Analysis and Prediction of Earthquake Impact-a Machine Learning approach

Anmol Gaba, Arnab Jana, Rahul Subramaniam, Yash Agrawal, Merin Meleet
Department of Information Science and Engineering
R.V College of Engineering
Bengaluru, India
yashagrawal.is16@rvce.edu.in

***Abstract*—An earthquake is a natural disaster known on account of the devastating effect it has on naturally occurring structures and manmade structures such as buildings, bungalows and residential locations to name a few. Earthquakes are measured using seismometers, that detect the vibrations due to seismic waves travelling through the earth's crust. In this work, the damage that is caused by an earthquake was classified into damage grades, ranging in values from one to five. A previously acquired data set was used, wherein a series of parameters were taken into consideration to predict the damage grade of a given building, which is associated with a Unique Identification String. The prediction was done using a survey of existing machine learning classifier algorithms. The machine learning algorithms used in this work were Logistic Regression, Naive Bayes Classifier, Random Forest Classifier and K-Nearest Neighbors. Based on an evaluation of a set of attributes, the most appropriate algorithm was considered. A detailed analysis was done on the predicted attribute by the given algorithm, followed by data analysis that provided details that could help mitigate the impact of an earthquake in future.**

Keywords—predictive analysis, Logistic Regression, K-Nearest Neighbors, Random Forest Classifier, Naive Bayes Classifier, Machine Learning.

I. INTRODUCTION

An earthquake is a calamitous occurrence that is detrimental to human interest and has an undesirable impact on the environment. Earthquakes have always caused incalculable damage to structures and properties and caused the deaths of millions of people throughout the world. In order to minimize the impact of such an event, several national, international and transnational organizations take various disaster detection and prevention measures. Time and quantity of the organization's resources are limiting factors, and organization managers face several difficulties when it comes to the distribution of the resources.

Leveraging the power of machine learning is a viable option to predict the degree of damage that is done to buildings post an earthquake. It can help identify safe and unsafe buildings which helps to predict damage prone areas and thus avoiding death and injuries resulting from the aftershock of an earthquake, while simultaneously making rescue efforts efficient.

This is done by classifying these structures on a damage grade scale based on various factors like its age, foundation, number of floors, material used and several other parameters. Then the number of families and the probable casualties ward-by-ward in a district are taken into account. This enables distribution of relief forces proportionately ward-wise and its prioritization based on the extent of damage.

Models of this kind can help save as many lives as quickly as possible and turn out to be an efficient and cost-effective solution. It can be further improved by the inclusion of distribution of resources like food, clothes, medical, monetary supplies based on the extent of human casualties and the damage incurred by the various structures.

II. LITERATURE SURVEY

Damage prediction due to earthquakes and other natural calamities is a field of predictive analysis that is gaining a lot of traction in recent times. A lot of research is being done in this field, as a result of which manifold approaches to predict damage on account of earthquakes have been worked upon and developed.

The prediction of earthquake damage was done using Fuzzy Analysis [5]. It used a parameter called the Seismic Damage Index as a measure of the damage. Other impact factors and modified index were considered to calculate the average damage index.

Another perspective to Earthquake damage prediction was done using Artificial Neural Networks [12]. This approach dealt with predicting seismic damage done to multi storied buildings by combining Geographic Information Systems and Artificial Neural Networks. A value known as peak acceleration value was utilized to perform the desired computation. Earthquake intensity was another measure that was considered in this given hypothesis.

A second neural network approach to damage prediction dealt with the use of NNARX model which is an improved version of the NARX type of Artificial Neural networks [13]. The input and output parameters dealt with real-time data. Although, this model was implemented in order to predict floods, it can be utilized to predict damage due to other natural disasters on retrieval of relevant data. The study was done with respect to the city of Kuala Lumpur in particular.

Damage prediction was also performed as a two stage process [14]. The first step involved a simple linear regression model, and a neural networks based model in the

978-1-7281-2619-7/19/$31.00 ©2019 IEEE

second step. The model was used to predict damage due to typhoons. The model enabled the prediction of the number of damage distribution poles and lines from the weather forecasts of a typhoon. The prediction was done in every district of the Kagoshima Prefecture.

The approaches mentioned above, dealt with earthquake (or other natural disasters) damage prediction. There are, however, few similar approaches to predict an earthquake itself, or its attributes like the magnitude of the earthquake, time of the earthquake etc. Earthquake prediction was done using Data Analytics, with the help of a map reduce model that was used to locate places with maximum tremors [1]. It was also done with the help of Artificial Neural networks [2], and a Back propagation neural network [3]. The latter was used to determine the magnitude of an earthquake, while the former dealt with the impact of an earthquake. A final approach used weighted factor coefficients to determine the probability of an earthquake [4].

