International Journal Software Engineering and Computer Science (IJSECS)
4 (1), 2024, 13-23
Published Online April 2024 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN : 2776-4869, E-ISSN : 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v4i1.2084.

RESEARCH ARTICLE Open Access

Classification of Potential Tsunami Disaster Due to Earthquakes in Indonesia Based on Machine Learning

Eri Mardiani *
Informatics Study Program, Faculty of Communication and Information Technology, Universitas Nasional, City of South Jakarta, Special Capital Region of Jakarta, Indonesia.
Corresponding Email: pipitn627@gmail.com.

Nur Rahmansyah
Animation Study Program, Politeknik Negeri Media Kreatif, South Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: nur_rahmansyah@polimedia.ac.id.

Sari Ningsih
Information Systems Study Program, Faculty of Communication and Information Technology, Universitas Nasional, City of South Jakarta, Special Capital Region of Jakarta, Indonesia.
Email: lectures.sariningsih@gmail.com.

Dhieka Avrilia Lantana
Digital Business Study Program, Faculty of Economics and Business, Universitas Nasional, City of South Jakarta, Special Capital Region of Jakarta, Indonesia.
Email: dhiekalantana12@gmail.com.

Nabila Puspita Wulandana
Accounting Study Program, Faculty of Economics and Business, UPN Veteran Jakarta, Central Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: 2110112084@mahasiswa.upnvj.ac.id.

Azzaleya Agashi Lombu
Accounting Study Program, Faculty of Economics and Business, UPN Veteran Jakarta, Central Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: 2110112086@mahasiswa.upnvj.ac.id.

Sisca Budyarti
Accounting Study Program, Faculty of Economics and Business, UPN Veteran Jakarta, Central Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: 2110112098@mahasiswa.upnvj.ac.id.

Received: January 11, 2024; Accepted: March 10, 2024; Published: April 1, 2024.

Abstract: Earthquakes and tsunamis pose significant threats to Indonesia due to its unique geological positioning at the convergence of four tectonic plates. This study focuses on classifying the potential occurrence of tsunami disasters following earthquakes using various data mining methods, including k-Nearest Neighbor (kNN), Naïve Bayes, Decision Tree and Ensemble Method, and Linear Regression. The research employs a qualitative approach to systematically understand and describe the context of natural disasters, utilizing both primary and secondary data collection techniques. Performance evaluation metrics such as Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, and Recall are utilized to assess the effectiveness of each method in predicting potential tsunami events. The findings reveal that the kNN method exhibits the highest performance, with an AUC of 94.4% and a precision of 82.8%, indicating robust predictive capabilities. However, misclassifications were observed, emphasizing the need for further refinement. Naïve Bayes also shows promising results with an AUC of 84.5% and precision of 78.6%. Decision Tree and Ensemble Method models, such as Random Forest and AdaBoost, demonstrate reasonable performance, with Random Forest achieving the highest AUC of 71.9%. Linear Regression is employed to explore the correlation between earthquake attributes and tsunami occurrence, revealing a weak relationship. Further research integrating advanced modeling approaches and additional earthquake attributes is recommended to enhance the predictive capabilities of tsunami risk assessment models. The study underscores the importance of employing diverse machine learning techniques and evaluating their performance metrics to refine the accuracy of tsunami prediction models, ultimately contributing to practical disaster preparedness and mitigation strategies.

Keywords: Data Mining; Earthquake; Tsunami; Orange Tools; K-Nearest Neighbor (KNN) Algorithm.

1. Introduction

Earthquakes are a natural phenomenon that often occurs in Indonesia due to the interaction of tectonic plates. Indonesia is located at the meeting point of the world's four tectonic plates, namely Eurasia, Australia, the Pacific, and the Philippines. The Australian and Pacific plates are a type of flexible oceanic plate, while the Eurasian plate is a type of rigid continental plate [1]. This meeting of tectonic plates causes active subduction and faulting on the seabed and land, potentially triggering earthquakes. If the earthquake's epicenter is in the sea, a tsunami can occur because of it. Most of Indonesia is prone to earthquakes and tsunamis, except for the Kalimantan area, where the potential for earthquake danger is relatively small [2]. Disaster preparedness is essential and must be built in every community group. Experience shows that disaster damage can be reduced significantly if everyone is better prepared to face disasters [3]. Therefore, disaster prevention becomes an important focus, where the public is asked to understand the warning signs of catastrophes and steps to reduce risks and prevent disasters from occurring [4]. One effort to improve community preparedness for earthquake disasters is to analyze earthquakes that have occurred before and look at the factors that differentiate the impacts resulting from each earthquake [5].

