International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 3 ǁ March 2023 ǁ PP. 108-116

Machine Learning Techniques for Forecasting the Damages Caused to BuildingsDue to Earthquake

Dr.J. Ravi kumar[1],Dr.Jyothi R[2]
Assistant Professor, Dr.Ambedkar Institute of Technology,VTU[1]Associate Professor, Global Academy of Technology,VTU[2]

Abstract---Earthquake is the deadliest natural disaster known to humankind. An earthquake is defined as an unanticipated violent tremble of the ground or earth's surface due to the sudden release of a large amount of energy. Earthquakes are normally created when the rock below abruptly breaks along a fault. This impromptu release of energy creates seismic waves that cause the ground to shake. Earthquake destroys both lives and properties. In fact, earthquakes kill more people than all the other natural disasters put concurrently. Many lives are lost due to the poor quality of building structures.

Our paper is based on an earthquake that occurred in Nepal in 2015. We are using data that was collected by the National Planning Commission Secretariat Kathmandu using the Central Bureau of Statistics and Living Labs of Kathmandu's surveys. The goal of our paper is to build an accurate Machine Learning which would predict the damage level a building would suffer if an earthquake would happen. We have tried to implement different Machine Learning models. We are considering an F-1 score to get the accuracy of models. We have also taken training and testing time for consideration. Our endgame is to use that model to predict what kind of buildings are safer.

The process begins with first collecting the data and applying Exploratory Data Analysis to it. Exploratory Data Analysis assists us to interpret the data well. By using Exploratory Data Analysis we were able to find the relationships between the parameters. We also found out about the outliers in Age, Height percentage and area percentage. To remove these outliers we have used the Winsorization technique. This technique helps to spread the data evenly. After winsorization, we trained the following models, Decision Tree, Random Forest, Gradient Boosting Classifier, XGBoost, Logistic Regression, Lightgbm, KNN, Adaboost.

From the aforementioned models, the top 3 models were XGBoost, Gradient Boosting Classifier, Lightgbm with 74.73%, 74.54% and 74.52% accuracy respectively. Total training time taken by these models were 1389 seconds, 5980 seconds and 1606 seconds respectively.

Date of Submission: 01-03-2023 Date of acceptance: 11-03-2023

I. INTRODUCTION

WHEN The earth's surface shakes, the phenomenon is referred to as an earthquake. Earthquake is regarded as one of the deadliest natural disasters. Immense damage and loss of property are caused by earthquakes. The most dangerous thing about an earthquake is that it is very unpredictable. Some earthquakes are not even felt, while few earthquakes kill thousands of people and cause billions of dollars ofdamages. The intensity of an earthquake is measured by Richter's scale. Generally, earthquakes occur due to the movement of tectonic plates under the earth's surface. Other causes of earthquakes are volcanoes, mines and nuclear blasts. On 25th April 2015 tectonic earthquake shook Nepal, parts of Bangladesh, India and China. The epicentre was near Lamjung, nearly 50 miles northwest of Kathmandu. It was measured to be 7.8Mw on Richter's scale. It was not a single earthquake. The ground under was not stable and daily there were many aftershocks. The most severe was on May 12th, it measured to be 7.3Mw on Richter's scale. These earthquakes are also called Gorkha earthquakes. This catastrophic event had killed nearly 9000 people, about 22,000 injured and many missing. More than 600,000 homes were decimated. The Nepalese economy became unstable. It is estimated that the earthquake had caused 5 to 10 Billion Dollars of damage.It had destroyed many Ancient and religious places. Mountain expedition to Mount Everest was halted.[14][15]

The whole world came to rescue Nepal, about 3 Billion dollars were pledged by donors. India was the first nation to respond. It responded with few hours, it deployed its militaryto provide rescue and relief operations. India is the largest donor to Nepal. It had given over a Billion dollar aid. India helped in the evacuation of both Indians and foreigners. China, Asian Development Bank and the United Kingdom were large donors.[16]

II. LITERATURE SURVEY/RELATED WORK

Earthquakes have been researched extensively. There have been many routes taken like Sertel et al.[1] (2007) used semivariograms to recognize and assess damages caused by the earthquake in urban areas. They worked on the 1999 Izmit earthquake which was devastedby the 7.4 Mw earthquake. It affected urban areas of Izmit, Adapazari (Sakarya), Golcuk, and Yalova. A semivariogram method was used to evaluate earthquake-induced spatial variation and thereby the degree of damage. The differences between pre-earthquake and post- earthquake semivariogram helped to measure the severity of the calamity.

