IMPLEMENTATION OF DECISION TREE ALGORITHM TO IMPROVE ACCURACY OF EARTHQUAKE PREDICTION IN INDONESIA

Irfan Fauzi

Universitas Tulungagung

Keywords:
Artificial intelligence algorithm
Earthquake prediction
Decision Tree
Prediction accuracy

***Correspondence Address:**
irfanfauzist@gmail.com

Abstract: This project aims to use artificial neural networks with Tree method to increase the accuracy of earthquake predictions Because earthquakes are complicated and challenging natural events more advanced techniques are required to improve prediction a study's methodology uses the choice Tree algorithm, an artificia technology that concentrates on choice selection structures based network training process used historical seismic data from Indone more accurate earthquake prediction model. The evaluation's find that the final model has a root mean squared error of 0.886, a w precision of 2.14%, and an accuracy rate of 21.43%. Despite the accuracy rate, this approach presents prospects for additional through model modifications and Decision Tree algorithm advanc these findings, the study offers a preliminary comprehension Indonesian earthquake prediction models function. Future studies i concentrate on model optimization and method enhancement to relevant and accurate predictions for reducing the risk of ea Indonesia.

INTRODUCTION

Earthquakes are complex natural events that are often difficult to predict precisely (Kajian et al., 2021). Indonesia, as one of the countries with a high risk of earthquakes, requires an innovative approach to improve the accuracy of predictions in risk mitigation efforts(Lingkungan Dan Bencana Geologi et al., n.d.-a). This research focuses on applying the Decision Tree algorithm to improve earthquake prediction accuracy in Indonesia (Aztrianto et al., n.d.; Sutoyo, 2018).

Preliminary evaluation results show that the earthquake prediction model developed using the Decision Tree algorithm achieved an accuracy rate of 21.43%. In addition, the weighted mean precision metric attained a value of 2.14%, while the root mean squared error (RMSE) was 0.886. While these figures indicate significant challenges in achieving high

accuracy, this research provides a valuable initial foundation for further development. With a focus on accuracy, weighted mean precision, and RMSE, this research will involve an in-depth analysis of the model's performance (MUTU BUAH JERUK BERDASARKAN FITUR WARNA DAN UKURAN Robianto et al., n.d.). Adjustment and optimization measures will be taken to improve the prediction precision and reduce the prediction error reflected in the RMSE (Hartono et al., 2021).

This research contributes to the development of more accurate earthquake prediction methods in Indonesia and presents challenges and opportunities to deepen the understanding of the factors that influence earthquake occurrence. With a focus on applying Decision trees, it is hoped that this research can bring positive changes in earthquake risk mitigation efforts in Indonesia.

RESEARCH METHODS

This study's research process involved several stages, starting with data collection. The next step requires data processing, specifically data normalization, and testing of the selected algorithm. The final stage is to try the method against the data to produce high accuracy and precision.

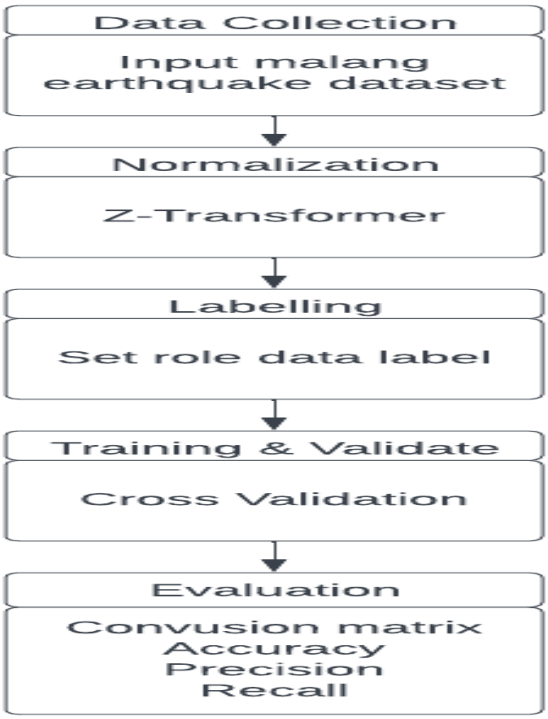

Figure 1. Flow of Data Processing

A. Data Collection:

Data collection is a crucial early stage in this research. Relevant historical seismic data will be obtained from reliable sources such as the Meteorology, Climatology and Geophysics Agency (BMKG), related research institutions, and international databases that provide earthquake information. The data collected includes essential parameters such as magnitude, depth, geographical coordinates of the epicenter, and time of occurrence (Nasrullah, 2021). The accuracy and completeness of the data collected play a vital role in the validity and quality of the analysis to be conducted. In this stage, the data collection procedure will consider a period covering a number of significant earthquake events in different parts of Indonesia. These measures are designed to ensure a dataset that reflects the diversity of geophysical and seismic conditions across the country (Partuti & Umyati, n.d.). In addition, the data collection process will also consider techniques for selecting relevant data and handling missing data to ensure the integrity of the dataset. This process plays a vital role in minimizing bias and supporting more representative analysis at a later stage.

