American Journal of Biological and Environmental Statistics
2020; 6(3): 58-63
http://www.sciencepublishinggroup.com/j/ajbes
doi: 10.11648/j.ajbes.20200603.14
ISSN: 2471-9765 (Print); ISSN: 2471-979X (Online)

Earthquake Damage Prediction Using Random Forest and Gradient Boosting Classifier

Sourav Pandurang Adi, Vivek Bettadapura Adishesha, Keshav Vaidyanathan Bharadwaj, Abhinav Narayan

Electronics and Communication Engineering, RNS Institute of Technology, Bengaluru, Karnataka, India

Email address:
souravadi1998@gmail.com (S. P. Adi), vivek.adishesha@gmail.com (V. B. Adishesha), keshavbharadwaj98@gmail.com (K. V. Bharadwaj), anarayan2016@gmail.com (A. Narayan)

To cite this article:
Sourav Pandurang Adi, Vivek Bettadapura Adishesha, Keshav Vaidyanathan Bharadwaj, Abhinav Narayan. Earthquake Damage Prediction Using Random Forest and Gradient Boosting Classifier. *American Journal of Biological and Environmental Statistics*. Vol. 6, No. 3, 2020, pp. 58-63. doi: 10.11648/j.ajbes.20200603.14

Received: September 26, 2020; **Accepted**: October 13, 2020; **Published**: October 21, 2020

Abstract: Earthquake is a major natural disaster that causes casualties in millions and leaving many more in trauma. Analyzing the consequences of such consequences gives one a better stand-in for potential catastrophe occurrences. It is important to establish a methodology that can assist in forecasting these earthquakes, as they can help prevent the severity of the damage. This paper discusses a machine learning model that can predict the damage grade severity caused by life-threatening earthquake that hit Nepal in the year 2015. The dataset is derived from the live competition hosted by Driven Data. The data was collected through the surveys conducted by the Kathmandu Living Labs and the Central Bureau of Statistics, which operates under the National Planning Commission Secretariat of Nepal. To accomplish the defined goal, we used the Random Forest Classifier and Gradient Boosting Classifier. The Random Forest Classifier algorithm demonstrated in this study was outperformed by the Gradient Boosting Classifier. With necessary parameter tuning using the Random Forest Classifier, the F1-Score achieved was 72.95%. The next technique was to perform winsorization on some attributes to handle outliers which improved the F1-score to 74.33% along with gradient boosting classifier. The last techniqueinvolved only hyper-parameter tuning with gradient boosting classifier achieved the best F1-Score of 74.42%.

Keywords: Random Forest Classifier, Gradient Boosting Classifier, Winsorizing, Earthquake

1. Introduction

Earthquakes almost always occur on faults and on the surfaces of the earth where one side is rising in relation to the other. Typically, earthquakes occur on faults, previously identified by geological mapping, which shows that motion across the fault has occurred in the past. Earthquakes that happen very near to the surface of the Earth bear an impact that is visible as fault lines on the land or ground.

Here are a few types of earthquakes:

A Volcanic earthquake is an earthquake that results when tectonic forces occur concurrently with volcanic velocity.

Tectonic Earthquake occurs when the rocks change their physical and chemical properties due to the geological forcescausing the break of the earth's crust

Collapse earthquakes are the smaller earthquakes that are a result of seismic waves and generally are observed in caverns and mines.

The outburst of either a nuclear or a chemical device or both simultaneously leads to an explosion earthquake.

Here, in this research, we are working on the tectonic earthquake which shook Nepal with a Richter Magnitude of 7.8Mw on April 25, 2015 [7]. This catastrophic life-threatening earthquake ended up killing over 8000 people and leaving 22000 injured. Century-old buildings (ancient ones) including Changu Narayan Temple and Dharahara Tower were demolished at UNESCO World Heritage Sites in Valley ofKathmandu. Hundreds of houses have been lost in many Nepal districts. It was the worst earthquake that hit Nepal in 80 years. An avalanche was triggered on Mount Everest slaughtering approximately 20 people. Many landslides were observed in steep valleys covering Ghodatabela, killing about 250 people. Reports at the time of the quake described the number of trekkers and climbers at

base camp as up to 1000.