According to research by Prathivi and Rastri (2020), this way, people will be more alert if an earthquake occurs in the future and can prepare themselves to evacuate if the earthquake has the potential to cause a tsunami. This analysis can be carried out using the Orange Data Mining application [6]. Data mining or information mining is a software tool used to discover hidden patterns, trends, or rules that exist on a large-dimensional basis and create rules that are used to predict future behavior. Classification in information mining is a method of developing information to predict the value of a group of attributes. Using data mining using the k-NN algorithm, Naive Bayes, Decision Tree, Ensemble Method, and Linear Regression produces a set of provisions called rules, which are used as markers to predict the class of information you want to predict [7].

The dataset is used to see the impact of categorical data (Gender, Race/ethnicity, Parental level of education, Lunch, and Test preparation course) on numerical data (math, reading, and writing scores). And the dataset here is about the potential for tsunamis resulting from earthquakes to occur and not occur. With this dataset, we will know more about predictions of what will happen and what potential results from earthquakes and a more specific algorithm for knowing predictions of tsunami disasters [8]. Based on previous research, namely Dhendra Marutho (2019) produced a comparison using three algorithm models, Decision Tree, Naïve Bayes, and K-NN, using a cross-validation testing model using k-fold ten, resulting in an accuracy of 96.56% for the Decision Tree algorithm, then Naïve Bayes was 94.32%. Finally, K-NN was 95.98%, and

from the comparison of the Root Mean Squared Error (RMSE), the value that was close to zero was in the Decision Tree method, so it can be concluded that for research on water level datasets in Jakarta, the best use of this model is Decision Tree algorithm [8].

There are many technological advances available with computers. Computers not only help us get work done but can also be a fun and valuable tool to use [9]. To further maximize work, internet-based technology supports online work. In addition, with internet technology, work has become much faster and easier to complete than traditional computers [10]. The development of information technology has become so rapid. Internet technology connects thousands of individual and organizational computer networks worldwide [11]. The development of the times provides an increase in data, which means every human being needs more and more data [12]. With the large amount of data available, it is increasingly difficult to analyze it manually. Therefore, humans need data mining to collect exciting information from available data. Using data mining, you will discover interesting knowledge from large amounts of data stored in databases and warehouses [13]. Big Data Analytics can offer huge benefits to businesses that use it effectively, including better and faster decision-making, more customer satisfaction, discovering new opportunities, optimizing processes, and creating of better goods and services [14]. Literature studies regarding the Naive Bayes Algorithm for Sentiment analysis of Windows PhoneStore Application Reviews can be concluded by analyzing Windows Phone Store application users by classifying reviews into positive or negative opinion categories [13]. Data Mining uses Classification Techniques with 5 Algorithm Models [15].

The K-Nearest Neighbor (k-NN) algorithm is a model used to classify objects based on learning data closest to the object [7]. It can also be understood that k-nearest neighbor is one of the simplest and most widely used algorithms. Data points will be classified based on the similarity of certain groups of adjacent data points. So, this algorithm will provide competitive results [4]. K-NN equation [8].

$$d_i = \sqrt{\sum_{i=1}^{n} (x_{ij} - p_j)^2}$$

Information
di = Sample distance
xij = Knowledge sample data
pj = Jth var input data
n = Number of samples

One of the data mining methods is Naive Bayes classification. The naive Bayes Classifier is a classification method rooted in Bayes' theorem [16]. Naive Bayes is a classification method based on simple probabilities and designed to be used with the assumption that explanatory variables are independent of each other. In this algorithm, learning is emphasized when estimating probabilities. The Naive Bayes method aims to find probabilities when we know specific other probabilities. The results of data mining calculations using the Naive Bayes classification method will be more useful if the presentation is attractive and well understood by the data recipients [13]—the Naive Bayes equation [15].