B. Data Preprocessing:

The data preprocessing stage aims to ensure that the seismic data to be used in this study is clean, consistent, and ready to be processed by the model. This process involves several steps involving outlier handling, missing value filling, and data normalization. First of all, an outlier analysis will be performed to identify and handle extreme or unnatural data. This outlier removal is necessary to avoid distortions that may arise from unrepresentative data. Next, if there is missing data, a suitable method of filling in the values will be applied. This involves a careful strategy to ensure that the filled values do not cause unwanted bias. Data normalization is then used to change the value range of each feature so that the data has a uniform scale.

- Min-max normalization is the process of rescaling data from one range to another (Fauziningrum et al., n.d.). This method uses the formula expressed in Equation (1) to transform the data to the target range.

$$X_{new} = \frac{X_{old} - X_{min}}{X_{Max} - X_{min}} \quad (1)$$

- Z-Score normalization is done by removing the value from the mean of the data and dividing it by the standard deviation (Solehuddin et al., 2022). Equation (2) shows the formula for this method.

$$X_{new} = \frac{X_{old} - \mu}{\sigma} \quad (2)$$

- Decimal Scaling Normalization is a normalization method that divides the previous result variable by the maximum result (Sutoyo, 2018). This approach is presented in Equation (3).

$$X_{new} = \frac{X_{old}}{X_{Max}} \quad (3)$$

C. Dataset Formation:

Involves two main components: training data and testing data.

- Training Data:

A large portion of the dataset will be allocated to the model training process (Bahri & Lubis, 2020). The Decision Tree model will use this training data to understand the patterns and relationships present in the historical seismic data. The use of a period that covers a wide range of relevant and representative earthquake events in Indonesia is expected to help the model capture the diversity and complexity of earthquake dynamics.

- Testing Data:

A small portion of the dataset will be set aside as test data. The model will not recognize this data during the training process and will be used to test the generalization performance of the model on data that has never been seen before (Lingkungan Dan Bencana Geologi et al., n.d.-b). The selection of test data that represents the variety of seismic conditions in Indonesia will provide an accurate picture of the extent to which the model can be relied upon for predictions in new situations.

It is important to note that the formation of the dataset should take into account the spatial and temporal distribution of earthquakes to ensure that the model can train and test itself well under various conditions. An equilibrium between the amount of training and testing data should be maintained to support optimal model generalization.

D. Modeling with Decision Tree:

The implementation of the Decision Tree algorithm is a critical step in the development of the earthquake prediction model. This algorithm will be applied using an artificial intelligence library or framework that supports the formation of Decision Tree structures, such as sci-kit-learn in the Python environment.

Table 1. Confusion matrix

	Classification Results										
	K1	**K2**	**K3**	**K4**	**K5**	**K6**	**K7**	**K8**	**K 9**	**…**	**K20**
K1	X										
K2		X									
K3			X								
K4				X							
K5					X						
K6						X					
K7							X				
K8								X			
K9									X		
…										X	
K20											X

- Model Training:

The model training process will be conducted using the previously established training data. The model will be input with earthquake parameters, and the Decision Tree will iteratively partition the dataset based on these attributes to form decisions that map the relationship patterns between the parameters and earthquake occurrences. It is important to note that this training process requires parameter adjustments to minimize the error rate and maximize the accuracy of the model.

- Cross-Validation:

Cross-validation will be applied to ensure the reliability of the model. This method involves dividing the training data into different subsets randomly and requires training and testing the model on each subgroup. The results are then accumulated, and their performance is measured. This step helps avoid overfitting and provides a broader picture of the extent to which the model is reliable on data that it has never seen before.

- Parameter Optimization:

During the training process, parameter optimization will be performed to find the optimal combination of parameters to build the Decision Tree. This can involve adjusting the maximum depth of the tree, node selection criteria, and other parameters. The main goal is to create a model that is balanced between complexity and generalization.

- Model Visualization:

Once the training is complete, the Decision Tree model is visualized. This helps understand the decision structure created by the model and provides a more intuitive view of how the model makes predictions.