Originally, the United States Geological Survey [7] (USGS) presented an estimate of economic losses of up to 9% to 50% of GDP, with the highest estimate of 35%. India and China provided economic assistance to Nepal, which totaled more than $1 trillion. Over 100 members of the search and rescue team (Lifesaving Troops), medical experts, and three Chinook helicopters were sent for use by the government of Nepal. Asian Development Bank (ADB) assisted Nepal with a $3 million grant as support measures and up to $200 million for initial recovery. The UK gave £73 million to which the government donated £23 million, and the public donated £50 million. The United Kingdom also assisted by supplying 30 tons of humanitarian assistance and 8 tons of supplies. In this research, we have used dataset given by driven data [19], performed EDA using Tableau [12], and finally developed a machine learning model that is capable of predicting the damage grade severity to the buildings caused by the earthquake [1, 2, 4, 9, 16, 17]. The models can also be used for forecasting the damage level to the buildings to a certain extent [6, 8, 20]. The performance of the models was measured using F1-Score [11].

2. Literature Survey/Related Works

Asim et al. [1] (2016) used different machine learning algorithms for earthquake magnitude prediction for theHindukush region. All algorithms behave differently than others, but the Linear Programming Boost Ensemble Classifier displays better sensitivity performance, while the Pattern Recognition Neural Network appears to deliver the least false alarms relative to the other classifiers. The researchers finally prove that the random phenomenon of earthquakes can be modeled using different machine learning techniques.

To achieve the necessary results, ensemble learning algorithms are used. The Random Forest Classifier is taken up first and then the Gradient Boosting Classifier. The results depict the Gradient Boosting Classifier algorithm is outperforming the Random Forest Classifier. Hosokawa et al. (2009) [2] proposed an earthquake damage prediction system that focused on a combination of earthquake data, accurate ground conditions, and multi-temporal SAR prediction.

Dezhang Sun et al.(2009), [3] centered on fuzzy mathematics, a membership approach has been developed to forecast earthquake damage to buildings to estimate earthquake risk reduction. They used the seismic risk index as an earthquake damage measure, the cumulative seismic damage index as a return index, the impact factors as a shift index.

Rapid assessment of damage severity to the buildings is an essential post-event recovery. To achieve this Sujith Mangalathu (2019) [4] et al. evaluated the possibility of using various machine learning techniques such as K-Nearest Neighbors, Random Forests, Decision Trees, etc. Data from the 2014 Napa earthquake was used for research in which the damage was graded based on the ATC-20 tag assigned to it. The machine learning model used spectral acceleration at 0.3s, fault size, building unique characteristics such as age, floor area, etc.

Hidenori Kawabe [5] et al. (2008) predicted the damage potential of steel and reinforced concrete high-rise buildings and constructed damage prediction maps for the Osaka basin using the long-period ground motions. For earthquake response analysis, one mass model (analytical model for the equivalent mass of one degree of freedom) is adopted. Using the maps, authors point out that the dynamic response of high- rise buildings exceed the present seismic design criteria.

David Vere-Jones (1995) [6] reviewed the issues that arise during earthquake prediction and the risk of forecasting earthquakes. Katsuichiro Goda (2015) [7] et al. summarized key findings of ground shaking damage in Nepal. With the available seismological data, building damage was linked by reviewing the seismotectonic setting of Nepal, earthquake rupture process, aftershock data which was provided by the U.S. Geological Survey (USGS).

Daniel Weijie Loi et al. [8] (2014) reported about the challenges in using earthquake data interpretation regression models. The report briefs out the critical mistakes between the expected data and the field data. The paper concludes byachieving a more practical prediction model for earthquake data.

