Prediction of Building Damage Caused by Earthquake with Machine Learning

[1]*Muhammed Ali Hasiloglu and [2]Tuba Tatar
*[1,2]Faculty of Engineering, Department of Civil Engineering Sakarya University, Türkiye

Abstract

Estimating structural damage after an earthquake remains crucial in preventing loss of lives and properties. Conventional methods of damage estimation require a large amount of time and financial resources. For this reason, in recent years, machine learning algorithms that produce faster and more economical results have become the research topics of interest in damage estimation. Within the scope of this study, machine learning models that predict the damage level of the structure after the earthquake have been developed.

In the models created, data sets containing structural and demographic information collected in 11 regions after the 2015 Gorkha, Nepal earthquake were used. Three classes of repair levels labeled by the engineers were chosen as the estimate label. The models were divided into Random Forest and XGBoost according to the classification algorithm they used, and models with and without demographic features according to the data they used.

When the general accuracy rates of the models were compared, the models containing demographic information were more successful. The most successful result is the random forest model with an accuracy rate of 70.83% and the highest damage class recall value of 76.36%.

Keywords: Damage class, Random Forest, XGBoost, Demographic, Earthquake

1. Introduction

An earthquake is an unpredictable natural disaster that affects communities and causes massive damage [1]. Determining the regional distribution and levels of damage on building damage that may occur during an earthquake in advance plays a very significant role in maintaining the speed of action and resource utilization in post-earthquake scenarios [2]. Visually identifying and classifying building damage requires significant time and personnel resources and can take months after the incident [3]. This article targets to form machine learning models that quickly predict earthquake-related damage using various structural and demographic characteristics.

Machine learning has become a frequently used tool in solving important problems in recent years. Compared to traditional approaches, machine learning offers advantages for overcoming complex problems, providing computational efficiency, and facilitating decision-making [4]. For this reason, machine learning is often used in many areas, as well as in the field of damage detection of buildings. Data collected after the 2015 Gorkha earthquake was used for the training of machine learning models.

On April 25, 2015, at 11:56 am local time, an earthquake of magnitude Mw= 7.8 occurred in Gorkha, Nepal. The main shock and the Mw= 7.3 aftershock that occurred 17 days later caused 9,000 deaths, 23,000 injuries, and an estimated $7 billion in economic loss in total [5]. The fact that a large part of the residential building typology consists of unreinforced masonry structures constructed using mud and stone as mortar has significantly contributed to the losses [6]. The Nepalese government conducted a large household survey using mobile technology to assess building damage and rebuild housing in 11 earthquake-affected areas [7]. As a result of these surveys, tags were added by the engineers stating that the building needs repair or needs to be rebuilt. The regions where the surveys were conducted, and the proportions of damaged structures

are given in Table 1.

Table 1. 2015 Gorkha Nepal Earthquake Damage Summary

Location	Population	Buildings	Requiring major repairs or reconstruction
Dhading	430,851	89122	86.07%
Dolakha	279,577	60639	92.54%
Gorkha	360,323	78074	84.74%
Kavrepalanchok	460,464	98019	80.99%
Makwanpur	464,263	90994	41.32%
Nuwakot	372,007	77148	93.10%
Okhaldhunga	190,248	39352	60.09%
Ramechhap	284,051	58623	84.65%
Rasuwa	57,209	12644	95.13%
Sindhuli	352,965	68750	59.70%
Sindhupalchok	425,175	88741	96.34%

About 78% of the building stock in the investigated region needs major repairs or reconstruction. This was a sign that the earthquake had caused a great deal of destruction. Within the scope of this study, none/minor repair, major repair, and reconstruction labels were used as estimation classes. It is known that the structures that will be severely damaged constitute a loss of life more important than economic losses. For this reason, when investigating the performance of the model that will make damage detection, it is important to determine how accurately the class of severe damage is determined, as well as its general accuracy. To see the effect of demographic characteristics on the models, the data were divided into two. Models were created with Random Forest and XGBoost algorithms to predict damage classes.

2. Materials and Methods

The models created in this article are constructed by following the flow chart shown in **Error! Reference source not found.**. First, the data sets are combined based on the building id and region id. Secondly, the labels of the data are freed from outliers for use in machine learning algorithms. To determine the class prediction successes more accurately, the data were balanced by subsampling. Then, the selected machine learning algorithms were validated with cross-validation, parameter selections that would give the best results were made, and the test data and models were tested. Feature selection was made to reduce model flexibility and training time. Finally, the models created were compared with the accuracy rates, Precision, and Recall values of the Reconstruction class.