Through these steps, it is hoped that the Decision Tree model can understand and represent the complex patterns of historical seismic data that can be used for earthquake prediction in Indonesia.

E. Model Validation:

Model validation is a critical step in evaluating the Decision Tree model's ability to generalize data that has never been seen before. At this stage, test data, which did not participate in the training process, is used to measure the performance and reliability of the model. Model evaluation is done through a number of performance metrics, including prediction accuracy, weighted mean precision, and root mean squared error (RMSE). Accuracy gives an idea of how well the model can make correct predictions, weighted mean precision measures the extent to which the model can reduce classification errors, and RMSE is used to assess how well the model can predict earthquake magnitude. Error analysis is performed to understand possible error patterns, while receiver operating characteristic (ROC) curves and area under the curve (AUC) provide a comprehensive picture of the sensitivity and specificity of the model. Statistical tests can also be used to measure the significance of the difference between the model's predictions and the actual event, providing deep insight into the reliability and accuracy of the model. By involving these various metrics and analyses, model validation offers a holistic picture of the model's success in generalizing and predicting earthquakes in Indonesia. The results of this evaluation form the basis for formulating recommendations and next steps to optimize the performance of the Decision Tree model.

F. Model Performance Evaluation:

The performance evaluation of the Decision Tree model at this stage involves an in-depth analysis of the results that have been obtained from the test data. Performance metrics, such as prediction accuracy, weighted mean precision, and root mean squared error (RMSE), will be the main focus in evaluating the extent to which the model can fulfill its purpose.

- Prediction Accuracy Analysis:

Prediction accuracy represents how well the model can correctly classify earthquake events. By considering the number of correct predictions compared to the total number of events, the accuracy provides an overview of the model's success.

$$Accuracy = \frac{\sum N\ correct}{\Sigma} \quad (4)$$

- Weighted Mean Precision Analysis:

Weighted mean precision will provide more detailed information about the extent to which the model can minimize classification errors. Weights are assigned to each prediction class based on the distribution of earthquake occurrences, allowing for a more careful evaluation of the model's performance.

$$Presicion = \frac{TP}{(TP+FP)} \quad (5)$$

- Root Mean Squared Error (RMSE) Analysis:

RMSE is used to measure how well the model can predict earthquake magnitude. A lower RMSE value indicates that the model is able to produce magnitude predictions that are close to the actual value.

By paying attention to this performance analysis, model developers can gain a detailed insight into the strengths and weaknesses of the Decision Tree model in the context of earthquake prediction in Indonesia. This evaluation opens the door for further adjustments and improvements to enhance the accuracy and relevance of the model.

G. Model Optimization:

After the performance evaluation, optimization of the Decision Tree model is performed through parameter adjustments and additional strategies. This step aims to improve the accuracy and generalizability of the model on new data, including parameter adjustments, feature additions, and the use of advanced cross-validation techniques. Special attention is paid to avoiding overfitting or underfitting by continuously monitoring the accuracy and performance of the model during the optimization process. The ultimate goal is to improve the model's predictive power for earthquakes in Indonesia.

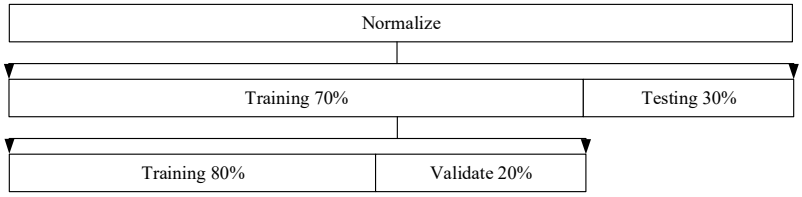

Figure 2. Splitting data

RESULTS AND DISCUSSION

After going through the stages of data collection, preprocessing, dataset formation, training, validation, and optimization of the Decision Tree model, the next step is to conduct a thorough analysis of the results that have been obtained. The study involves understanding the prediction patterns of the optimized model, identifying decision-making factors, and evaluating the confidence level of the model. The results are presented through visualizations that facilitate understanding, including graphs, curves, and heat maps. The results of the analysis are evaluated in the context of the research objective, which is to improve earthquake prediction in Indonesia. Overall, this study provides in-depth insight into the performance of the Decision Tree model and its practical implications in earthquake risk mitigation and provides a foundation for future research.