K Chaurasia et al.(2019) [9] predicted the level of damage caused by the earthquake that hit Nepal in the year 2015. The researchers have used Neural Networks and Random Forest classifier techniques to achieve the goal.

Khaleed Talab et al.(2018) [10] developed a data mining methodology for the development of Landslide Susceptibility Maps (LSM's) for areas that are highly susceptible to be affected by landslides. The authors use Random Forest algorithm to produce more reliable maps.

3. Dataset

There are 39 columns in the dataset [19] consisting of binary and categorical datatypes. The dataset includes geo_level_id's, floor count before the earthquake, age of the building, normalized area and height as integer data, land surface condition, foundation type, roof type, etc. as categorical data and secondary uses like school, institution, industry, etc. as binary data.

The goal was to predict damage level severity labeled from 1-3. The variable is of ordinal type. The valuation was done based on the F1-Score [11] which balances precision and re-call/sensitivity by considering the harmonic mean between the two. The problem can be classified as a classification or an intermediate problem statement between classification and regression.

4. Approach

4.1. Data Preparation

To start with we performed exploratory data analysis to understand the relations between different attributes that

were provided. To perform the EDA, we used Tableau [12]. Tableau is a data visualization tool used for data analysis. The first move involved testing the missing and duplicate values in the data. We used the Pandas library of Python to achieve this. The observation was that there were no missing values and duplicate values.

4.2. Exploratory Data Analysis

A preliminary check is performed on the distribution of target variable damage grade. The observation we could make was that about 56.89% of the damage grade on the building had a severity level of 2, 33.47% had a severity level of 3 and 9.64% had a level of 1. Figure 1.

Next, the relationship between damage grade and number of floors is checked. In comparison to single-floor buildings, buildings with 2 floors had significant damage followed by 3-floor buildings. It is also notable that 2-floor buildings have a damage grade of 2 followed by 3. The same was observed for 3-floor buildings. Figure 2.

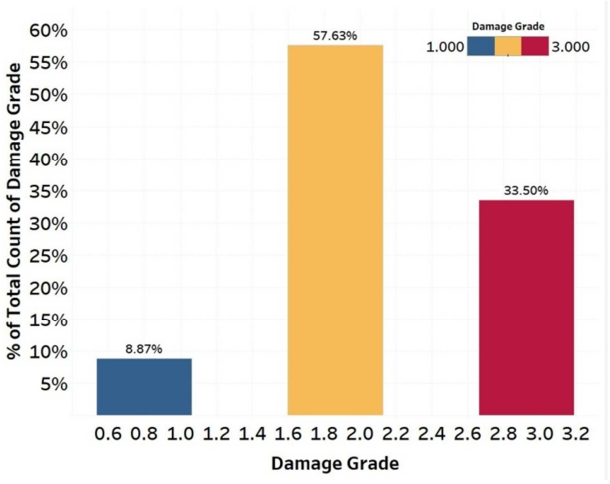

Figure 1. *Percentage Distribution of Damage Grade.*

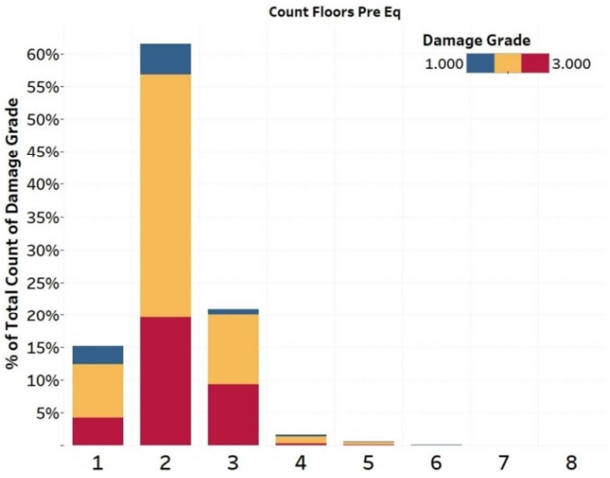

Figure 2. *Percentage Damage Grade vs Floor Count of the Building before mishap.*

The next process is to find a relation between age and dam- age grade. Tableau informs that, buildings aged less than 50 years has a dominating damage grade. Buildings aged form 0-20 saw a significant rise in damage grade and those of 15 and above saw a steady decline. Most of the buildings aged 10 years had a damage grade of 2 followed by 3, this trend was also observed for buildings of different ages. Figures 3 and 4.