Hosokawa et al. [2] (2009) are about earthquake intensity estimation and damage detection, the methodology they follow is using remote sensing data. They predict the damage by using the magnitude of the earthquake, location, ground condition, distance attenuation equation together with change detection using multi-temporal SAR data.

Dezhang Sun et al. [3] (2009) used fuzzy mathematics to estimate the damage caused by an earthquake. This was a unique approach to forecast the damage. They relied upon the seismic risk index as metrics to earthquake damage and the cumulative seismic damage index was derived from the return index, the impact factors came from the shift index, impact factors were used as modification index. They implemented random sampling and closeness functions.

Hidenori Kawabe et al. [4] (2008) worked on Nankai and Tonankai earthquakes that occurred on the Nankai Trough. They tried to forecast using the empirical Green's function method. They worked on broad-band powerful ground motions. They used 3-D finite-difference simulation. Their results concluded that investigation is required for high-rise buildings and base-isolated to check if they are seismic-safe.

Lee et al. [5] (2018) work is based on Wenchuan Earthquake. They used the technique nonlinear regression method which is a new method where closeness degree is given. The damage grade is also taken into consideration for finding member function. This method greatly simplifies the calculation.

H Chen et al. [6] (2016) did extensive research onthe Nepal earthquake and its damages. They collected various data points from the tectonic summary, seismic course and strong motion data. They saw that Nepal's building infrastructure was underdeveloped and lacked asteel industry. They saw that number of buildings having mud bounded bricks was 50% more than cement bricks. Next, they found out that mud bounded masonry was the worst affected with 36% of them collapsed and 23.6% of them were severely damaged whereas only 7% of cement bounded masonry collapsed and 22% were severely damaged. R C frame was the best only 1.38% were collapsed and 8% were severely damaged.

Katsuichiro Goda [7] (2015) et al. analysed important results of damages caused by the earthquake. They analysed all the seismological data, damages to the building. They also recorded the complete rupture process, aftershock data. They submitted data to U.S. Geological Survey (USGS).

Sourav Adi (2020) [8] et al. worked on the Nepal earthquake. They predicted the damage caused by the earthquake using the Random Forest classifier and Gradient Boosting Classifier. They also used hyper-parameter tuning to improve the accuracy.

Khaleed Talab et al. [9] (2018) developed a nascent method of data mining for the development of Landslide SusceptibilityMaps.

These were used in all the areas which were highly prone to landslides. They used the Random Forest algorithm to producemore effective maps.

Anand et al. [10] (2017) used deep learning synergistically. They took 3D point cloud features from high-quality images. They also implemented multiple kernel learning. They used convolutional neural networks, 3D features both separately and together. They found that convolutional neural network alone gave 91% average accuracy but pairing it with 3D point increased it to 94%. It was concluded that integrating them both was a practical decision.[11][12] are few other examples of other disasters.

III. DATA RESOURCES

The CSV file contains 40 columns and 260602 rows. This dataset contains more than 1 crore data points. It is one of the largest post-disaster datasets ever collected. It contains both binary and categorical data. This file was derived from the driven data website. Kathmandu Living Labs and the Central Bureau of Statistics surveys helped in collecting the data. The file contains a unique id for each building, it is called building_id.