$$P(H|X) = \frac{\mathrm{P(X|H).P(H)}}{\mathrm{P(X)}}$$

Explanation
X =Data with an unknown class
H = Hypothesis that the data belongs to a specific class
P(H|X) = Probability of hypothesis H given condition X (Posterior Probability)
P(H) =Probability of hypothesis H (Prior Probability)
P(X|H) = Probability of X given condition H
P(X) =Probability of X

A Decision Tree is a machine learning algorithm that employs a set of rules to make decisions with a tree-like structure that models possible outcomes, resource costs, utilities, and potential consequences or risks. The concept involves presenting the algorithm with conditional statements, including branches representing decision-making steps that can lead to favorable outcomes. This classification utilizes observations at nodes to determine targets at leaves. Decision Trees are one of the most popular classification methods because

they are easy for humans to interpret and can break down complex decision-making processes into simpler ones [15]. Decision Tree Equation [14].

$$Entropy(S) = \sum_{i=1}^{n} pi * \log_2 pi$$

Explanation:
S = Set of cases
A = Attribute
n = Number of partitions in S
Pi = Proportion of Si to S

The ensemble method is a machine learning algorithm that seeks the best predictive solution compared to other algorithms because this ensemble method uses several learning algorithms to achieve better predictive solutions than algorithms that can be obtained from just one constituent learning algorithm [17]. Unlike statistical ensembles in statistical mechanics that are usually always unbounded, Ensemble Learning consists only of a set of limited alternative models. Still, it usually allows for even more flexible structures among the alternative models. Evaluating predictions from an ensemble typically requires much more computation than evaluating forecasts from a single model [17]. The equation for Ensemble Method.

$$F\ (x) = \sum_{t=1}^{T} \propto_t h_t(x)$$

$h(x)$ = Base or weak classifier
$\propto$ = Learning rate
F(x) = Output, in the form of strong or final classification

Linear regression is one of the algorithms that models an equation to calculate estimation. This method aims to find patterns in numeric values, requiring numeric data to be processed with this algorithm model. Linear regression predicts the value of unknown data using other related and known data values. Mathematically, it models the unknown or dependent variable and the known or independent variable as a linear equation. Linear Regression Equation [18].

$$Y = a + bX$$

In this study, the following variables are defined:
Y = Dependent variable.
A = Constant.
B = Regression coefficient (slope); the magnitude of the response generated by the variable.
X = Independent variable.

Indicator Variables:
1) Title: Name/title given to the earthquake.
2) Magnitude: Magnitude of the earthquake.
3) Date_Time: Date and time.
4) Cdi: Maximum reported intensity for the event range.
5) Mmi: Estimated maximum instrumental intensity for the event.
6) Alert: Alert level.
7) Sig: Number describing the significance of the event. A larger number indicates a more significant event.
8) Net: Identifies the network considered as the preferred source of information for this event.
9) Nst: Total number of seismic stations used to determine the earthquake's location.
10) Dmin: Horizontal distance from the earthquake's epicenter to the nearest station.
11) Gap: Largest azimuthal gap between nearby stations (in degrees).
12) MagType: Method or algorithm used to calculate the desired magnitude for the event.
13) Depth: Depth at which the earthquake begins to rupture.

14) Latitude / Longitude: Coordinate system used to determine and describe the position or location of a place on the Earth's surface.
15) Location: Location within the country.
16) Continent: Continent where the country affected by the earthquake is located.
17) Country: Country affected by the earthquake.

This research aims to classify the potential of tsunami and earthquake disasters by comparing disasters occurring outside by applying methods such as kNN (k-Nearest Neighbour), Naïve Bayes, Decision Tree and Ensemble Method, and Linear Regression. Furthermore, a comparative analysis of the models will be conducted to ensure the accuracy level and contribute to determining the accuracy performance of several data mining methods, including kNN (k-Nearest Neighbour), Naïve Bayes, Decision Tree and Ensemble Method, and Linear Regression. This analysis provides essential information on the role of policies, insurance companies, and community participation in disasters before and after. For instance, poverty, job access, and political access lead some communities to reside in earthquake-prone areas with inadequate housing construction for earthquake resilience. Lack of trust in early warnings of earthquake and tsunami hazards leads many people to disregard them.