The plot between damage grade vs land surface condition showed that there was a severe damage grade observed for those buildings whose land condition was of type ‘t’ followed by ‘n’ and ‘o’ respectively. Figure 5.

Buildings with foundation type of ‘r’ had a greater damage than other categories (w, u, h, i). Figure 6.

Those buildings which had ground floor type of ‘f’ suffered a greater damage followed by ’x’, ’v’,’z’ and ’m’ respectively. Figure 7.

Buildings with other floor type of ‘q’ suffered the highest damage than compared to ’x’, ’j’ and ’s’ types. Figure 8.

Plan configuration type of ‘d’ had a higher damage than all other types. Figure 9.

Position ‘s’ and ‘t’ had greater damage than compared to ’j’ and ’o’ Figure 10.

‘n’ and ‘q’ roof type buildings were damaged badly than ’x’ Figure 11.

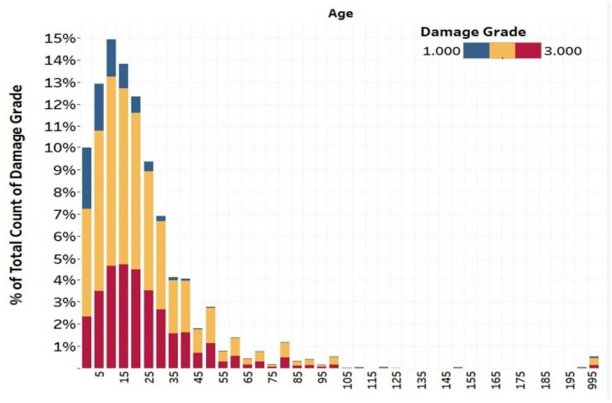

Figure 3. *Percentage Damage Grade vs Age of the Building (complete).*

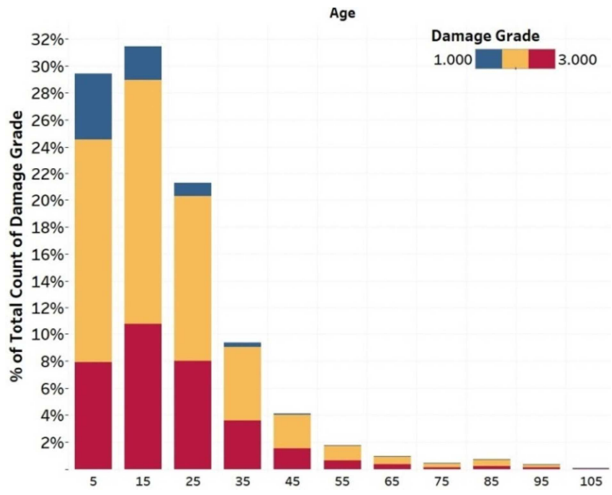

