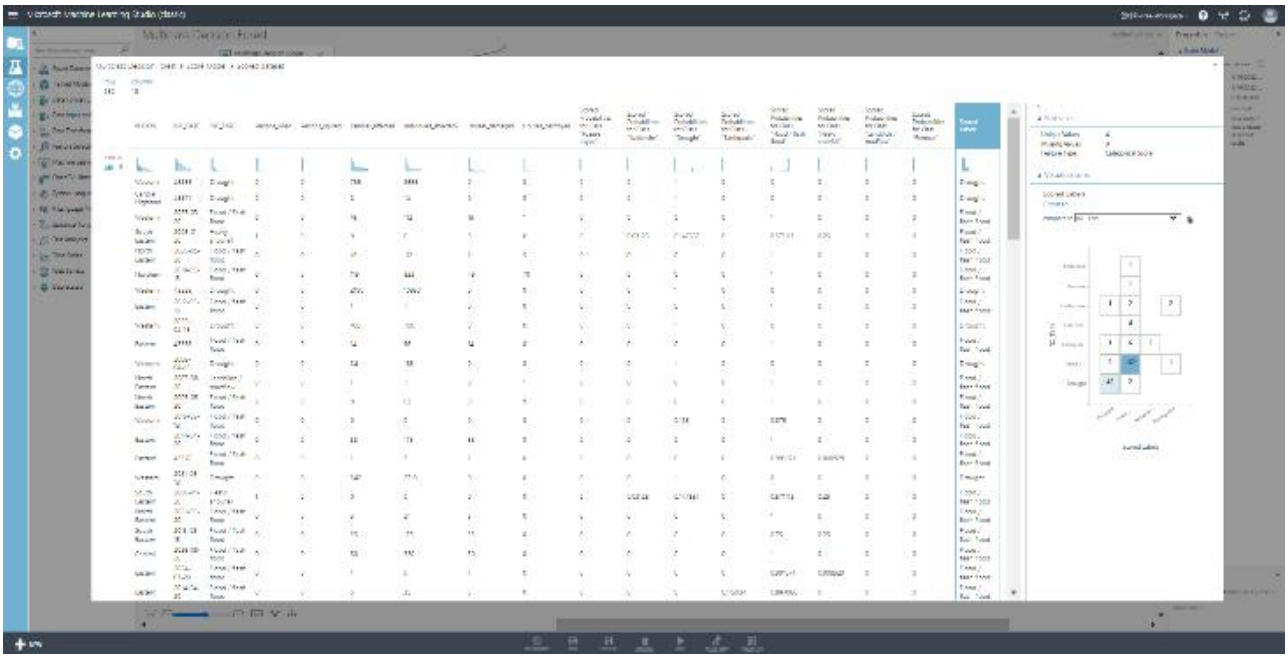


Figure 4: Score Model

Using the Multi-class Decision Forest algorithm, the predicted value was probability. For some data, the probability was 1 and some data were accurately predicted, but for other data, 0.57% was predicted, resulting in a higher weight than for other natural disasters.

Score Labels also provided figures for the types of natural disasters predicted for the original data. In the case of drought, 42 out of 44 were predicted, 162 out of 169 were predicted, 2 out of 9 were predicted, and 3 out of 5 were predicted.

The results of the evaluation of the predictive model shown in the Evaluated Model show high accuracy with overall accuracy of approximately 89% and average accuracy of approximately 97%. The average accuracy is about 89 percent, and the average reproduction rate is about 89 percent. The visualization of the prediction results visualized the accuracy of how accurately the actual data were predicted. The forecast was 95.5 percent for drought and 96.4 percent for floods, but 40.0 percent, 22.2 percent and 0 percent for earthquakes, heavy snow, and landslides. As can be seen from the classification of the number of predicted results in Fig. 4, the study progressed at a rate of 7:3 and found that the accuracy of the prediction increased due to the high frequency of learning of floods and droughts among 30% of the data. However, in the case of heavy snow and earthquakes, the accuracy was low because it was few to learn.

5. Conclusions

The result was obtained using the Decision Forest algorithm. To prevent over-fitting, not only the data are divided in a 7:3 ratio, but also the data to be predicted in the preprocessing of the data are selected so as not to utilize unnecessary data. To predict natural disasters, the model was completed according to the learning method, and the accuracy was slightly higher at 89%. The average accuracy was 97 percent, showing very high accuracy. This shows that the Decision Forest algorithm was good for learning data. I tried machine learning through Microsoft Azure Machine Learning Studio, and it was not only easy to use, but also visually good data learning and experimental results. The results were shown in graphs or tables, and the accuracy and accuracy of reproduction were easily understood by the user. Unlike other Machine Learning tools, the method of use is Drag & Drop, as described above, and works in a simple way, not in a coding and difficult way. I think I can learn a new Machine Learning tool that I used for the first time and use it easily when studying other data.