III. Data Description

As shown in the below diagram, the process of obtaining and defining the data is mainly composed of three essential steps: Data Sourcing, Damage State determination and determining the State Variables.

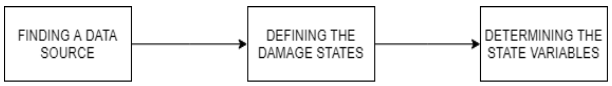

Fig. 1. Flow diagram of the data description step

A. Data Source

The dataset has been fetched from [9]. The dataset includes 4 files, namely the train dataset (used to train the machine learning model), the test dataset (used to test the trained machine learning mode), the file with data about ownership and use case of the building and the file that contains data with respect to the structure and materials of the buildings. The dataset was found to have a total of 1,038,900 records, with the number of records in the test dataset being 421,175, while that of the train dataset was 617,725. The training data and the test data percentages has been observed to be 59.46% and 40.54% (approximated to 60% and 40% respectively).

The extent of damage caused by the earthquake to the buildings was classified from Grade 1 (least damage) to Grade 5 (maximum damage) based on the survey studies and evaluation conducted by appropriate agencies. The primary task was to grade the extent of damage that would likely be caused by an Earthquake (irrespective of the magnitude). This was done considering the definition of magnitude by the United States Geological Survey [10].

B. State Variables

There were many quantifiable variables that accounted for the extent of damage caused by the Earthquake. The two main category of variables are structural variables and non-structural variables. The non-structural variables can be further divided into string factors and boolean factors.

Some of the structural variables are: area of the building assessed, number of families in the building (in square feet), number of floors in the building recorded pre and post the earthquake, age of the building in years, plinth area of the building (in square feet), height of the building before and after the earthquake (in feet).

Some of the non structural string factors considered are: surface condition of the land on which the building was built, type of foundation used in the building, type of roof used in the building, type of the ground floor used in the building, type of construction used in other floors (except ground floor and roof), position of the building, building plan configuration, actual condition of the building after the earthquake.

The non-structural boolean value i.e. true-false value factors are: whether the building was used for any secondary purpose such as agricultural, hotel, rental, institutional, school, industrial, health post, government office, police station; whether the building superstructure type used materials such as adobe - mud, mud mortar – stone, cement mortar – stone, mud mortar – brick, timber, bamboo, RC (non-engineered), RC (engineered).

The above mentioned variables contributed to the necessary parameters that were considered as the attributes for the respective tables in the dataset.

IV. Methodology

This covers the technique and flow of events that were used to perform the prediction process. The prediction methodology itself is composed of three integral steps: data preprocessing, model selection and the final prediction process.

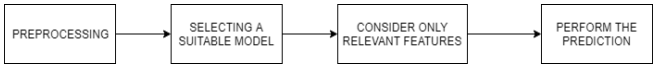

Fig. 2. Flow diagram of the methodology step

A. Preprocessing

In the dataset, a building was uniquely identified by 4 attributes: Building Identification, District Identification, Municipality Identification, Ward Identification. These attributes were added to the training data for identifying the building damage grade.

Many attributes in the files related to the building structure and ownership details of the buildings had string data which were converted to their vectorized representation using Label Encoding technique. E.g. the ownership status in the file with building ownership details is comprised of 3 categories i.e. public, private, others which were converted to 0,1,2 integer data values [11].

There were 33,417 entries in the attribute pertaining to whether building repairs on earthquake affected buildings had started or not, that were found to be blank. Based on the assumption that since there was no formally documented record of the commencement of the repairs, the blank values were assumed to be ‘not repaired’. Such filling was done on the basis of the worst case scenarios possible to get optimal results.

B. Model Selection

For any classification problem, a plenitude of machine learning algorithms are available and thus to choose the best among them, it is necessary to evaluate them on the same data based on a suitable parameter.

Here we chose F1 score with 'weighted' average, as calculated in accordance with the content mentioned in the Hackerearth website [7], as our evaluation metric. The F1 score is a weighted average of precision and recall as mentioned below:

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

Here, TP refers to True Positive, i.e. When the model rightly predicts a positive result.
FP refers to False Positive i.e. When the model wrongly predicts a positive result
FN refers to False Negative i.e. When the model wrongly predicts a negative result.