Damage_grade is the target value, it is of 3 levels(1,2,3). 1 means low damage, 2 means moderate damage, 3 means severe damage. This dataset contains 38 parameters. It has the number of floors, age, area percentage, height percentage, land surface condition, type of foundation, roof type. It also has the information on whether the building was using superstructure_timber or hassuperstructure_mud_mortar_stone etc. It also has details of usage like if it was a just house or had secondary usage like school, hotel, industry, agriculture etc.[13]

IV. METHODOLOGY

In this project, we are following a classic approach. Weare going to collect the data first, then process the data. After processing and cleaning the data were are going to analyse the data to find the patterns and relationships among the parameters. It will enable us to improve the model. Then we are going to divide the data for testing and training in a specific ratio. After splitting the data, we are going to build a model and train it using the training data. Then we feed it testing data and get the prediction. We calculate the accuracy and then we tune the model. We apply the same approach for all the different algorithms and get the final output. Finally, we would tabulate the results and come to a conclusion of which algorithm is better.

A. Pre-Processing the Data

After getting the data, we will process it using Matplotlib, Pandas and Seaborn. Using Pandas and the data frame we found out that there are no missing values in the data. Also, there were no duplicate data. The data which we have got were completely cleaned. CSV file more than 79 MB. 32 parameters were integer type and 8 parameters were object type.

B. Exploratory Data Analysis

Exploratory data analysis is a statistical procedure where we perform initial investigations on data so as to discover patterns, spot anomalies, test a hypothesis and find a relationship between the attributes. Here our data is highly imbalanced, damage of type 2 (moderate damage) is more than 56%, damage of type 1(little damage) is about 33.47% and damage of type 3(severe damage) is less than 10%.Figure 1.

Then we find if there is any correlation between any of the columns to see if something stands out explicitly. From Figure2, we can see there are not a lot of co-related fields, has_secondary_use is correlated with has_secondary_use_agriculture, has_secondary_use_hotel, height_percentage is highly correlated with count_floors_pre_eq and area_percentage and height_percentage are correlated with has_super_structure features and secondary use of buildings. It is important to find correlations otherwise the model will give skewed or incorrect results.

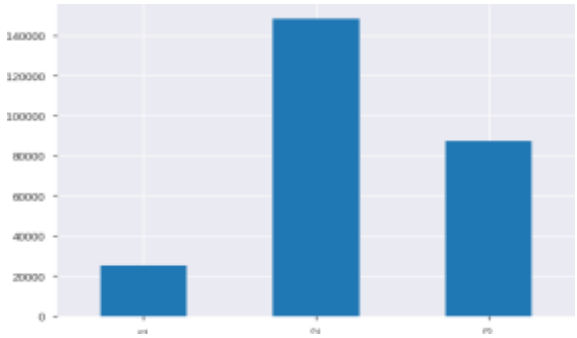

Fig 1 Distribution of Damage Grade.

Fig 2. Correlation between All the Columns.

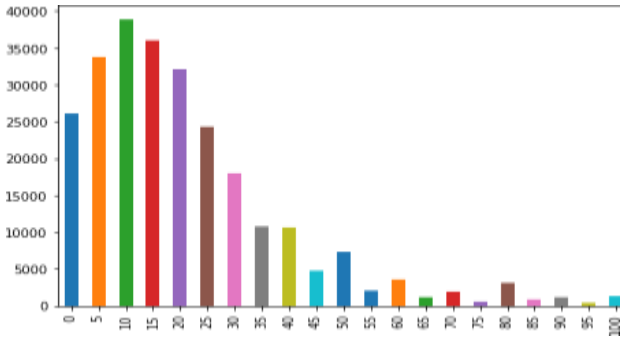

Fig 3. Distribution of Building's Age.

Next, we analyse the impact of age on our dataset. We can clearly see that there is more number of newer buildings than older ones. Figure 3 and 4 We can also see that damage1 was more than damage 3 in buildings ageing between 0 to 5, in every other age group it is opposite, it implies newer buildingswere sturdier.

We also analyse the impact of building material on the damage. The top 5 types of structures that got damaged the most were made up of timber, bamboo and some form of mud.stone_flag, cement_mortar_stone and rc_engineered type were affected the least. This show that these buildings either withheld the earthquake well or they were far away from the epicentre. Figure 5

We also see if there are any relationships among them. Bamboo and timber are correlated so one of them can be dropped before feeding the data into the ML model. Many columns are negatively correlated, the first column and third column are positively correlated so one of them can be dropped as well. Figure 6.