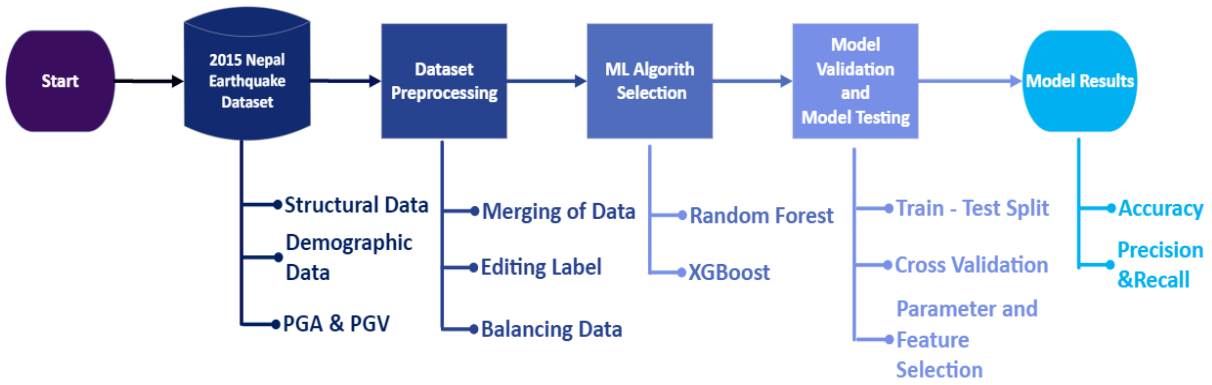


Figure 1. Flowchart

2.1. Datasets

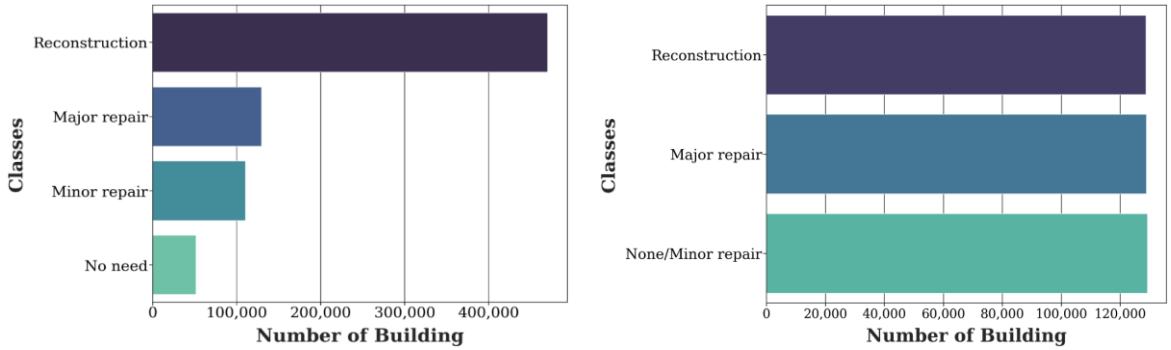
Following the 2015 Gorkha Nepal earthquake, Nepal's Household Registration for Housing Reconstruction Program (HRRP) conducted a large household survey to assess building damage in earthquake-affected areas. Although the initial purpose of this survey was to identify eligible owners who would benefit from government assistance for the reconstruction of housing, other useful socio-economic information was also collected at the census level [8]. The data includes detailed information on 800,000 households belonging to the 11 affected regions [9]. The information collected includes the following features: Information on the physical condition of buildings before and after the earthquake, household composition (age, gender), household income, education level, resources used (tap water, cooking method, toilet facilities, and other facilities) Table 2.

In addition to data from surveys, U.S. Using ShakeMap maps provided by the Geological Survey (USGS), PGA and PGV information of the main and major aftershocks of 110 regions were used. ShakeMap provides near-real-time maps of ground motion and shaking intensity following significant earthquakes [10].

The data to be used for machine learning and their explanations are given in Table 2. Technical solution suggestions labeled by engineers as a result of surveys conducted in Nepal were used as the class to be estimated in the models Figure 2.a.