2. Research Method

This research uses qualitative methods to explain systematically [19]. This research is qualitative, where qualitative research aims to understand, find meaning, and define a context by describing in detail and depth the context being researched. Qualitative research is a mechanism for searching and collecting, as well as comprehensive analysis and interpretation to produce new understanding. Descriptive analysis will be used in this research to provide an understanding [11]. The data collected for this research is primary and secondary. This research's secondary data collection techniques were obtained from online media and other sources. Researchers will use literature studies to describe and analyze the data collected and related to this research [18].

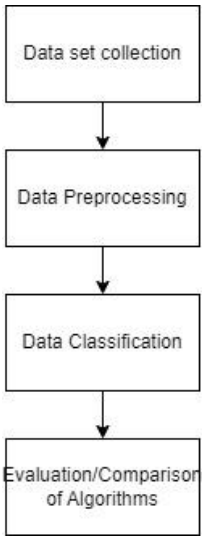

Figure 1. Research Stage

This research has four stages: dataset collection, data testing, prediction process, performance evaluation, and method comparison results [20].

3. Result and Discussion

3.1 Results

The method comparison results are presented in a Confusion Matrix, Precision, Recall, and F-measure table for each method.

3.1.1 K-Nearest Neighbor (K-NN)

In the image below, you can see that we used Fix Proportion Data of 10%. We use 2 neighbor approaches. Then, using Data Train, as much as 358 data with 15 features and 0 missing data. In the preprocess section, this project selects the standard normalized features. Predictions are used to determine whether the data used is accurate and reasonable.

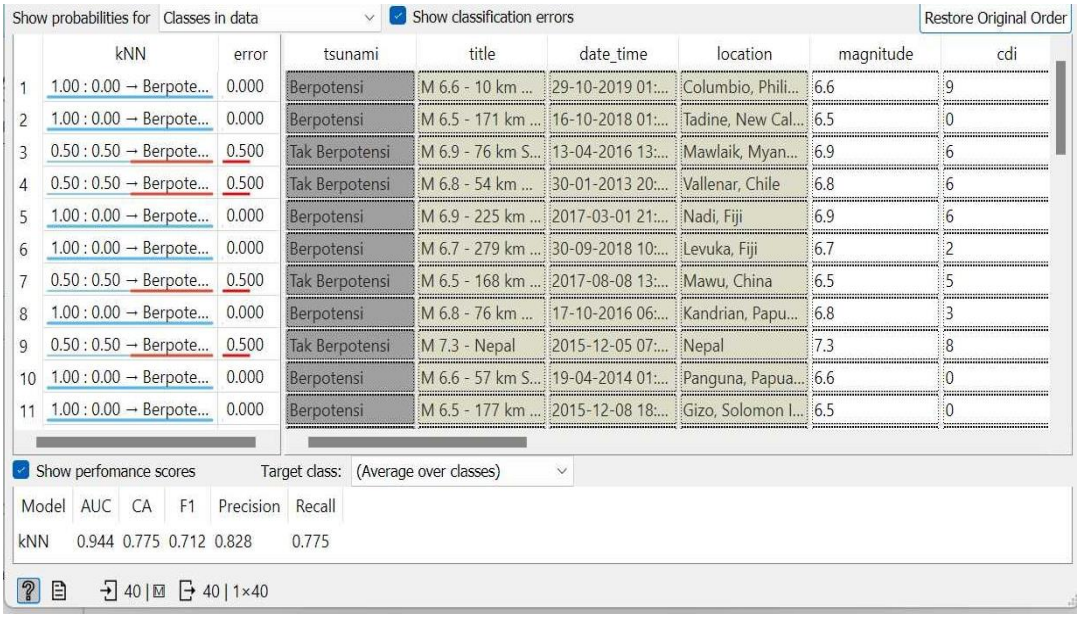

Figure 2. K-Nearest Neighbor (K-NN)