Figure 4. *Percentage Damage Grade vs Age of the Building (0-105 years only).*

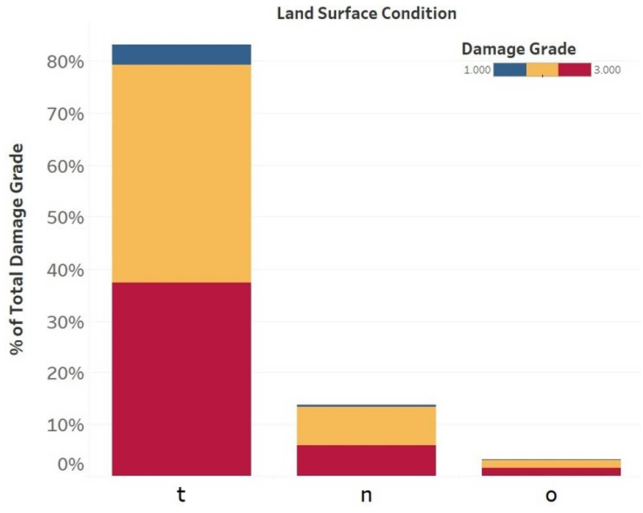

Figure 5. *Percentage Damage Grade vs Land Surface Condition.*

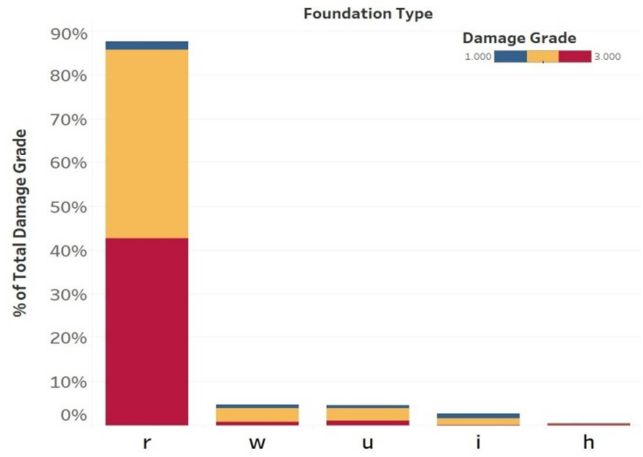

Figure 6. *Percentage Damage Grade vs Type of the Foundation.*

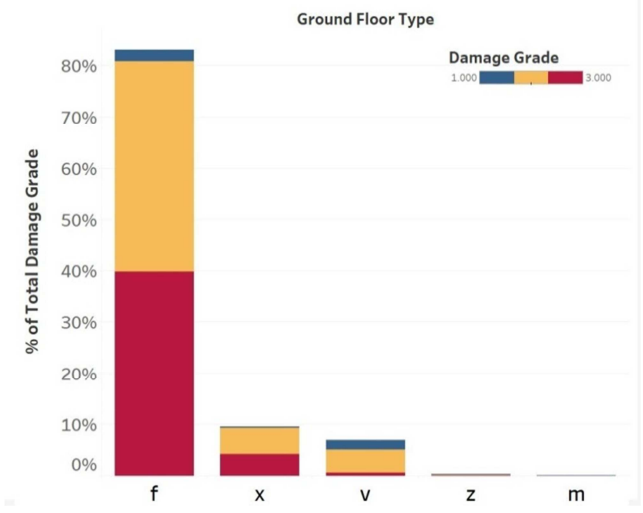

Figure 7. *Percentage Damage Grade vs Ground Floor Type.*

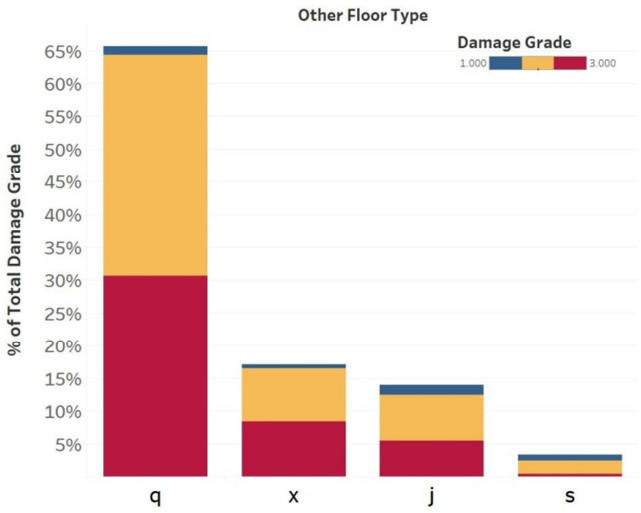

Figure 8. *Percentage Damage Grade vs Other Floor Type.*

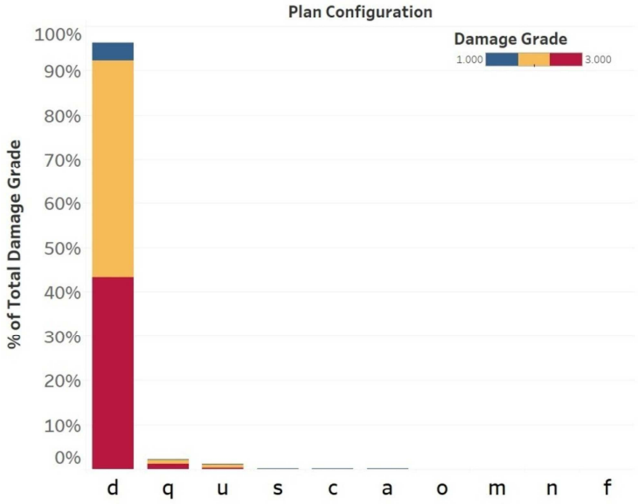

Figure 9. *Percentage Damage Grade vs Plan Configuration.*

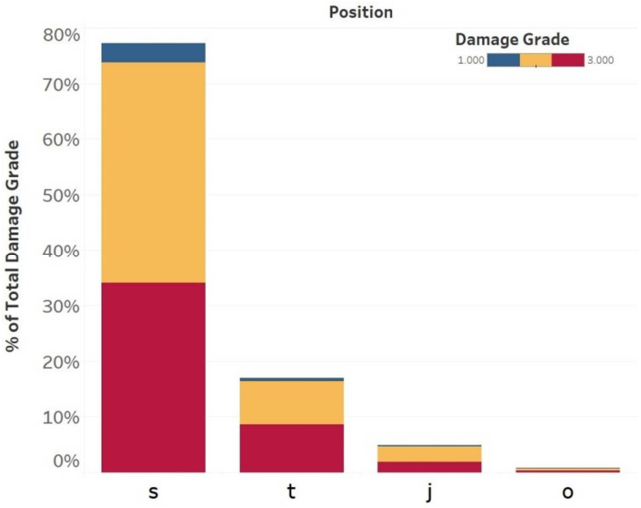

Figure 10. *Percentage Damage Grade vs Position.*

5. Training and Testing the Model

Here we train the Machine Learning model with all listed features to predict the target variable damage grade. Before the model is fitted on the data, transformations have to be applied since data have categorical variables that cannot be directly