On running the predefined classification algorithms on the preprocessed data, the following F1 results were obtained, as simulated by the Hackerearth platform

TABLE I. ALGORITHM AND CORRESPONDING F1 SCORES

Algorithm	**Score**
Logistic Regression	0.39167
Naive Bayes Classifier	0.50435
Random Forest Classifier	0.75127
K-Nearest Neighbors	0.62280

C. Prediction

The models developed by the individual algorithms were trained on the training dataset and then test data was used for final prediction of damage grades and for evaluation. Since on evaluation, Random Forest Classifier algorithm was found to possess the highest F1 score, the model was considered for the prediction process.

The feature importance method was used to obtain the importance of the features in our model [15]. Out of all the features that were considered in the initial model, the least important features were dropped, thus producing a more accurate prediction of the damage grade. The threshold value for a feature to be relevant to our model was set to be 0.000200, i.e. any parameter with feature importance less than this value would be dropped from our dataset.

On dropping the following Boolean attributes from the dataset we found an increase in the model score from **0.75127 to 0.76503** – whether the building has secondary use as an institution, whether the building has other geotechnical risks, whether the building has secondary use as a school, whether the building has secondary use as a government office, whether the building has secondary use as a health post, whether the building has secondary use as a police station.

V. RESULTS AND DISCUSSION

The library that has been used to plot the below figures is an open source Python library [8].

The graphs in Fig 3, Fig 4 and Fig 5 compare the number of affected buildings (count) for a particular Damage Grade to their corresponding foundation type, roof type and ground floor type respectively. The corresponding tables Table II, Table III and Table IV depict the ratio of number of buildings of Damage Grade 5 to Damage Grade 1 for a corresponding foundation type, roof type and ground floor type respectively. The ratios indicate the likelihood of buildings with the given material's ability to sustain damage against earthquakes. The ratios are considered here instead of directly comparing the number of affected buildings as the former can take into account the variation observed among the Damage Grades in each case, irrespective of the number of affected buildings.

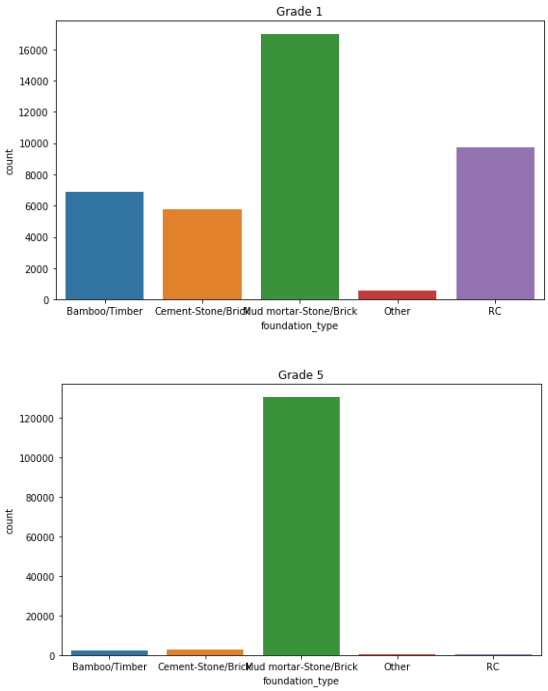

Fig. 3. Graph of number of buildings versus the building material for the building foundation for damage grade 1 and damage grade 5 respectively

TABLE II RATIO OF NUMBER OF BUILDINGS IN DAMAGE GRADE 5 AND DAMAGE GRADE 1: FOUNDATION

Foundation Type	**Count (Grade 5)**	**Count (Grade 1)**	**Ratio (Grade 5 / Grade 1)**
Bamboo/ Timber	2249	6869	0.3274
Cement Stone / Brick	3090	5575	0.535
Mud Mortar – Stone / Brick	130281	16968	7.678
Reinforced Concrete	348	9715	0.0358
Others	756	583	1.296

In Fig 3 it is observed that the buildings having the foundation type 'Mud mortar – Stone/Brick' has the highest

ratio 7.678 from Table II indicating high likelihood of damage. Additionally, it even has the most count among affected buildings of Damage Grade 1 and Damage Grade 5. The foundation type Reinforced Concrete has the least ratio of 0.0358 from Table II, indicating the least amount of damage and the highest sustainability among the materials used. It is also the least frequent among the Damage Grade buildings. Foundation types 'Bamboo / Timber' and 'Cement Stone / Brick' with ratios of 0.3274 and 0.535 from Table II respectively can also be considered as cheaper alternatives to Reinforced Concrete.