Similarly, we see how height percentage and area percentage affect the damage. We see that the distribution is not equal, there are few outliers in both. Finally, we look into foundation type, surface condition, roof type and others'impacton the type of damage. Figure 7 and Figure 8.

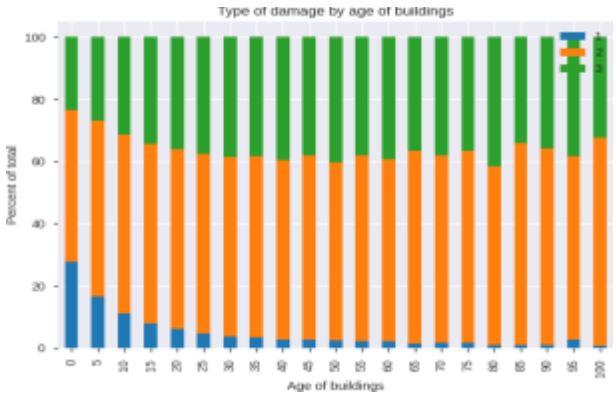

Fig 4: Distribution of Building's Age with Respect to Damage

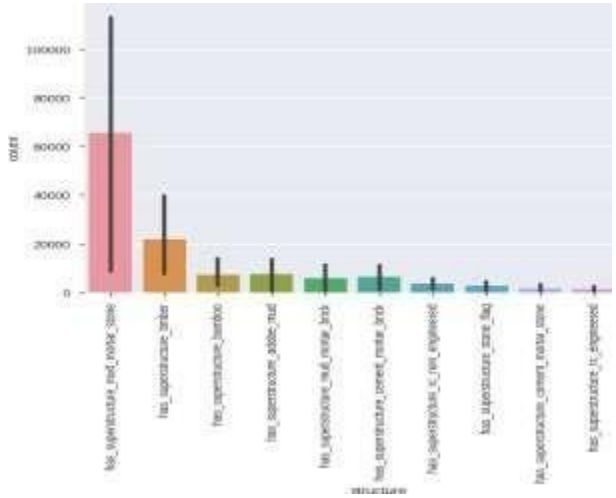

Fig 5. Type of Building Material.

