Al-Kitab Journal for Pure Sciences (2025); 9(1):1-13.
https://doi.org/10.32441/kjps.09.01.p1

Al-Kitab Journal for Pure Sciences
ISSN: 2617-1260 (print), 2617-8141(online)
https://isnra.net/index.php/kjps

Soft Voting Classifier of Machine Learning Algorithms to Predict Earthquake

Oqbah Salim Atiyah

Department of Computer Science, College of Computer Science and Mathematics, University of Tikrit, Iraq.

*Corresponding Author: oqbah_salim@tu.edu.iq

Citation: Atiyah OS. Soft Voting Classifier of Machine Learning Algorithms to Predict Earthquake. Al-Kitab J. Pure Sci. [Internet]. 2024 Dec. 28;9(01):1-13.DOI: https://doi.org/10.32441/kjps.09.01.p1

Keywords: Earthquakes, Machine Learning, Soft Voting Classifier.

Article History

Received	16 Aug.	2024
Accepted	22 Oct.	2024
Available online	28 Dec.	2024

Abstract:

Earthquakes are among the most dangerous natural disasters that can cause major losses to buildings and threaten human lives. The research community is very interested in the topic of earthquakes because they occur suddenly and predicting them is very important for human safety. Creating accurate earthquake prediction techniques by applying machine learning (ML) approaches will help save people's lives and prevent damage. To identify important features and analyze the correlation between these features before submitting them to classification models, we proposed a new feature selection approach in this paper which combines two filtering ways: Normalization which is based on the Chi-square approach and analysis of variance, and the correlation approach based on the logistic regression technique (CLR-AVCH). Accordingly, three algorithms are applied. Then a facilitated voting classifier is created that combines the two best models with the highest prediction accuracy (histogram-based gradient boosting, adaptive boosting) to create a single technique that includes the strengths of the techniques that were combined to help find important patterns in the acquired data to obtain a model capable of early prediction of earthquakes. The proposed work achieved higher accuracy, F1_score, recall, and precision (0.94, 0.92, 0.94, 0.92), respectively.

Keywords: Earthquakes, Machine Learning, Soft Voting Classifier.

مصنف التصويت الناعم من خوارزميات التعلم الآلي للتنبؤ بالزلازل

عقبه سالم عطية
قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة تكريت، العراق
oqbah_salim@tu.edu.iq

الخلاصة:

تعتبر الزلازل من أخطر الكوارث الطبيعية التي يمكن أن تسبب خسائر كبيرة في المباني وتهدد حياة الإنسان. يهتم المجتمع البحثي كثيراً بموضوع الزلازل لأنها تحدث بشكل مفاجئ والتنبؤ بها مهم جداً لسلامة الإنسان. إن إنشاء تقنيات دقيقة للتنبؤ بالزلازل من خلال تطبيق أساليب التعلم الآلي (*ML*) سيساعد في إنقاذ حياة الناس ومنع الضرر. لتحديد الميزات المهمة وتحليل الارتباط بين هذه الميزات قبل تقديمها إلى نماذج التصنيف، لقد اقترحنا نهجًا جديدًا لاختيار الميزات في هذا البحث والذي يجمع بين طريقتين للتصفية: التطبيق الذي يعتمد على نهج مربع كاي (*Chi-square*) وتحليل التباين، ونهج الارتباط المعتمد على تقنية الانحدار اللوجستي (*CLR-AVCH*). وبناءً على ذلك، تم تطبيق ثلاث خوارزميات. وتم إنشاء مصنف التصويت الميسر الذي يجمع بين أفضل نموذجين يتمتعان بأعلى دقة للتنبؤ (*AdaBoost* , *HGBoost*) ، لتكوين تقنية واحدة تضم نقاط قوة التقنيات التي تم دمجها للمساعدة في العثور على الأنماط المهمة في البيانات المكتسبة للحصول على نموذج قادر على التنبؤ المبكر بالزلازل. حقق العمل المقترح اعلى النتائج بدقة ٠,٩٤ و *F1_score* ٠,٩ و*recall* ٠,٩٤ و*precision* ٠,٩٢.

الكلمات المفتاحية: الزلازل ، التعلم الآلي ، مصنف التصويت الناعم.