Table 2. Survey Data After the 2015 Gorkha Nepal Earthquake.

	Datasets Name	**Description**	**Raw Shape**
Buildings	Structural Data	*Physical and material properties of structures before and after the earthquake.*	*(760.000x30)*
	Damage Assessment Data	*Structural damage levels after the earthquake.*	*(760.000x78)*
	Building Ownership and Use	*Features related to the purpose of use of the building.*	*(760.000x16)*
Households	Demographics	*Features such as household size, household income level, and age of the household head.*	*(750.000x10)*
	Households Resources	*Features about the assets owned by the household.*	*(750.000x35)*
Earthquake Parameters	PGA & PGV	*Peak ground acceleration and peak ground velocity values of earthquakes over Mw= 6 that Dec between 2015 and 2016. (USGS ShakeMap)*	*(110x12)*



a) Raw Distribution b) Balanced Distribution

Figure 2. Technical Solution Proposal Label

2.2. Data Preprocessing

To improve model success and create meaningful models in the study, various operations were performed on the data before the model was trained. First of all, the discrete data were combined according to their common features such as structure id and region id. Then, the rows with empty observations in the data were cleaned as they constitute a small part of the data. The prediction label used in this study consists of four-class categorical variables. In these variables, the tags "none" were used when no repair was required in the structure after the earthquake, Minor repair when "minor repairs" were required in non-structural elements, "Major repair" when repairs were required in non-structural and partial structural elements, and "Reconstruction" tags were used when the structure was no longer in use. As seen in Figure 2, the numerical distribution of the labels is unbalanced. Since the classification algorithms to be used will perform better in balanced data, first of all, data balancing is implied. Random down sampling, one of the best simple methods, can be used when there are enough samples [11]. In this study, these two classes were combined because the number of None and Minor Repair labels is small, and they are classes with similar characters. Then the other classes are down-sampled according to the minority class Major Repair [Figure 2.b].

Two data sets were created to be used in the models over the balanced data set. The first dataset created includes the pre-earthquake physical condition of the building, earthquake parameters, and various geotechnical properties. In the second dataset, household data was added in addition to the first data [Figure 3. Splitting Data]. The purpose of creating two different data is to see the effect of demographic characteristics on the success of damage level estimation.

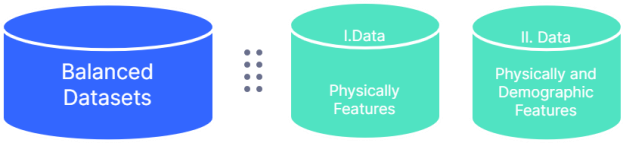


Figure 3. Splitting Data

Most of the features in the data consist of categorical variables. Machine learning models require all input and output variables to be numeric. If there are categorical variables in the data, it must be numerically coded before evaluating the model [12]. Each class in the categorical variable is represented as a binary variable. In addition, the dummy variable trap was avoided by extracting a class from each categorical variable. Relationships of some of the features used with their damage levels are given in Figure 4-Figure ***14***.