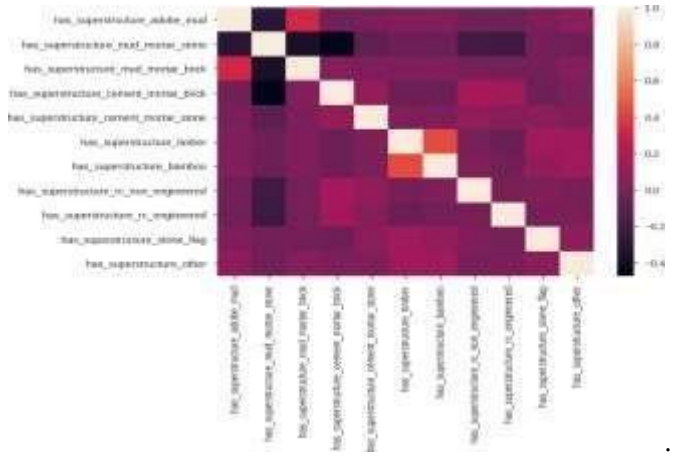

Fig 6. Correlation among Types of Materials

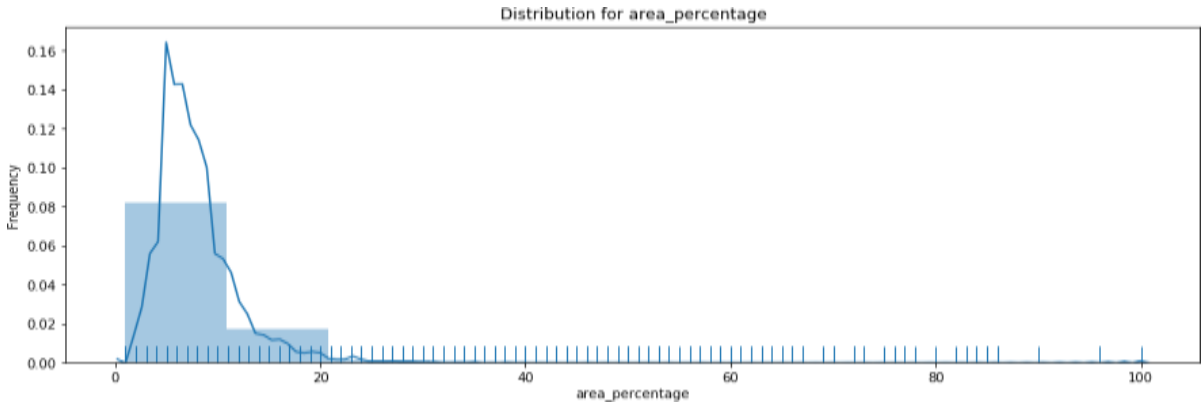

Fig 7. Distribution of Area Percentage

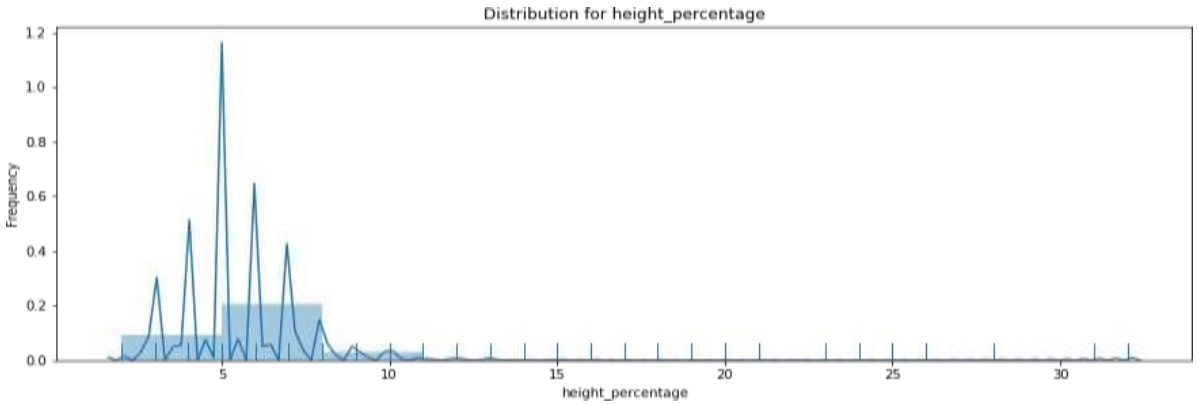

Fig 8. Distribution of Height Percentage

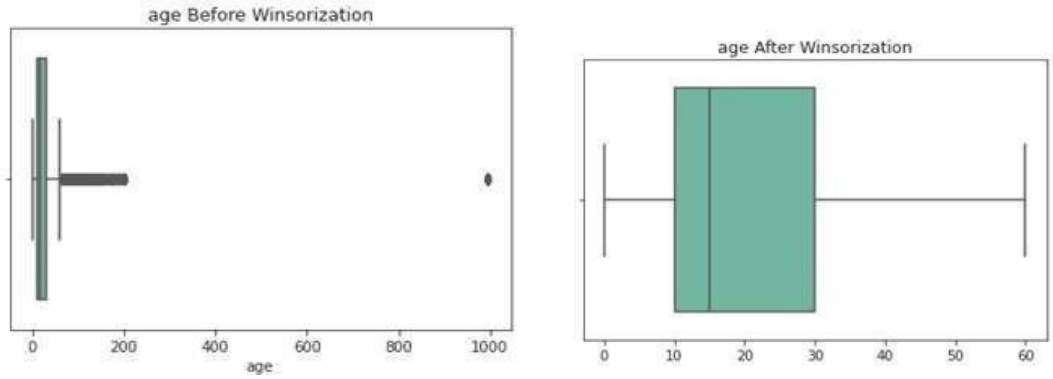

Fig 9. Winsorization of Age

C. Winsorization

Winsorization is the process of transforming the statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. It is a way to minimize the influence of outliers in the data. Here in our data age, area_percentage and height_percentage have outliers, we apply 5% winsorization to the data and remove the outliers. Winsorization has a different effect on different algorithms. For some it increases accuracy for others it decreases accuracy[19].Figure 9.