1. Introduction:

An earthquake is a mild to severe tremor caused by the sudden movement of underground rocks. There are four types of earthquakes: tectonic, collapse, explosive, and volcanic. Earthquakes occur at tectonic fault boundaries. A tectonic earthquake occurs due to the breaking of the Earth's crust due to geological forces affecting adjacent plates, causing a change in the chemical and physical structure. Often tectonic blocks move slowly, so they are trapped at their edges due to friction. Thus, when the pressure on the edge of friction increases, an earthquake occurs, and there are energy waves that are transmitted to the Earth's crust and cause the vibration that forms the earthquake [1], many losses and injuries due to earthquakes. Every day, there are natural disasters in all countries of the world. The countries most vulnerable to earthquakes are Taiwan, Southern California, Iran, Indonesia, Turkey, and Japan.

People feel an earthquake if its magnitude is greater than 2.5, and people do not feel it if its magnitude is less than 2.5. Destructive earthquakes have a magnitude greater than 4.5 [2]. Sometimes earthquakes cause serious deaths and great material damage, so researchers make a great effort to predict earthquakes to stop such negative events. To notify people of the danger promptly. Humans cannot prevent earthquakes, but preventive measures can be taken to reduce negative events, using machine learning approaches, the strength and danger of earthquakes can be predicted [3]. Many technologies can be used, such as sensors, magnetic and electrical waves, and other devices that can estimate the size of an earthquake based on seismic indicators by analyzing available data on previous earthquakes [1]. Until now, there is no optimal model that gives results with 100% accuracy, but there are still attempts to increase the accuracy of the results provided [4]. Many factors affect the machine learning prediction process, such as the amount of features, the number of constraints in the data set, and the nature of the classification or regression problem. Therefore, we will use several algorithms in ML and compare their results to discover the most suitable one for the specific situation [5]. The main goal of earthquake forecasting is to determine three important things: the future earthquake, where it will occur, when, and its size, to reduce losses resulting from earthquakes. Predicting earthquakes can significantly reduce seismic damage, and this is the primary goal. Accordingly, there is a great interest in conducting research studies on earthquake prediction [6].

2. Related Work:

- Koehler, J., Li, W., Faber, etc. (2023). This study was conducted to predict earthquakes by using deep learning (DL) to discover whether a time series lasting more than two years generates an earthquake larger than five magnitudes or not. The trained technique was evaluated, and the accuracy was approximately 72.3%. Therefore, it should be developed by using more available data [7].
- Sajan, K. C., Bhusal, A., (2023). The paper applies four techniques of decision tree, random forest, logistic regression, and eXtreme gradient boosting (XGBoost) in machine learning to predict the degree of damage and rehabilitation. By comparison with other techniques, XGBoost predicts the collapse of institutions and buildings with better accuracy. However, prediction techniques must be developed to achieve better accuracy [8].
- Researchers Yang, F., and Kefalas, M., (2022), presented a regression model in machine learning. They developed an automated regression model based on laboratory seismic data, which helps predict earthquakes. This automated model included modeling, optimization, feature extraction, and selection techniques. The Bayesian approach is used to develop hyper-parameters. The results achieved for the model on test and training data mean square error (MSE) 1.48, 1.51, and mean absolute error (MAE) 1.52, 1.59. However, this model needs to be developed to achieve the best results for predicting laboratory earthquakes [9].
- Another study by (Berhich, A., Belouadha, F. Z., & Kabbaj, M. I), applied recurrent neural network technology to predict earthquakes based on location. K-Means is used to aggregate data based on geographical parameters to provide a prediction depending on location. The data set was divided into two groups. The first group includes seismic events whose

magnitude is between 2 and 5, while the second group includes events whose magnitude is greater than 5. The model needs to be developed to provide accuracy in performance [10].

● An, Z., et al. (2023), presented a model that combines the multilayer perception (MLP) and deep interest network (DIN) models and simulated data tests to predict earthquakes. In this study, the data were processed, and the DIN model was used to predict earthquakes. The results achieved by the proposed approach are 0.69, and this indicates an improvement rate of about 11% over the original DIN technology. With this improvement, there is still a need to develop the models used to enhance the monitoring accuracy and model efficiency [11].

3. Methodology:

This stage introduces some classification techniques in ML in this study:

3.1 Hist Gradient Boosting (HGBOOST): Histogram-based gradient boosting, or Hist Gradient Boosting (HGBoost) [12], is a boosting ensemble that utilizes the histograms of features for accurate selection and fast for best splits. It is characterized by fast processing. It reduces the number of features by pooling with graphs to increase the speed of the algorithm.

3.2 ADABOOST: Adaboost (Ada) or adaptive boosting, is a popular technique that fits into the basic series of algorithm classifiers to update the weight of samples by giving the misclassified samples more weight and then adapting the learner with the newly updated weights. By combining the results utilizing the majority voting technique in the classifier base, the final prediction is reached, and by reaching a higher weight, the highest performance of the basic classifier is determined [13].