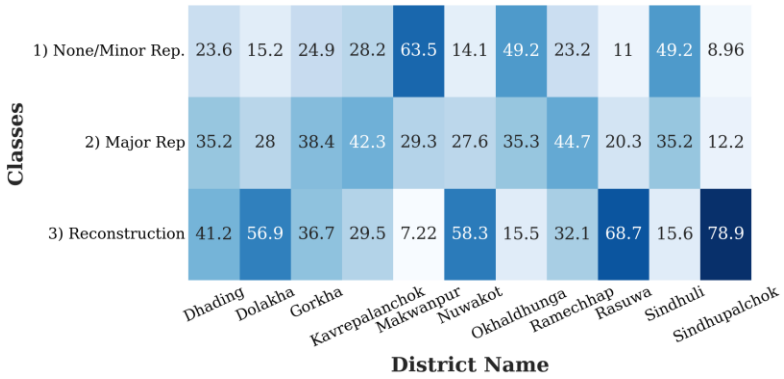


Figure 4. Damage Distribution of Regions

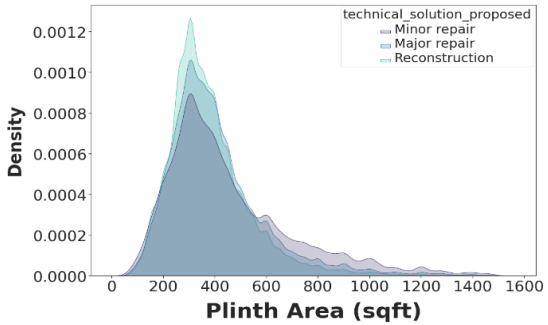


Figure 5. Plinth Area Distribution

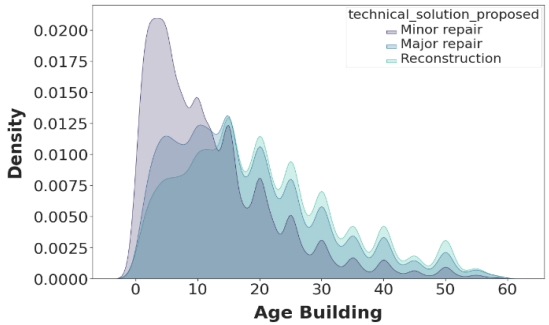


Figure 6. Age Building Distribution

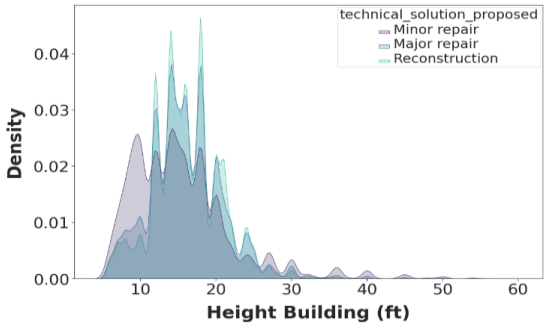


Figure 7. Height Building Distribution

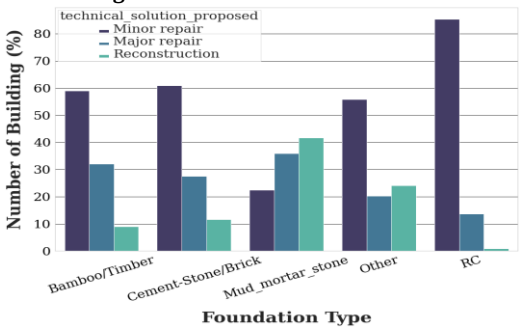


Figure 8. Foundation Type Distribution

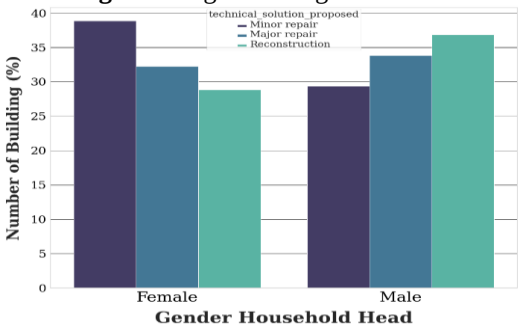


Figure 9. Gender Household Head Distribution

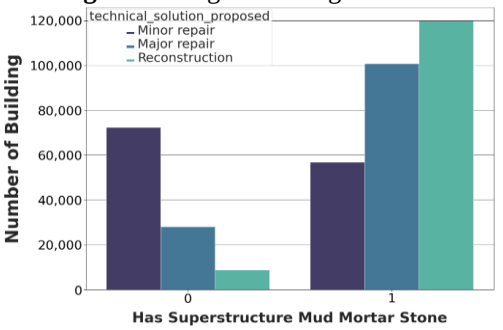


Figure 10. Mud-Mortar Stone Distribution

Figure 11. Ownership Status Distribution

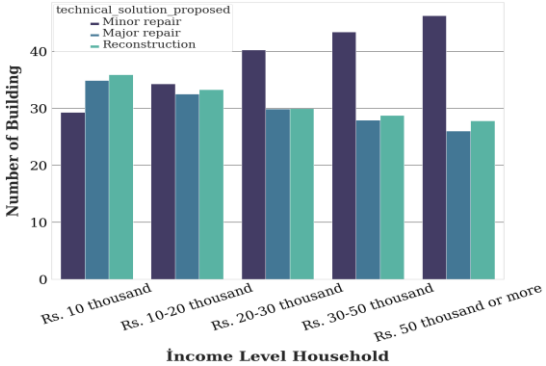


Figure 12. Income Level Distribution

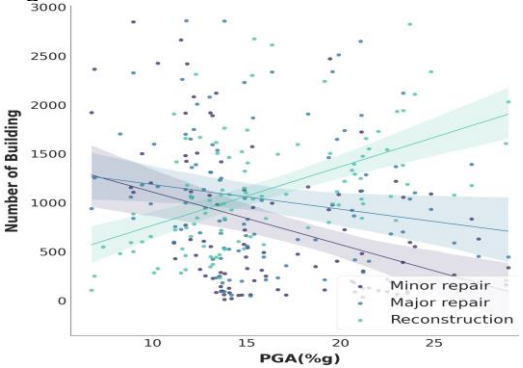


Figure 13. PGA and Classes Relation

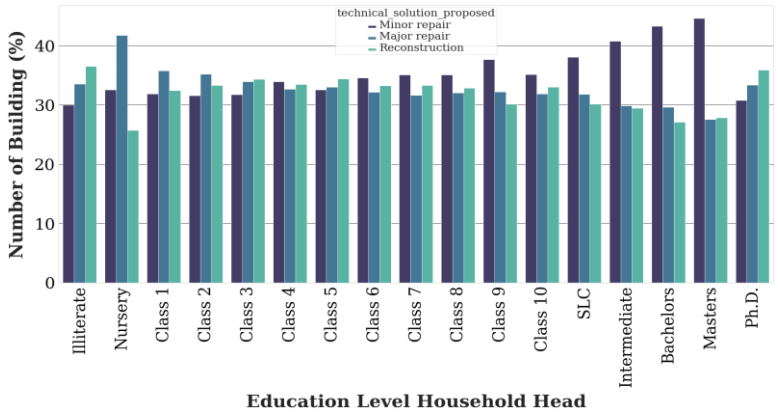


Figure 14. Education Level Household Head Distribution

2.3. Machine Learning Algorithms

In this study, Random Forest Classification and XGBoost algorithms were used as machine learning algorithms. These algorithms were chosen because the data tags used are multi-class and there are too many assets. In addition, these algorithms are algorithms that have often proven their success in the literature [13]. Both algorithms use gradient boosting. Gradient Boosting tries to create a strong predictor by combining weak models while minimizing the loss function using gradient descent in the model.

2.3.1. Random Forest

Random forest is an ensemble algorithm and it is formed by the combination of multiple decision trees. The working logic is based on generating subsets by choosing random samples from the training set. It creates decision trees from these clusters. It votes by making predictions for the samples selected from these decision trees and increases the scores of the trees whose predictions are correct. Random forest algorithms include parameter values such as number of estimators, min samples split, max depth, and min samples leaf [14].