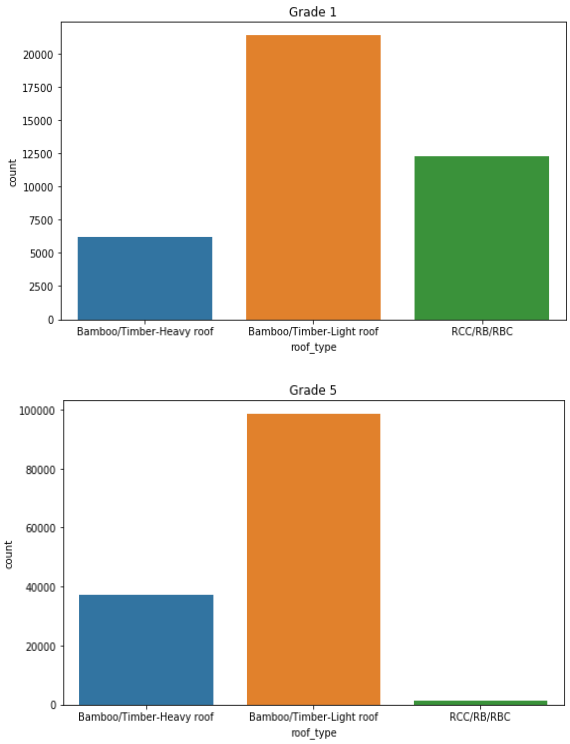

Fig. 4. Graph of number of buildings versus the building material for the building rooftop for damage grade 1 and damage grade 5 respectively

TABLE III RATIO OF NUMBER OF BUILDINGS IN DAMAGE GRADE 5 AND DAMAGE GRADE 1: ROOFTOP

Roof Top Type	Count (Grade 5)	Count (Grade 1)	Ratio (Grade 5 / Grade 1)
Bamboo / Timber - Heavy roof	37167	6229	5.967
Bamboo / Timber - Light roof	98426	21380	4.603
Reinforced Cement Concrete / Reinforced Brick / Reinforced Brick Concrete	1131	12301	0.0919

In Fig 4, it can be observed that the buildings having the roof type 'Bamboo/Timber – Heavy roof' has the highest ratio of 5.967 from Table III depicting the futility of its use against earthquakes. It also has the most building count among both, Damage Grade 1 and Damage Grade 5 graphs. 'Bamboo/Timber – Light Roof' is likely to resist damage more than its heavier counterpart, with a ratio of 4.603 from Table III. RCC/ RB/ RBC or Reinforced Cement Concrete / Reinforced Brick / Reinforced Brick Concrete has the least ratio of 0.0919 from Table III, indicating it to be a good roof top material to sustain the impact of earthquake. It even has the least building count among Damage Grade 5 buildings.

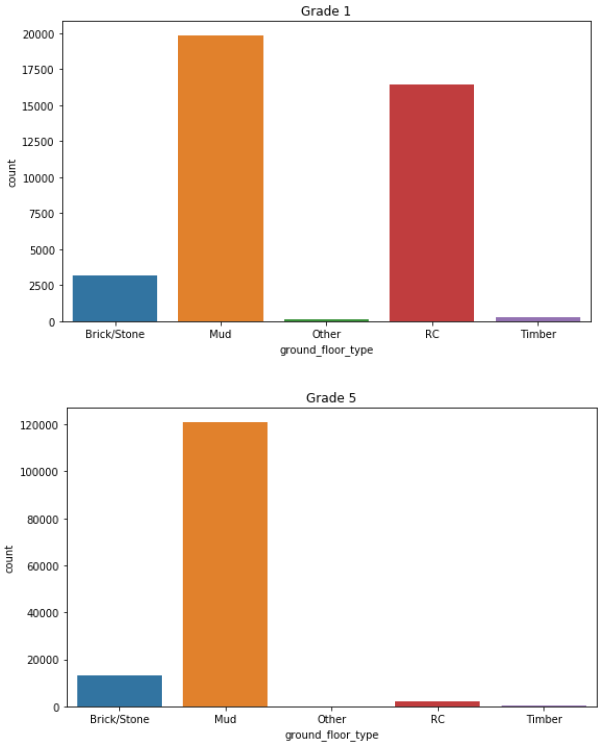

Fig. 5. Graph of number of buildings versus the building material for the building ground floor for damage grade 1 and damage grade 5 respectively

TABLE IV. RATIO OF NUMBER OF BUILDINGS IN DAMAGE GRADE 5 AND DAMAGE GRADE 1: GROUND FLOOR

Ground Floor Type	Count (Grade 5)	Count (Grade 1)	Ratio (Grade 5 / Grade 1)
Brick/Stone	13131	3202	4.1008
Mud	121023	19856	6.0950
Others	114	126	0.9047
Reinforced Concrete	2014	16459	0.1223
Timber	442	267	1.655

The material used in the ground floor is indicative of the material that should be used in the subsequent floors above as well. In Fig 5, it can be observed that for ground floor type 'Mud', the ratio is 6.0950, the highest from Table IV, implying that it is more susceptible to a higher grade of damage. The ratio of ground floor type 'Reinforced concrete' from Table IV is the least i.e. 0.1223, depicting again that it holds up well against earthquake damage. Meanwhile the ratio for the commonly used ground floor type 'Brick /Stone' is 4.704, implying that it is also prone to damage at the time of an earthquake. The ground floor type 'Timber' is comparatively less susceptible to damage, with a ratio of 1.655.