3.3 K-nearest Neighbour (KNN): It is one of the common techniques used in classifying data, as it works to create another case that is added to the sample that currently exists in a specific space, and the nearest neighbor is found by calculating features similar to the sample. Where k is a constant value, through which features similar to the new case are calculated, this leads to selecting new samples and classifying them into similar categories [12] [14].

In this paper, the researcher proposes using a classification voting technique, which combines multiple methods to provide the best prediction accuracy for earthquake detection.

4. Proposed method:

The proposed work consists of three steps: pre-processing, feature selection step, and prediction models. A dataset of earthquakes was downloaded from Kaggle. Which includes two possibilities: tsunami (1) and other (0). It contains categorical data as well as numerical data and is stored in CSV format. The 782 samples consist of 2001–2023 [15]. Figure 1 depicts a system workflow.

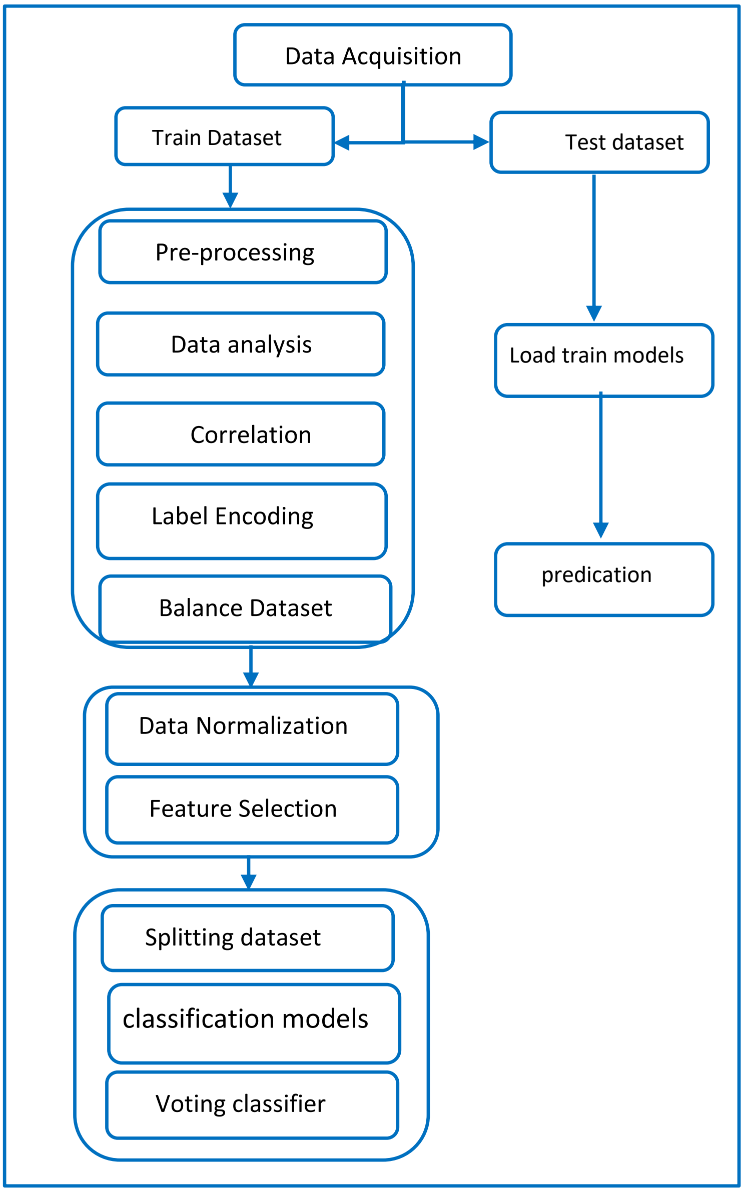

Figure 1: System workflow

4.1 Pre-processing Phase: We apply a series of preliminary operations on the data to improve its efficiency so that the classification techniques work optimally. The basic operations in this step are cleaning, and processing missing values. If the missing value is numeric, the column average will be calculated and replaced with the missing value, and if the missing values are nominal, it will be replaced with neighborhood values [16]. Now, to produce the dataset free of missing values, ready for later use, and encoding the data in the data set, as Table 1. And identify useful features for developing the model [17].

Table 1: Summary of the analyzing dataset.