2.3.2. XGBoost

It is an ensemble method like Random Forest, but it uses the gradient boosting algorithm it uses more efficiently. Unlike the Random Forest algorithm, it is more resistant to overlearning. If the trees used to give the correct prediction with similar examples, the model can overlearn. To prevent this, he prunes trees with a similarity score. By subtracting the similarity ratio of a node and the similarity ratio of the structure (child) under the node, the gain of that node is obtained. If this difference is low, the tree stops deepening. With XGBoost, trees are created in parallel, not sequentially, so the wider community can train its trees faster [15].

2.4. Models Validation and Testing

To improve the performance of the models and ensure their reliability of the model, the hyperparameters of the models were first selected with five-fold cross-validation. Then the models were tested with the selected parameters. For this, 80% of the data set (300,203 samples) created was used in the training and validation of the model. The remaining 20% (75051 samples) of data was used only to test the models. The parameters selected for the models are given in **Error! Reference source not found.**.The fact that there are too many features in the created models, especially in the data containing both structural and demographic characteristics, negatively affects

the flexibility of the use of the model. Since the purpose of the models is to predict the effects of future earthquakes, reducing the number of features will facilitate the acquisition of new data. For this reason, new models were created by removing the features with low effect in the models.

The Random Forest model found numerical characteristics to be more important. The XGBoost model has attached considerable importance to categorical data and the Mud-Mortar Stone feature among them [Figure 15Figure ***16***]. The results of the models were compared using the best 40 features of both models.