V. BUILDING THE MODEL

Here we train a model using an algorithm to predict the damage using all the features. Since all the data is not in integer format we have to transform the data through a process of encoding. Before we build a model we have to divide the data into two parts, one for training and the other for testing. In our project, we have divided it 80:20.

We have used Decision Tree, Random Forest, Gradient Boosting, XGBoost, KNN, Lightgbm, Logistic Regression and Adaboost to train the models. We have used GridSearchCV, RandomizedSearchCV to tune parameters.It helped in getting better accuracy. Winsorization increased the accuracy of KNN, XGBoost, Logistic Regression anddid not help other models.

VI. RESULTS

There are various ways to find the correctness of a model, like Accuracy, Recall, Precision, F1 score. We have used both Accuracy and micro averaged F1 scores. Accuracy is denoted as the fraction of right forecasts for the given test data. It can be obtained simply by dividing the number of accurate predictions by the cumulative predictions. F1 Score is the weighted average of Precision and Recall. Hence, this score considers both false positives and false negatives into consideration. Usually, the F1 score is used to judge the performance of a binary classifier, but since we have three plausible descriptions we will use an alternative called the micro averaged F1 score. In our project, the values of the F1 score and Accuracy were almost identical.

$$F_{micro} = \frac{2.\,P_{micro}.R_{micro}}{P_{micro} + R_{micro}}$$

Where

$$P_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{K=1}^{3}(TP_k + FP_k)}$$

$$R_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{K=1}^{3}(TP_k + FN_k)}$$

Here, TP means true positive, FP means false positive and FN means false negative.K=1, 2, 3 are the classes

First, we began our project by training the model using the Decision Tree classifier. We used GridSearchCV to tune the model to get better accuracy. We used depth 30, with min_samples_split equal to 3. We did not use winsorization to it. It gave an accuracy of 71.35%, it just took 2.47 s to train. It is the fastest model. Then we implemented Random Forest using RandomizedSearchCV without winsorization. The number of trees was set to 400. It gave an accuracy of 73.64 %. It took 527 seconds to train the model.[17]

Then we tried Gradient Boosting Classifier using warmstart, number of trees with 300 and depth of 30. It gave an accuracy of 74.54%. It took more than 1550 seconds to train. Then XGBoost was tried, here winsorization helped in increasing accuracy to 74.73% from 74.62%.[18] Next, we tried KNN with Manhattan Distance and Minkowski metrics. It gave 71.98% accuracy. It suffered from overfitting. It took less than 4 seconds to train. Then we tried Lightgbm, it took 5980s (about 100 min) to train but gave 74.52% accuracy. Here Winsorization did not help. Then we tried Logistic regression with Multinomial multiclass and the Newton-cg solver gave 59.16%. At last, we worked on the Adaboost classifier. Here we trained the Adaboost classifier on the base estimator of Gradient Boosting Classifier. It gave an accuracy of 72.21%. But it took 7 hours 24 minutes to train. We had also tried SMOTE and undersampling since the data is highly imbalanced but itdecreased the accuracies of the models.