fed to the model. One hot encoding was used to encode the categorical variables. Data is then split into dependent and independent variables to make a train test split of 80% of data for model preparation (training), and 20 percent for model checking (testing).

Three methods were tried to achieve a better result, first one was to fit the Random Forest classifier [13] to the train data by varying different parameters. Once trained, prediction on the 20% test data using the model is done. The Second method was to apply Winsorizing [15] on different attributes and Gradient Boosting Classifier [14] was fit to the data. The third method was to remove Winsorizing and perform hyperparameter tuning to the classifier. All our methods were later validated on new test data which was supplied by the driven data platform.

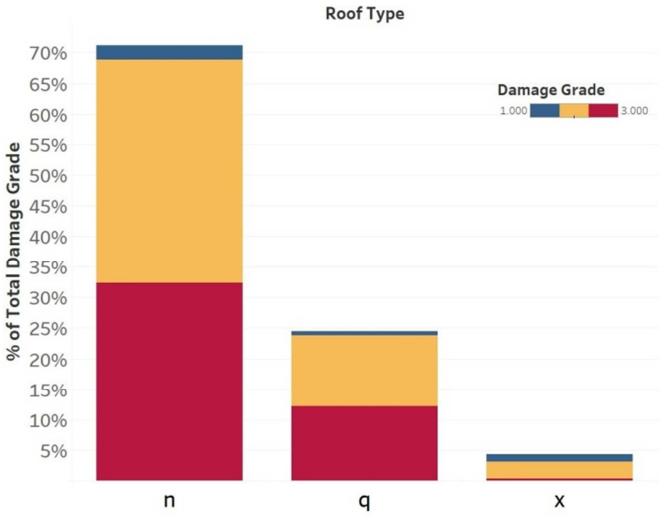

Figure 11. *Percentage Damage Grade vs Roof Type.*

The overall flowchart is shown in Figure 12.