Table 3. Model Parameters

Models	Tried Parameters	Best Score Parameters
Random Forest	*Number of Estimators : [10, 100, 500, 1000],* *Max Depth : [None, 2, 5, 8, 10],* *Min Samples Split: [2, 5, 10]*	*Number of Estimators: 100,* *Max Depth: None,* *Min Samples Split: 2*
XGBoost	*Number of Estimators : [500, 1000,2000],* *Max Depth : [None, 2, 5, 6, 8, 10],* *Min Samples Split: [2, 5, 10] ,* *Learning Rate : [0.1 0.01, 0.02, 0.05],* *Subsample : [0.6, 0.8, 1.0]*	*Number of Estimators:2000,* *Max Depth: 6,* *Min Samples Split: 2,* *Learning Rate: 0.1,* *Subsample: 0.8*

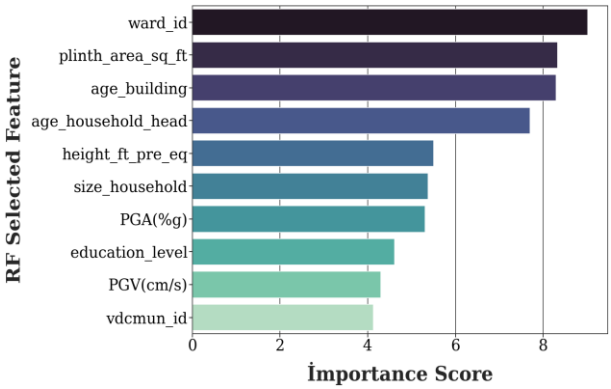


Figure 15. Top 10 Features of Random Forest

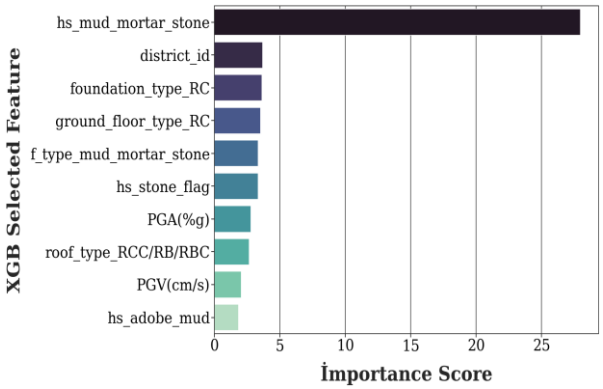


Figure 16. Top 10 Features of XGBoost

2.5. Model Evaluation Metrics

In machine learning, if the prediction task is to identify a categorical class, this task is called the classification task. The task in this study is a multi-class classification task, as the prediction label contains more than two classes. Many metrics have been put forward to test the capabilities of multiclass classification models. In this study, we will discuss the performance of the models using the Accuracy Rate and Precision & Recall values.

The accuracy rate is one of the most used metrics in classification problems. Accuracy is a general measure of how accurately the model predicts on the test set [16]. It is calculated as the ratio of the number of correctly guessed samples in all classes to the number of tests used. It is a more useful metric for problems that are balanced and all tags have the same severity rating. In this study, the accuracy rate alone is not a measure of performance, as the importance levels of the tags are not equal.

The Reconstruction class on forecast tags is a more important class than the None/Minor Repair and Major Repair classes. Because the buildings belonging to this class belong to the buildings that have been heavily damaged in the earthquake and are hazardous . Identifying buildings belonging to this class before an earthquake occurs will be the most important step in preventing loss of life and property. It is desirable to predict these classes as high as possible in the models. For this reason, Precision and the more important Recall values were calculated to compare this class performance. Briefly, if we call the Reconstruction class, which is the class we are interested in, positive and the other classes are negative, True Positive value is the number of correct predictions of the Reconstruction class, False Positive is the number of predictions that are actually in the negative class but predicted as a positive class, False Negative is classes that are actually positive classes but are predicted as negative classes. The metrics are calculated as in Eq. 1-Eq**. 2**.

$$Precision = TP / (TP + FP) \qquad \textbf{Eq. 1}$$

$$Recall = TP / (TP + FN) \qquad \textbf{Eq. 2}$$

3. Results and Discussion

The trained machine learning models were tested with approximately 75,000 sample buildings. The test results are shown in Table 4. All Model Results. Model performances were tried to be determined by the accuracy rate showing the average prediction success and the recall values showing the prediction success of the Reconstruction class.