Seq.	features	Missing data	After handling	types
0	title	0	0	object
1	magnitude	0	0	float64

2	date time	0	0	object
3	cdi	0	0	int64
4	mmi	0	0	int64
5	alert	367	0	object
6	tsunami	0	0	int64
7	sig	0	0	int64
8	net	0	0	object
9	nst	0	0	int64
10	dmin	0	0	float64
11	gap	0	0	float64
12	magType	0	0	object
13	depth	0	0	float64
14	latitude	0	0	float64
15	longitude	0	0	float64
16	location	0	0	object
17	continent	0	0	object
18	country	0	0	object

4.2 Correlation: Correlation indicates the relationship between changes with other pairs. So, correlations are displayed to illustrate low and high correlations between variables [18]. The correlation technique is widely used in datasets to identify relationships that help to understand the importance of attributes with respect to the target group to be predicted, as in Figure 2. Thermal correlations measure the degree of correlation between features, where the values are [-1,1]. Value (1): represents the presence of correlation between the two features. And (0): represents the absence of correlation between the two features. The value (−1): represents the inverse correlation between the two features, the equation below is used to calculate correlation [19]. n is a sample size, x_i , y_i are data points in dataset, and the $\bar{x}$ is mean of x-values, $\bar{y}$ is mean of y-values.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n} \sqrt{(x_i-\bar{x})^2 \sum_{i=1}^{n} (y_i-\bar{y})^2}} \quad (1)$$

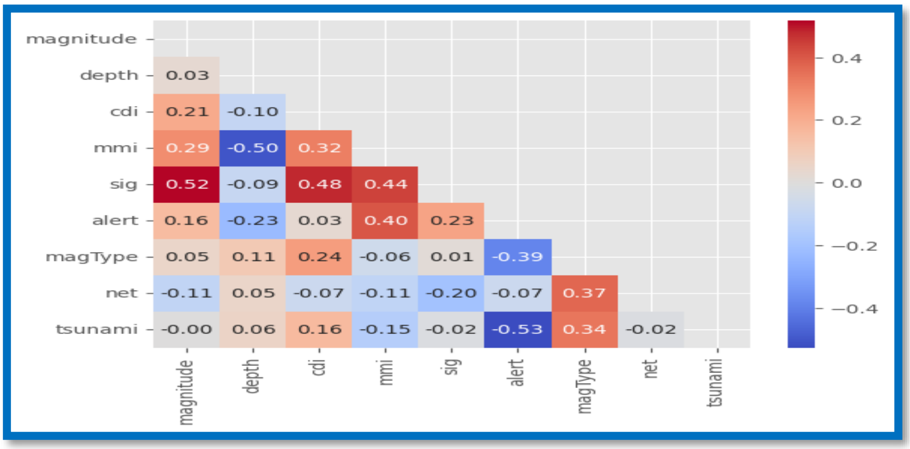

Figure 2: The correlations.

4.3 Data Encoding: Variables are affected by the scale of qualitative factors. Therefore, these variables must be converted into numerical values through encryption techniques, as some machine learning algorithms only deal with numerical variables [20]. Therefore, categorical data must be converted to numerical values (**Table 2**). In this paper, the label coding approach was used, the labels of categorical data were converted into numerical values to be used in modeling and analysis. This approach is effective and simple. The working principle of the label encoding approach into each class value in the class variable a unique integer is assigned. The three categorical variables in the data set were coded to improve the functioning of the model.

Table 2: depicts the data encoding.

a. data before labeling

	magnitude	date_ time	cdi	mini	alert	tsunami	sig	net	nst	dmi	gap	Mag Typ	depth	latitude	longitude
0	7.0	2-11-2022 02:03	8	7	green	1	768	us	117	0.5	17.0	mww	14	-10	159.6
1	6.9	8-11-2022 13:37	4	4	green	0	735	us	99	2.2	34.0	mww	25	-5.0	100.8
2	7.0	2-11-2022 07:09	3	3	green	1	755	us	147	3.1	18.0	mww	579	-20.1	-178.3
3	7.3	1-11-2022 10:48	5	5	green	1	833	us	149	1.8	21.0	mww	37	-19.3	-172.1
4	6.6	9-11-2022 10:14	0	2	green	1	670	us	131	5.0	27.0	mww	624	-25.6	178.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
777	7.7	3-01-2001 17:33	0	8	red	0	912	us	427	0	0.0	mwc	60	13	-88.7
778	6.9	0-01-2001 16:02	5	7	red	0	745	ak	0	0	0.0	mw	36.4	56.8	-153.3
779	7.1	9-01-2001 16:49	0	7	red	0	776	us	372	0	0.0	mwb	103	-15.0	167.2
780	6.8	1-01-2001 08:54	0	5	red	0	711	us	64	0	0.0	mwc	33	6.6	126.9
781	7.5	1-01-2001 06:57	0	7	red	0	865	us	324	0.	0.0	mwc	33	6.9	126.6