TABLE I
RESULTS

MODEL USED	PARAMETERS USED	Accuracy	Training Time (s)
Decision Tree	Max features = None Maximum depth=30 Min samplesSplit=3 Min samplesLeaf=16	71.35%	2.47
Random Forest	N estimators=400 max_features = None Max depth=30 min_samples_ Split=3 Min_samples_ Leaf=16 N_jobs=4	73.64%	527
Gradient Boosting Classifier	max_depth = 10 n_estimators = 300 warm_start =True	74.54%	1561
XGBoost	n_jobs=-1 n_estimators= 600 max_depth= 10 learning_rate = 0.15	74.74%	1389

TABLE II
RESULTS

MODEL USED	PARAMETERS USED	Accuracy	Training Time (s)
KNN	n_neighbors =18 weights = distanceMetric =MinkowskiP=1	71.98%	3.88
LightGBM	N_estimators =400 num_leaves=65 objective=multiclassboosting_ type=dart N_jobs=6	74.52%	5980
Logistic Regression	multi_class =multinomialSolver= Newton-cg	59.16%	184
AdaBoost Classifier	base_estimator= Gradient Boosting Classifier	72.22%	26579

VII. CONCLUSION

This study has successfully built models which can predict damages to buildings caused by the earthquake. We have examined various algorithms and learnt that XGBoost was the most accurate model with 74.74% accuracy, Gradient Boosting and LitghtGBM were very close. We see that ensemble models gave better results than other types of models. The Decision tree and KNN were very quick and gave decent accuracy. These models can be used if time a constraint. We also saw the undersampling and oversampling does not help.

Further, we can do research if these models can be applied to various other natural disasters.

REFERENCES

[1]. Elif Sertel, Sinasi Kaya and Paul J. Curran Use of Semivariograms to Identify Earthquake Damage in an Urban Area, 2017
[2]. M Hosokawa, B. P Jeong, and O Takizawa. Earthquake intensity estimation and damage detection using remote sensing data for global rescue operations, 2009.
[3]. D Sun and B Sun. Rapid prediction of earthquake damage to buildings based on fuzzy analysis, 2010.
[4]. Kawabe Hidenori, Kamae katsuhiro, and Irikura Ko- jiro. Damage prediction of long-period structures during subduction earthquakes -Part 1: Long-period ground motion prediction in the Osaka basin for future Nankai Earthquakes, 2008.
[5]. Lee J, Lin J, and Liu J. An Improved Method for Earthquake Damage Prediction of Highway Subgrade and Pavement Based on Closeness Degree Method,2018.
[6]. Hao Chen, Quancai Xie, Zhiqiang Li, Wen Xue and Kang Liu. SeismicDamage to Structures (2015)
[7]. T K Katsuichiro Goda "The 2015 Gorkha Nepal earthquake:insights from earthquake damage survey". Frontiers in Built Environment, 2015.
[8]. Sourav Pandurang Adi, Vivek Bettadapura Adishesha, Keshav Vaidyanathan Bharadwaj and Abhinav Narayan "Earthquake Damage Prediction Using Random Forest and Gradient Boosting Classifier" 2020.
[9]. Khaled Taalab, Tao Cheng, and Yang Zheng "Mapping landslide susceptibility and types using Random Forest". Big Earth Data, 2 (2), 2018.
[10]. Anand Vetrivel, Markus Gerke, Norman Kerle and Francesco Nex "Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning" 2017.
[11]. RouetĜLeduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A. (2017). Machine learning predicts laboratory earthquakes. Geophysical Research Letters, 44(18), 9276-9282.
[12]. Martínez-Álvarez, F., Troncoso, A., Morales-Esteban, A., & Riquelme,
[13]. J. C. (2011, May). Computational intelligence techniques for predicting earthquakes. In International Conference on Hybrid Artificial Intelligence Systems (pp. 287-294). Springer, Berlin, Heidelberg
[14]. .https://www.drivendata.org/competitions/57/nepal- earthquake/.
[15]. https://en.wikipedia.org/wiki/April_2015_Nepal_earthqua ke
[16]. https://www.britannica.com/topic/Nepal-earthquake-of- 2015
[17]. https://www.worldvision.org/disaster-relief-news-stories/2015-nepal- earthquake-facts

[18]. https://xgboost.readthedocs.io/en/latest/
[1] https://scikit-learn.org/stable/
[2] Alan Reifman and Kristina Garrett "Winsorize". En- cyclopedia ofresearch design, pages 1636–1638, 01 2010.