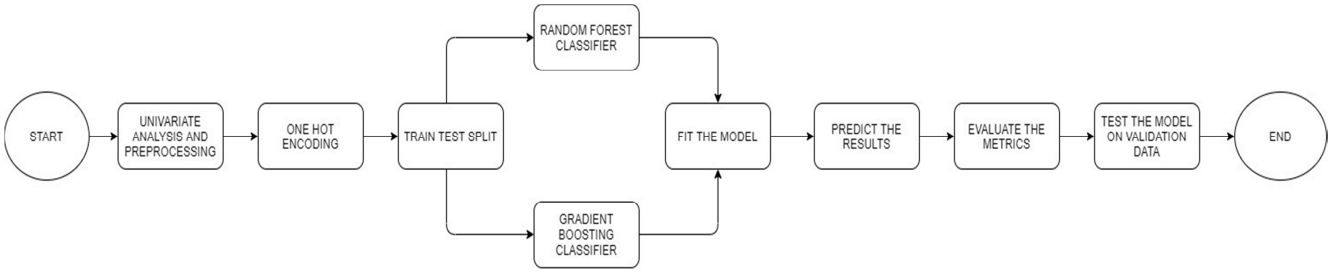

Figure 12. *Flowchart.*

6. Results and Evaluation

The competition results were evaluated based on F1-Score. F1-Score is the formula combining accuracy with recall (harmonic mean between accuracy and recall/sensitivity). Mathematically, F1-Score is given byEquation (1). Generally, F1-Score is preferred whenever there is a need for a balance between precision and recall, also when the data is unevenly distributed. The fact is that F1-Score is used for performance evaluation in the case of binary classifier but since we are working on the problem having more than twolabels, the result will be evaluated based on micro averaged F1-Score.

Referencing to the table 1, we can clearly say that the gradient boosting classifier algorithm outperforms Random Forest Classifier. For the Random Forest Classifier, we kept the default values for most of the parameters and changed the number of decision trees (estimators) to 500, maximum depth to 5 and kept the criterion as Gini. Advantage of theRandom Forest Classifier is that the more the number of trees more accurate is the output. The problem is that with increase in trees more complex the solution becomes. Gini entropy avoids the complex logarithmic calculations. With the parameters mentioned in the table we achieved the F1-Score of 0.7295 (72.95%).

The Gradient Boosting Classifier was applied with some more parameter tuning process and this time winsorizing [15] was removed. The process involved setting the estimators to 800, learning rate to 0.2, minimum samples split to 1600, minimum samples leaf to 250, maximum depth to 9 and subsample to 0.9. With this tuning we could achieve abest result of 0.7442 (74.42%).

$$F_{micro} = \frac{2 \times Precision_{micro} \times Sensitivity_{micro}}{Precision_{micro} + Sensitivty_{micro}} \quad (1)$$

where,

$$Precision_{micro} = \frac{\sum_{b=1}^{3} TP_b}{\sum_{b=1}^{3} (TP_b + FP_b)} \quad (2)$$

and

$$Sensitivty_{micro} = \frac{\sum_{b=1}^{3} TP_b}{\sum_{b=1}^{3} (TP_b + FN_b)} \quad (3)$$

Here, the abbreviations TP represents True Positive, FP rep- resents False Positive, FN represents False Negative and 'b' is an indication of number of classes which is 1, 2 and 3 respectively representing the damage grade severity.