Table 4. All Model Results

	Algorithms					
	Random Forest			**XGBoost**		
	Accuracy	**Reconstruction Precision**	**Reconstruction Recall**	**Accuracy**	**Reconstruction Precision**	**Reconstruction Recall**
I. Dataset (43 Features)	66.93%	74.26%	72.61%	69.19%	75.75%	73.97%
II. Dataset (79 Features)	70.83%	77.25%	76.36%	70.72%	77.40%	75.92%
I. Dataset (Top 30 Features)	66.90%	73.30%	72.34%	68.65%	75.36%	73.05%
II. Dataset (Top 40 Features)	70.68%	77.48%	76.22%	70.26%	77.28%	75.32%

When compared in terms of the data set used, the prediction success of the models that do not use demographic features lagged behind other models. In other words, demographic information contributed to model success. The highest accuracy rate is the model in which the Random Forest algorithm and all the features are used. The accuracy rate of this model is 70.83% and the Reconstruction class Recall value is 76.36%. The model that followed this model as a success was the XGBoost model, which used the same data set and gave 70.72% accuracy and 75.92% Recall value.

The models created by feature selection provided close prediction successes with the best models, although fewer features were used. The Random Forest model, which uses the top 40 features, outperformed XGBoost by a small margin with an accuracy rate of 70.68% and a Recall value of 76.22%.

Another important parameter when comparing the performances of the models is that a building belonging to the Reconstruction class is mistakenly included in the None/Minor Repair class. The loss that occurs here can be thought of as the loss that would occur if a sick person was mistakenly appointed as healthy. Even if the model's Recall value is high, this wrong estimation may result in a greater financial loss against the other model. Confusion matrices of the models created by feature selection are given in Figure 17Figure **18**. As can be seen in these models, even though the number of correct predictions for the Reconstruction class in the Random Forest model is higher than XGBoost, the None/Minor repair prediction as an incorrect prediction remained somewhat high in XGBoost. For this reason, which algorithm will be successful may vary depending on the financial losses caused by incorrect predictions.