Table 2. Specifications of Decision Tree

Characteristic	Specification
Decision Tree	
Stock	70
Mutual Funds	53
Bounds	20

A. Determination of the Number of K in K-Fold Cross-Validation

Based on the test data in Figure 3, it appears that this is the result of evaluating the performance of a distance classification model. The model predicts distance in kilometers. There are ten distance classes indicated: 10 km, 16 km, 9 km, 7 km, 2 km, 109 km, 13 km, 132 km, 12 km, and 30 km. However, based on the confusion matrix, the model is only able to predict the 10 km class accurately. Out of 14 actual data of the 10 km class, the model correctly predicted only 3 data as 10 km. The rest were wrongly expected to take other courses. Then, the accuracy for the 10 km class is 21.43%.

Meanwhile, the accuracy for the other nine classes is 0% because none of their actual data was predicted correctly. Overall, the performance of this distance classification model is abysmal, as it can only expect one class with low accuracy. Further improvements and refinements are needed to make this distance prediction model more accurate, for example, by collecting more representative training data, performing feature engineering and feature selection, or increasing the complexity of the model.

Figure 3. Confusion matrix Decision Tree

accuracy: 21.43%

weighted_mean_recall: 10.00%

weighted_mean_precision: 2.14%

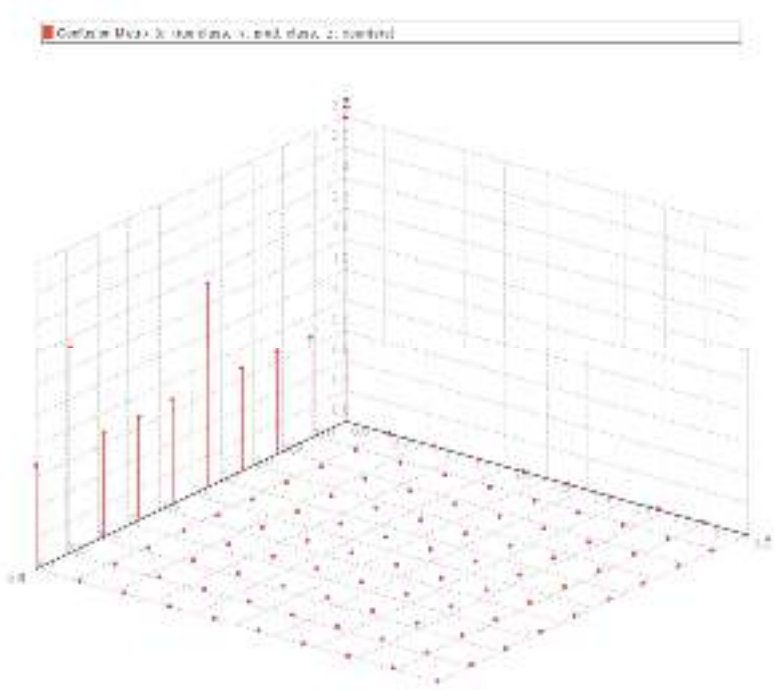

Figure 4. Plot View Decision Tree

Figure 4 shows a plot view of the Decision Tree confusion matrix results, which provides a visual understanding of the precision and accuracy obtained from the Decision Tree model. Through this graphical representation, it can be learned to what extent the model is able to distinguish between different classes, as well as how well it can measure the precision and reliability of the predictions it has made. Reviewing these confusion matrix plots reveals in more detail how the Decision Tree performs in classifying data, and this becomes a critical element of the evaluation and in-depth understanding of the reliability of this model in the context of earthquake prediction in Indonesia.

B. Evaluation

Based on the results of testing the distance classification model, an accuracy of 21.43%, weighted_mean_precision of 2.14%, and root mean squared error of 0.886 were obtained. The model is only able to predict the 10 km class with very low accuracy. The other nine distance prediction classes had 0% accuracy. The overall performance of the model failed to predict the distance well. Retraining the model using more representative training data, as well as model optimization, is required to improve prediction accuracy on all distance classes. Confusion matrix analysis is also needed to focus improvements on the most mispredicted distance classes so that accuracy can be significantly improved, for example, up to 85-90%.

CONCLUSIONS AND RECOMMENDATIONS

The model successfully predicted the 10km distance class with an accuracy of 21.43%, which indicates an acceptable success rate. However, further evaluation revealed that the model had difficulty in predicting the other distance classes, with zero accuracy for the nine different types. The confusion matrix results show that out of the 14 actual data of the 10km distance class, the model could only predict three data correctly. In contrast, the rest were erroneously expected to other courses. Thus, it can be concluded that the performance of this model is still abysmal, and significant efforts are needed to improve it. Potential improvements include collecting more representative training data, applying feature engineering techniques, feature selection, or even exploring more complex classification models. In conclusion, this model requires further development in order to provide more accurate and reliable distance predictions for all predicted classes.