782 rows x 15 columns

data after labeling

	magnitude	date_ time	cdi	mini	alert	tsunami	sig	net	nst	dmi	gap	Mag Type	depth	latitude	longitude
0	7.0	22-11-2022 02:03	8	7	0	1	768	9	117	0.5	17.0	8	14	-10	159.6
1	6.9	18-11-2022 13:37	4	4	0	0	735	9	99	2.2	34.0	8	25	-5.0	100.8
2	7.0	12-11-2022 07:09	3	3	0	1	755	9	147	3.1	18.0	8	579	-20.1	-178.3
3	7.3	11-11-2022 10:48	5	5	0	1	833	9	149	1.8	21.0	8	37	-19.3	-172.1
4	6.6	09-11-2022 10:14	0	2	0	1	670	9	131	5.0	27.0	8	624	-25.6	178.3
…	…	…	…	…		…	…		…	…	…		…	…	…
777	7.7	13-01-2001 17:33	0	8	2	0	912	9	427	0	0.0	7	60	13	-88.7
778	6.9	10-01-2001 16:02	5	7	2	0	745	0	0	0	0.0	5	36.4	56.8	-153.3
779	7.1	09-01-2001 16:49	0	7	2	0	776	9	372	0	0.0	6	103	-15.0	167.2
780	6.8	01-01-2001 08:54	0	5	2	0	711	9	64	0	0.0	7	33	6.6	126.9
781	7.5	01-01-2001 06:57	0	7	2	0	865	9	324	0.	0.0	7	33	6.9	126.6

782 rows x 15 columns

4.4 Data Normalization: Normalization is the process of preparing data for machine learning. Normalization aims to reduce the number of features to a similar extent. Which stabilizes the training, improves its function, leads to more data, and increases its safety **[21]**. The MinMaxScaler function was used, it a function converts data to a specified range and rescales the data to set the minimum value to 0, and the maximum value to 1, which scales each feature individually and has a maximum and minimum value, with values of 1 and 0.

4.5 Feature Selection Phase: In this step, appropriate features are selected to obtain the best results. Less important data gives fewer opportunities to make decisions due to noise. Less frequent data gives greater accuracy, and this speeds up the work of the algorithms [22]. Which reduces training time [23]. The proposed method combines the filtering approach, and their association through logistic regression with the normalization process, works on analysis of variance (ANOVA), and uses the chi-square technique (CL-ANCH), as in Figure 3 which shows the proposed method for our study. Correlation analysis uses the logistic regression function (logis. coef) to determine the target value and measure the correlation between features. Where these values are treated as a set of correlation values. Highly related features are grouped into a single group that includes similar elements. The common features that have the best values are kept in a group and the rest of the features are neglected. The correlation of numerical variables is measured using the logistic regression model, the normalization technique, and ANOVA, which compare the differences through the average values of the variables. The P value represents the result of ANOVA. These values represent the difference between the variance within the group, which produces values that show that the null hypothesis is supported or rejected. The null hypothesis is rejected if the difference between variables is large. The Chi-squared technique performs a statistical test to examine the variance between categorical features that were randomly selected. The candidate variable for the feature is neglected and is not related to the problem. This approach shows that all categorical variables are important values, as the P value for the Chi-squared technique is <0.05. As in Table 3. This approach searches for variables that have correlation values less than 1 by analyzing the thermal correlations of the variables identified in Figure 2. It combines the existing set of correlated variables into a single group. AVONA and Chi-squared are applied to find variances through the mean values of the groups and find the P value for variables whose values are less than 0.05 as in (Table 3). While the remaining variables in each group are neglected. Figure 3 shows the proposed method for our study. Unimportant variables were removed, and variables were kept that enabled the model to be trained correctly to give better results.

Table 3: display the feature selection phase.

Features	With feature selection	With proposed method
magnitude	1.44	0.022
depth	3.02	0.002
mmi	-0.25	-0.003

sig	3.08	-0.01
alert	-0.078	-0.03
magType	-3.40	-0.04
net	-5.8	-0.07
tsunami	0.457333	-0.074