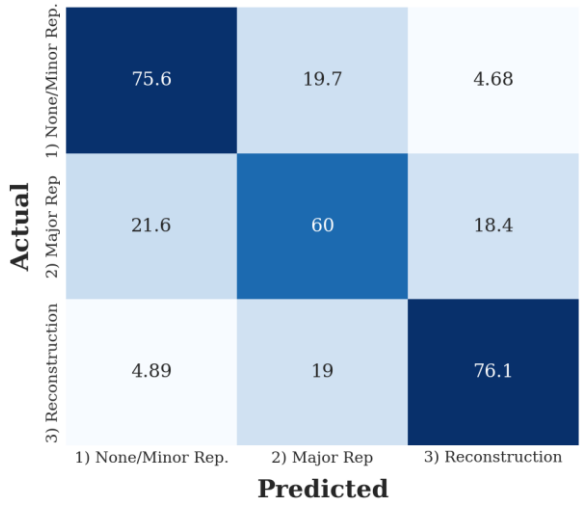


Figure 17. II. Data Top 40 Features RF Model

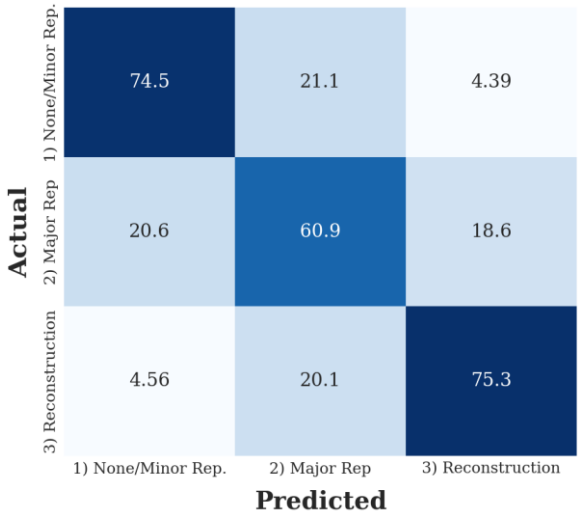


Figure 18. II.Data Top 40 Feature XGBoost Model

Conclusion

As a result, in this study, machine learning models were created that predict the results of a previous earthquake using various structural, geotechnical, and demographic information. In the models created, the effects of demographic information on forecasting success were examined. It has been observed that models with demographic characteristics produce more successful predictions. Considering that Nepal is one of the least developed countries, including demographic features in the models may have a more decisive influence in models on the country's economic classes may be observed clearly. The model tests were repeated with the best 40 features selected by the models, and the results were close to the models in which all the features were used. Even though the two algorithms used have advantages over each other, the Random Forest model gave the highest Reconstruction class prediction accuracy with 76.1%. The fact that the model is fast and can make damage estimation based on fewer features has proven that it can be used as an alternative to traditional methods. Whether the success rates of the models are sufficient in real applications can be determined by detailed loss analysis. The performance of machine learning is highly dependent on the datasets used, so collecting larger damage datasets will allow future damage prediction models to make predictions with less loss.

References

[1] K. A. R. V. D. Kahandawa, N. D. Domingo, K. S. Park, and S. R. Uma, "Earthquake damage estimation systems: Literature review," *Procedia Engineering*, vol. 212, pp. 622–628, 2018, doi: 10.1016/J.PROENG.2018.01.080.

[2] S. Mangalathu, H. Sun, C. C. Nweke, Z. Yi, and H. v. Burton, "Classifying earthquake damage to buildings using machine learning:," *https://doi.org/10.1177/8755293019878137*, vol. 36, no. 1, pp. 183–208, Jan. 2020, doi: 10.1177/8755293019878137.

[3] S. Mangalathu, H. Sun, C. C. Nweke, Z. Yi, and H. v. Burton, "Classifying earthquake damage to buildings using machine learning," *Earthquake Spectra*, vol. 36, no. 1, pp. 183–208, Feb. 2020, doi: 10.1177/8755293019878137.

[4] Y. Xie, M. Ebad Sichani, J. E. Padgett, and R. DesRoches, "The promise of implementing machine learning in earthquake engineering: A state-of-the-art review:," *https://doi.org/10.1177/8755293020919419*, vol. 36, no. 4, pp. 1769–1801, Jun. 2020, doi: 10.1177/8755293020919419.

[5] B. Adhikari *et al.*, "Earthquakes, Fuel Crisis, Power Outages, and Health Care in Nepal: Implications for the Future," *Disaster Medicine and Public Health Preparedness*, vol. 11, no. 5, pp. 625–632, Oct. 2017, doi: 10.1017/DMP.2016.195.

[6] R. K. Adhikari and D. D'Ayala, "2015 Nepal earthquake: seismic performance and post-earthquake reconstruction of stone in mud mortar masonry buildings," *Bulletin of Earthquake Engineering*, vol. 18, no. 8, pp. 3863–3896, Jun. 2020, doi: 10.1007/S10518-020-00834-Y/FIGURES/29.

[7] "About the project." http://eq2015.npc.gov.np/docs/#/about (accessed Aug. 11, 2022).

[8] "Introduction — NHRP Open Data Portal 1.0.0 documentation." https://open-hrrp.readthedocs.io/en/latest/1_introduction.html (accessed Aug. 07, 2022).

[9] "2015 Nepal Earthquake: Open Data Portal." https://eq2015.npc.gov.np/#/ (accessed Aug. 09, 2022).

[10] "USGS Earthquake Hazards Program." https://earthquake.usgs.gov/ (accessed Aug. 07, 2022).

[11] A. Estabrooks and N. Japkowicz, "A mixture-of-experts framework for learning from imbalanced data sets," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2189, pp. 34–43, 2001, doi: 10.1007/3-540-44816-0_4/COVER.

[12] J. Brownlee, "Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python," 2020, Accessed: Aug. 07, 2022. [Online]. Available: https://books.google.com/books?hl=tr&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=Jason+Brownlee+data+pre&ots=Cl5NAjeOrU&sig=stVN3hXjy-CvvuykeUVGs6vPuCc

[13] S. Jhaveri, I. Khedkar, Y. Kantharia, and S. Jaswal, "Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns," *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, pp. 1170–1173, Mar. 2019, doi: 10.1109/ICCMC.2019.8819828.

[14] L. Breiman, "Random Forests," *Machine Learning 2001 45:1*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[15] "XGBoost Documentation — xgboost 2.0.0-dev documentation." https://xgboost.readthedocs.io/en/latest/ (accessed Aug. 07, 2022).

[16] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," Aug. 2020, doi: 10.48550/arxiv.2008.05756.