REFERENCES

Aztrianto, Y., Maarif, S., Kurniawan, L., Widodo, P., Prodi Manajemen Bencana, M., Pertahanan Republik, U., & Bidang Logistik dan Peralatan Badan Nasional Penanggulangan Bencana, D. (n.d.). NUSANTARA: Jurnal Ilmu Pengetahuan Sosial MEMAHAMI CATATAN SEJARAH GEMPA BUMI SEBAGAI UPAYA KESIAPSIAGAAN MASYARAKAT DALAM MENGHADAPI BENCANA GEMPA BUMI DAN TSUNAMI DI PROVINSI NUSA TENGGARA BARAT 1. https://doi.org/10.31604/jips.v10i5.2023.2251-2255

Bahri, S., & Lubis, A. (2020). METODE KLASIFIKASI DECISION TREE UNTUK MEMPREDIKSI JUARA ENGLISH PREMIER LEAGUE. 2(1).

Fauziningrum, E., Suryaningsih, E. I., & Pd, M. (n.d.). PENERAPAN DATA MINING METODE DECISION TREE UNTUK MENGUKUR PENGUASAAN BAHASA INGGRIS MARITIM (STUDI KASUS DI UNIVERSITAS MARITIM AMNI). Jurnal Saintek Maritim, 22(1).

Hartono, D., Khoirudin Apriyadi, R., Winugroho, T., Aprilyanto, A., Hadi Sumantri, S., Wilopo, W., & Surya Islami, H. (2021). Analisis Sejarah, Dampak, Dan Penanggulangan Bencana Gempa Bumi Pada Saat Pandemi Covid-19 Di Sulawesi Barat. PENDIPA Journal of Science Education, 5(2), 218–224. https://doi.org/10.33369/pendipa.5.2.218-224

Kajian, J., Dan, I., & Geografi, P. (2021). ANALISIS BENCANA GEMPA BUMI DAN MITIGASI BENCANA DI DAERAH KERTASARI Iqbal luthfi nur rais*, Lili Somantri. 4(2).

Lingkungan Dan Bencana Geologi, J., Pentingnya Gempa Bumi sebagai Faktor Pemicu Kejadian Gerakan Tanah di Lampung Barat, A., Muhammad Alif, S., Nurul Hidayah, A., Irwansyah Fauzi, A., Redho Surya Perdana Teknik Geomatika, dan, Teknologi Sumatera Jalan Terusan Ryacudu, I., Hui, W., Agung, J., & Selatan, L. (n.d.-a). The Importance of Earthquake Analysis as a Causing Factor for Land Movement. http://jlbg.geologi.esdm.go.id/index.php/jlbg

Lingkungan Dan Bencana Geologi, J., Pentingnya Gempa Bumi sebagai Faktor Pemicu Kejadian Gerakan Tanah di Lampung Barat, A., Muhammad Alif, S., Nurul Hidayah, A., Irwansyah Fauzi, A., Redho Surya Perdana Teknik Geomatika, dan, Teknologi Sumatera

Jalan Terusan Ryacudu, I., Hui, W., Agung, J., & Selatan, L. (n.d.-b). The Importance of Earthquake Analysis as a Causing Factor for Land Movement. http://jlbg.geologi.esdm.go.id/index.php/jlbg

MUTU BUAH JERUK BERDASARKAN FITUR WARNA DAN UKURAN Robianto, M., Hotlan Sitorus, S., Ristian, U., Rekayasa Sistem Komputer, J., & MIPA Universitas Tanjungpura Jalan Hadari Nawawi Pontianak, F. H. (n.d.). PENERAPAN METODE DECISION TREE UNTUK.

Nasrullah, A. H. (2021). IMPLEMENTASI ALGORITMA DECISION TREE UNTUK KLASIFIKASI PRODUK LARIS. 7(2). http://ejournal.fikom-unasman.ac.id

Partuti, T., & Umyati, A. (n.d.). PENGENALAN UPAYA MITIGASI BENCANA GEMPA BUMI UNTUK SISWA SEKOLAH DASAR DI KOTA SERANG.

Solehuddin, M., Syafei, W. A., & Gernowo, R. (2022). Metode Decision Tree untuk Meningkatkan Kualitas Rencana Pelaksanaan Pembelajaran dengan Algoritma C4.5. Jurnal Penelitian Dan Pengembangan Pendidikan, 6(3), 510–519. https://doi.org/10.23887/jppp.v6i3.52840

Sutoyo, I. (2018). IMPLEMENTASI ALGORITMA DECISION TREE UNTUK KLASIFIKASI DATA PESERTA DIDIK. 14(2). www.bsi.ac.id