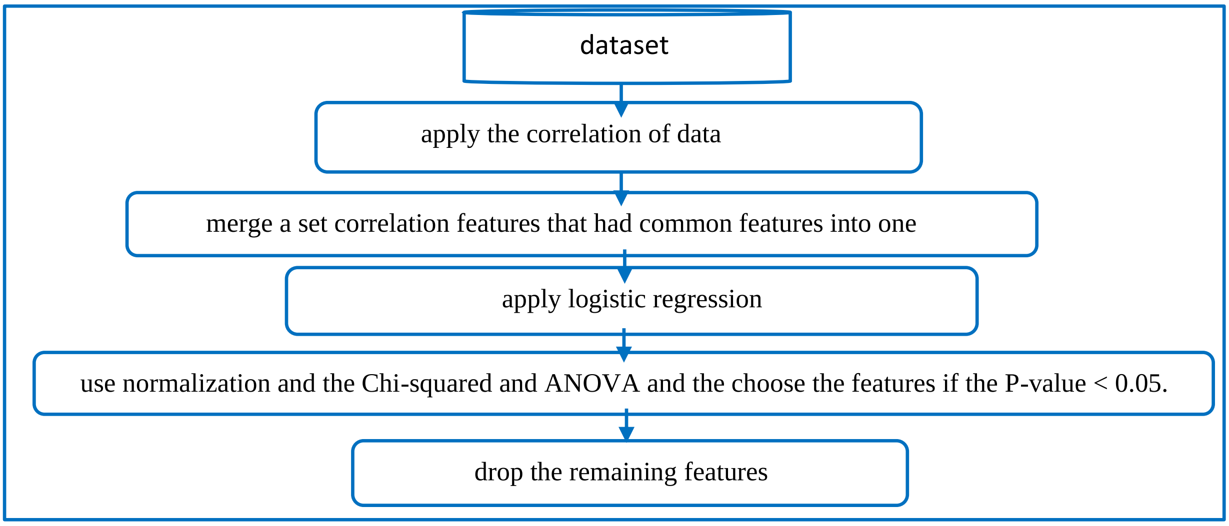

Figure 3: shows the proposed method

4.6 Split the Dataset: After performing the previous operations, the data becomes ready for training. At this stage, the data set is divided into two parts (training data = 0.8 and testing data = 0.2) (k=5). The K-Fold (k=5) class is used. It is given as arguments to the number of splits.

5. Applying Models and Results:

Data collection was implemented in Jupyter Notebook using Python code after pre-processing. Missing values and outliers were handled. The data set used in the study includes two tsunami cases (1) and (0) otherwise. Categorical data was digitized because some algorithms only deal with numerical values. The data set was segmented, and the machine learning algorithms were trained, identifying which ones were highly accurate.

5.1 Voting Classifier: It is one of the types of classification techniques known as group classifiers that rely on machine learning approaches as it works by combining some techniques to create a single technique that carries the power of the collected techniques, which gives a better prediction result [24]. A simplified voting technique was used by introducing two machine learning classifiers (HGBoost, and AdaBoost), which gave the best results in our study based on experiments. This new classifier works on a cumulative probability basis, as it uses the highest probability rate for the input models, which produces a probability value for class 0 or 1, as shown in Figure 4. The proposed method implements some processing steps to develop the data set, after which the best features are identified, and the soft voting classifier technique is used to provide the best results in classifying earthquake types with the best results.

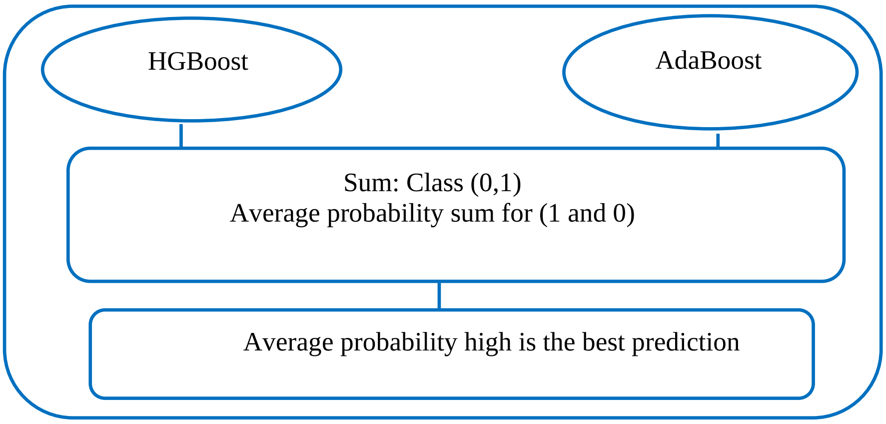

Figure 4: Display the proposed model (Soft Voting classifier).