In the case of drought and flooding in Afghanistan, the snow that falls on the Hindus and Pamir Plateau melts in spring and floods rivers, causing natural disasters such as flooding and frequent drought in desert areas. And because it is a desert area, it shows unpredictable nature. For this reason, we thought that the data for 2018, 2019 and 2020 were insufficient to learn the data, so we went through the process of matching them. From 2018 to 2020, the dataset has shown many times about drought and flooding in Afghanistan, which is sufficient for learning. However, floods and droughts with high accuracy of predictions have a lot of data to learn from Dataset, but this study showed significantly low data on other natural disasters such as earthquakes, avalanches, and landslides, which are not sufficient for learning.

In the case of natural disasters with small data, the accuracy of natural disasters such as earthquakes, avalanches, and landslides can be improved by adding data pre-2018 and learning and complementing it with enough. Therefore, in this study, to improve the accuracy of predicting natural disasters, the amount of data learned in landslides, avalanches, and earthquakes that do not occur frequently is increased and supplemented. Therefore, natural disasters such as droughts and floods, as well as natural disasters such as landslides, avalanches, and earthquakes, are expected to minimize not only casualties but also property damage.

References

Afghanistan (2022). Retrieved April 26, 2022, from https://ko.wikipedia.org/wiki/%EC%95%84%ED%94%84%EA%B0%80%EB%8B%88%EC%8A%A4%ED%83%84

An, S. H., Yeo S. H., & Kang, M. S. (2021). A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree, *Korea Journal of Artificial Intelligence*, *9*(1), 9-14.

Choi, B. J., Park, C. W., Cho, Y. H., Kim, D. S., & Lee, K. W. (2020). A Proposal of New Breaker Index Formula Using Supervised Machine Learning. *Korean Society of Coastal and Ocean Engineers*, *32*(6), 384-395.

Choi, E. S., Yoo, H. J., Kang, M. S., & Kim, S. A. (2019). Applying Artificial Intelligence for Diagnostic Classification of Korean Autism Spectrum Disorder. *Korean Neuropsychiatric Association, 17*(11), 1090-1095.

Choi, S. A. (2010). Natural Disaster Damage Cost Prediction Model based on Neural Network and Genetic Algorithm, *Proceedings of the Korean Information Science Society Conference, 37*(1), 380-384.

data world (2020). afghanistan-natural-disaster-incidents-from-january-to-september. Retrieved April 26, 2020, from https://data.humdata.org/dataset/afghanistan-natural-disaster-incidents-in-2020

Han, S. H., & Yang, K. C. (2007). Investigation of Standardization for Natural Disaster Classification, *The Journal of the Korea Contents Association, 7*(11), 309-319.

Jeong, Y. C., Ryu, H. Y., Lee, S. J., Seo, D. J., & Park, C. G. (2021). Identification recidivism risk factors study based on machine learning: Using decision tree analysis and random forest algorithm, *Korean Police Studies Association*, *20*(1), 323-350.

Kang, M. S., Kang, H. J., Yoo, K. B., Ihm, C. H., & Choi., E. S. (2018). *Getting started Machine Learning with Microsoft AZURE ML*. Seoul, Korea: Hanti Media.

Kim, J. Y. (2017). AlphaGo Case Study: On the Social Nature of Artificial Intelligence, *Journal of Science and Technology Studies, 17*(1), 5-39.

Kwak, Y. S., Kang M.S. (2020). A Study on Methods to Prevent the Spread of COVID-19 Based on Machine Learning, *Korea Journal of Artificial Intelligence, 8*(1), 7-9.

Microsoft Azure Machine Learning (2020). Retrieved April 20, 2020, from https://docs.microsoft.com/ko-kr/azure/machinelearning/studio/what-is-ml-studio

Nam, Y. J., & Shin, W. J. (2019). A Study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning, *Korea Journal of Artificial Intelligence*, *7*(2), 19-24.

Yun, J. M., Song, Y. S., Bak, G. J., & You, S. k. (2017). Prediction of Landslide Probability around Railway using Decision Tree Model, *Journal of the Korean Geosynthetics Society. 16*(4), 129-137.