Thus, from the results obtained from the analysis of the above plotted bar graphs, it can be safely concluded that Reinforced Concrete is a material that can be used in the construction of the building foundation, the building rooftop, as well as the building ground floor. Reinforced concrete is a known material in the field of building construction on account of its ability to withstand high tensile stress, particularly when reinforced with steel. This premise is observed in accordance with [6]. It can be reinforced with various materials ranging from Steel to a few other polymers.

VI. Conclusion

This work presents that the Random Forest Classifier algorithm has the highest accuracy in predicting the damage due to earthquakes, based on the F1 score calculated for each of the four algorithms previously mentioned in this work. K-Nearest Neighbors has been observed to be the second most preferred algorithm for earthquake damage prediction. On analysis of the materials that help curb damage to buildings during an earthquake, the work concludes that Reinforced Concrete is the material most suited to the cause. Earthquakes are well known to excite electromagnetic pulse, that cause tremors under the Earth's crust. These electromagnetic pulses are shielded effectively by Reinforced Concrete. Reinforced concrete has a low tensile strength, and hence Steel bars are used, which are embedded in the concrete sets. This provides Reinforced Concrete with immense ability to withstand natural calamities such as Earthquakes. This fact justifies the reason for the widespread presence of Reinforced Concrete among the buildings with Earthquake Damage grade 1, and its minimal presence in buildings with Earthquake Damage grade 5. The applications of this work can be further extended to predict damage caused by Earthquakes in areas for which a similar and relevant dataset can be obtained.

References

[1] C.P Shabariram and K.E Kannammal "Earthquake prediction using map reduce framework" 2017 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6, 2017.

[2] You-Po Su and Qing-Jie Zhu "Application of ANN to Prediction of Earthquake Influence" 2009 Second International Conference on Information and Computing Science, vol. 2 pp. 234-237, 2009.

[3] Cao Li and Xiaoyu Liu, "An improved PSO-BP neural network and its application to earthquake prediction," in 2016 Chinese Control and Decision Conference 2016, pp. 3434–3438.

[4] Xueli Wei, Xiaofeng Cui, Chun Jiang, Xinbo Zhou "The Earthquake Probability prediction based on weighted factor coefficients of principal components "2009 Fifth International Conference on Natural Computation, vol. 2, p. 608-612, 2009

[5] Dezhang Sun, Baitao Sun "Rapid prediction of earthquake damage to buildings based on fuzzy analysis"2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, vol. 3, p. 1332-1335, 2010

[6] Jianwen Zhang, Changyong Wang "Sheer Capacity computation of steel reinforced lightweight concrete beams" International Conference on Mechanic Automation and Control Engineering, 2010. p 1502-1506

[7] The Hackerearth platform official website "https://www.hackerearth.com"

[8] Python Seaborn Library pydata official website "https://seaborn.pydata.org/generated/seaborn.scatterplot.html"

[9] A Competitive Machine Learning series 'Machine Learning Challenge #6' problem 'Predicting the damage to a building' held on HackerEarth "https://www.hackerearth.com/problem/machinelearning/predict-the-energy-used-612632a9-3f496e7f/"

[10] United States Geological Survey glossary definitions "https://earthquake.usgs.gov/learn/glossary/?term=magnitude"

[11] Scikit-Learn documentation on Label Encoder "https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html"

[12] Long Wang, Xiaoqing Wang, Aixia Dou, Dongliang Wang "Study on construction seismic damage loss assessment using RS and GIS" International Symposium on Electromagnetic compatibility, 2014.

[13] Ramli Adnan. Abd Manan Samad, Zainazlan Md Zain, Fazlina Ahmat Ruslan "5 hours flood prediction modeling using improved NNARX structure: case study Kuala Lumpur", IEEE 4th International Conference on System Engineering and Technology, 2014.

[14] H Takata, H. Nakamura, T Hachino "On prediction of electric power damage by typhoons in each district in Kagoshima Prefecture via LRM and NN", SICE Annual Conference, 2004.

[15] Feature importance with a forest of trees "https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html