5.2 Evaluation: Three popular algorithms in this stage are used to help identify data patterns resulting from the testing process. At this stage, the results will be verified by algorithms and their presentation. To identify classification errors specific to each algorithm to test its accuracy, relying on the confusion matrix. The new model (soft voting classifier) is built by selecting the two best algorithms in terms of results. Where the performance of the techniques used (HGBoost, AdaBoost, and KNN) was compared and the model (soft voting classifier) was built, the features gained from the proposed feature selection approach (CLR-ANCH) were passed to the data set segmentation process to be divided into a training part and a test part for the classifiers to be used. The performance of the new classifier was compared with other techniques used. **Table 4** represents the comparison results for these models. The soft voting classifier gave the best classification accuracy. **Table 4** displays the soft voting classifier gave the highest degree of accuracy (.094), F1 score (0.92), recall (0.94), and precision (0.92) because the voting classifier depends on integrating the two models into one model that carries the strength of these combined models, which leads to the best prediction accuracy. **Figure 5** displays the receiver operating characteristic (ROC) curves for HGBoost, AdaBoost, KNN, and the soft voting classifier.

Table 4: displays the result of models.

Models	confusion matrix		accuracy	F1_score	recall	precision
	TP	FP				
	FN	TN				
KNN	57	6	0.91	0.89	0.89	0.90
	7	87				
AdaBoost	59	7	0.92	0.90	0.92	0.89
	5	86				
HGBoost	61	8	0.92	0.91	0.95	0.88
	3	85				
Voting model	60	5	0.94	0.92	0.94	0.92
	4	88				

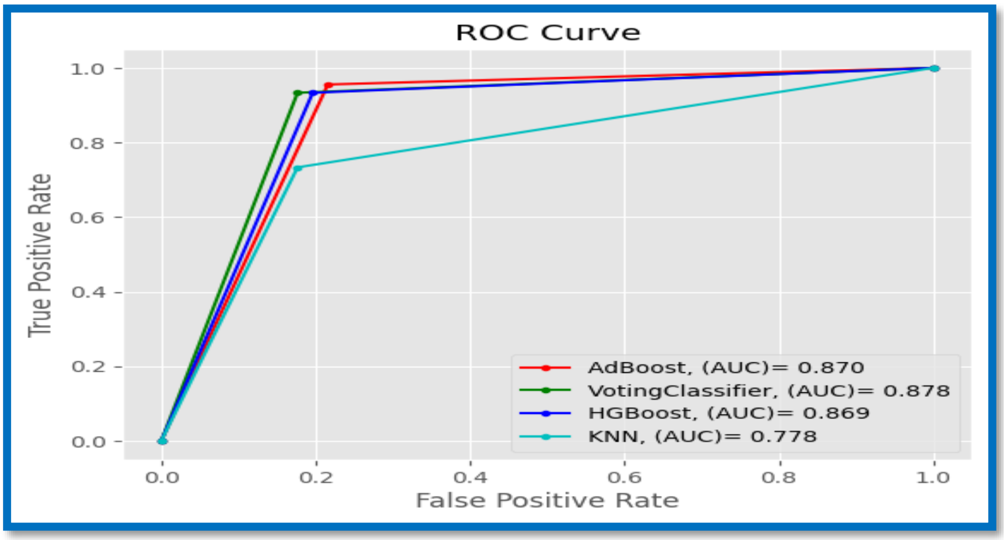

Figure 5: Display the ROC curves

6. Conclusion:

The proposed method consists of data pre-processing, normalization, and feature selection process through a correlation based on logistic regression as well as a normalization process with an analysis of variance (ANOVA) and chi-square (CL-ANCH), to select the best features to improve samples in the data set. The models proposed for earthquake prediction are (KNN, AdaBoost, HGBoost, and soft voting). The new model (Soft Voting) includes two techniques (HGBoost, and AdaBoost). The models are based on seismic indices calculated statistically and mathematically from the data set and later used as input for the proposed algorithms. Several metrics were utilized to evaluate the power of each algorithm. When comparing the performance of the techniques used and the new model, the highest performance and most power in the prediction process was for the proposed new model. It achieved an accuracy of 0.94, a recall of 0.94, precision of 0.92. and F1 score of 0.92. My future work for the study is to apply more feature selection models to a large dataset to improve diagnosis. In addition, using deep learning models.

7. References

[1] Acula, D. D. Detection of Earthquake Damages from Satellite Images using Gradient Boosting Algorithm with Decision Trees as Base Estimator. Acta Manilana, 2022; 70, 13-28.

[2] Chittora, P., Chakrabarti, T., Debnath, P., Gupta, A., Chakrabarti, P., Praveen, S. P., ... & Elngar, A. A. Experimental analysis of earthquake prediction using machine learning classifiers, curve fitting, and neural modeling. 2022; . doi.org/10.21203/rs.3.rs-1896823/v2

[3] Al Banna, M. H., Taher, K. A., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S., & Cho, G. H. Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges. IEEE Access, 2020; 8, 192880-192923.