Table 1. *Results.*

Classifier	Parameters	Micro Averaged F1-Score
Random Forest Classifier	Estimators = 500 Depth = 5 Criterion = Gini	0.7295 (72.95%)
Gradient Boosting Classifier with Winsorizing	Estimators = 300 Depth = 10 Warm_Start = True Learning Rate = 0.1	0.7433 (74.32%)
Gradient Boosting Classifier without winsorizing	Estimators = 800 Depth = 9 Learning Rate = 0.2 min_samples_split = 1600 min_samples_leaf = 250 subsample = 0.9	0.7442 (74.42%)

7. Conclusion

To conclude, a simple machine learning model that was able to properly classify the damage severity to the buildings caused by the life-threatening Gorkha earthquake is developed. In this research, a machine learning model using Random Forest Classifier algorithm and Gradient Boosting Classifier algorithm with and without Winsorizing was built which was able to achieve the F1-score of 0.7295 (72.95%), 0.7433 (74.33% (with Winsorizing)), and 0.7442 (74.42% (without Winsorizing) respectively for the described problem. The main drawback of the above-mentioned methods is the time constraint involved. The further development is to build a more optimal model so as to overcome the time constraint and also with improved accuracy.

References

[1] K. M. Asim, F. Martı´ nez A´ lvarez, A. Basit, and T. Iqbal "Earthquake magnitude prediction in Hindukush region using machine learning techniques". Nat Hazards, 2016.

[2] M Hosokawa, B. P Jeong, and O Takizawa. Earthquakeintensity 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] Sujith Mangalathu, M. EERI, Chukwuebuka C. Nweke, Han Sun, Henry V. Burton, and Zhengxiang Yi. Classifying earthquake damage to buildings using machine learning. Earthquake Spectra, 2020.

[5] 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.

[6] David Vere-Jones "Forecasting earthquakes and earthquake risk". International Journal of Forecasting, 11 (4): 503–538, 1995.

[7] T K Katsuichiro Goda "The 2015 Gorkha Nepal earthquake: insights from earthquake damage survey". Frontiers in Built Environment, 2015.

[8] D W Loi, M E Raghunandan, M Shanmugavel, andV Swamy. Data analytic engineering and its application in earthquake engineering: An overview, 2014.

[9] K. Chaurasia, S. Kanse, A. Yewale, V. K. Singh, B. Sharma, and B. R. Dattu. Predicting Damage to Buildings Caused by Earthquakes Using Machine Learning Techniques. In 2019 IEEE 9th International Conference on Advanced Computing (IACC), pages 81–86, 2019.

[10] Khaled Taalab, Tao Cheng, and Yang Zheng "Mapping landslide susceptibility and types using Random Forest". Big Earth Data, 2 (2), 2018.

[11] M S Szpakowicz. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. Springer, Berlin, Heidelberg, 2006.

[12] Inseok Ko and H C. Interactive Visualization of Healthcare Data Using Tableau, 2017.

[13] Qiong & Ren, Hui & Cheng, and Hai Han "Research on machine learning framework based on random forest algorithm". AIP Conference Proceedings, 2017.

[14] Alexey & Natekin and Alois Knoll. Gradient Boosting Machines, A Tutorial. Frontiers in neurorobotics, 2013.

[15] Alan Reifman and Kristina Garrett "Winsorize". En-cyclopedia of research design, pages 1636–1638, 01 2010.

[16] https://medium.com/swlh/predicting-damage-to-building-due-to-earthquake-using-data-science-e85a62adc0c0.

[17] https://towardsdatascience.com/earthquake-prediction-faffd7160f98.

[18] https://arxiv.org/ftp/arxiv/papers/1702/1702.05774.pdf.

[19] https://www.drivendata.org/competitions/57/nepal-earthquake/.

[20] Dr. P. Vishnu Raja, Dr K. Sangeetha, S. Sibikrishna, C. Shwetha, M. Vijaykumar (2020). Earthquake Prediction Using Machine.

[21] Learning Using Support Vector Machine Algorithm. *International Journal of Advanced Science and Technology, 2020.*