[4] Salam, M. A., Ibrahim, L., & Abdelminaam, D. S. Earthquake prediction using hybrid machine learning techniques. International Journal of Advanced Computer Science and Applications, 2021; 12(5), 654-6652021. DOI: 10.1109/ICEEE.2019.8884562

[5] Atiyah, O. S., & Thalij, S. H. A comparison of Coivd-19 cases classification based on machine learning approaches. Iraqi Journal for Electrical and Electronic Engineering, (2022). 18(1), 139-143.

[6] Li, J., Zhang, C., Chen, X., Cao, Y., & Jia, R. Improving abstractive summarization with iterative representation. In 2020 International Joint Conference on Neural Networks (IJCNN), 2020; (pp. 1-8). IEEE. DOI: 10.1109/IJCNN48605.2020.9206950.

[7] Koehler, J., Li, W., Faber, J., Ruempker, G., & Srivastava, N. Testing the Potential of Deep Learning in Earthquake Forecasting. arXiv preprint arXiv, 2023; 2307.01812. doi.org/10.48550/arXiv.2307.01812

[8] Sajan, K. C., Bhusal, A., Gautam, D., & Rupakhety, R. Earthquake damage and rehabilitation intervention prediction using machine learning. Engineering Failure Analysis, 2023;144, 106949.

[9] Yang, F., Kefalas, M., Koch, M., Kononova, A. V., Qiao, Y., & Bäck, T. Auto-rep: an automated regression pipeline approach for high-efficiency earthquake prediction using lanl data. In 2022 14th International Conference on Computer and Automation Engineering (ICCAE), 2022; (pp. 127-134). IEEE.

[10] Berhich, A., Belouadha, F. Z., & Kabbaj, M. I. A location-dependent earthquake prediction using recurrent neural network algorithms. Soil Dynamics and Earthquake Engineering, 2022; 161, 107389.

[11] An, Z., Si, G., Tian, P., Li, J., Liang, X., Zhou, F., & Wang, X. Research on Earthquake Data Prediction Method Based on DIN–MLP Algorithm. Electronics, 2023; 12(16), 3519.

[12] Cho, S. B., & Won, H. H. Machine learning in DNA microarray analysis for cancer classification. In Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003; 19 (pp. 189-198).

[13] Freund, Y. Boosting a weak learning algorithm by majority. Information and Computation, 1995; 121(2), 256-285.

[14] Abd, N. S., Atiyah, O. S., Ahmed, M. T., & Bakhit, A. Digital Marketing Data Classification by Using Machine Learning Algorithms. Iraqi Journal for Electrical And Electronic Engineering, 2024; 20(1).

[15] C. Chauhan. Earthquake dataset, Kaggle Dataset, 2023. Available: https://www.kaggle.com/datasets/warcoder/earthquake-dataset.

[16] Atiyah, O. S., & Thalij, S. H. (2022). Evaluation of COVID-19 Cases based on Classification Algorithms in Machine Learning. Webology, 19(1), 4878-4887.

[17] Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., & Burton, H. V. Classifying earthquake damage to buildings using machine learning. Earthquake Spectra, 2020; 36(1), 183-208.

[18] Seeram, E. An overview of correlational research. Radiologic Technology, 2019; 91(2), 176-179.

[19] Real Statistics Using Excel, "Basic Concepts of Correlation", retrieved 22 February 2015. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#cite_note-RealCorBasic-13 .

[20] Lopez-Arevalo, I., Aldana-Bobadilla, E., Molina-Villegas, A., Galeana-Zapién, H., Muñiz-Sanchez, V., & Gausin-Valle, S. A memory-efficient encoding method for processing mixed-type data on machine learning. Entropy, 2020; 22(12), 1391.

[21] Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. Data normalization and standardization: a technical report. Mach Learn Tech Rep, 2014; 1(1), 1-6.

[22] Yu, L., Liu, H., Efficient Feature Selection via Analysis of Relevance and Redundancy, Journal of Machine Learning Research, 2004; 5, 1205-1224,

[23] Büyükkeçeci, M., & Okur, M. C. A comprehensive review of feature selection and feature selection stability in machine learning. Gazi University Journal of Science, 2022, 36(4), 1506-1520.

[24] Wang. H., H. Moayedi, and L. Kok Foong, Genetic algorithm hybridized with multilayer perceptron to have an economical slope stability design. Engineering with Computers, 2021; (37), 3067-3078.