From Figure 2 above, it can be seen that predictions in the kNN model show AUC of 94.4%, CA of 77.5%, F1 of 71.2%, and Recall of 77.5%. This shows that the kNN Predictions model shows good predictions because it can be seen from its AUC of 94.4%. The overall prediction is quite strong, and the data model is good because it is above 50%. This means that the analysis of earthquakes with potential for tsunamis in the kNN model has good predictions. And the results of the Data Table, which is connected to the Confusion Matrix, show that the KNN prediction of potential earthquakes with actual potential earthquakes is 29 data. The results of the Data Table connected to the Confusion Matrix show a prediction error of 9 data because kNN predicts potential, the Actual should not have potential. The results of the Data Table which is connected to the Confusion Matrix show that the kNN prediction of an earthquake that has no potential with the Actual earthquake does not have the potential of 2 pieces of data. From the Scatter Plot above, the variables depth (depth) and magnitude (strength) influence whether an earthquake has the potential for a tsunami. From the Scatter Plot results, it can be interpreted that there is more potential for a tsunami at a depth below 200 km and at a magnitude of 6.4 – 8.2 SR. So, these two variables can influence whether or not an earthquake has the potential to cause a tsunami.

3.1.2 Naive Bayes

Figure 3 shows the results of the Test and Score from the Naïve Bayes method. Obtained AUC of 0.845, CA of 0.776, F1 of 0.780, Precision of 0.786, and recall of 0.776. The precision figure shows that this method has a prediction level of up to 79%, so it is almost accurate.

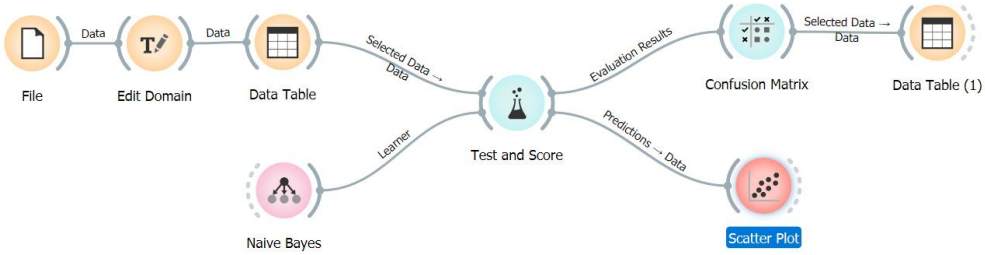

Figure 3. Naïve Bayes

In the Confusion Matrix, there is a similarity between the prediction (Predicted) and the actual (Actual), namely 232. This value shows that the prediction results for potential tsunamis have been proven correct (actual). There is an error in the prediction, namely 52. This value predicts that 52 has no potential for a tsunami, contrary to the actual value, which has the potential for a tsunami. The expected value of 37 is similar to the actual value. This means that the predicted value of an earthquake with the potential for a tsunami is the same as the actual event. There is an error in the prediction, namely 77. This value predicts that 77 has no potential for a tsunami, contrary to the actual value, which has the potential for a tsunami. The Scatter Plot below shows the distribution of earthquakes that can cause tsunamis. This Scatter Plot uses depth as an Axis.

3.1.3 Decision Tree and Ensemble Method

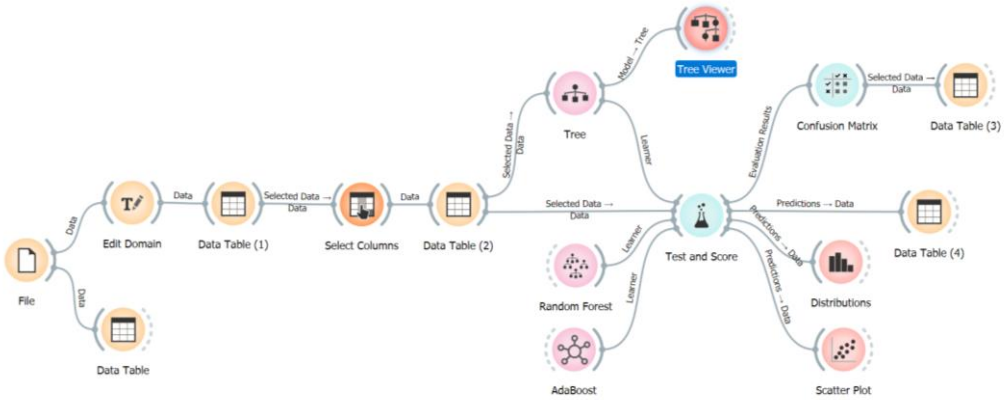

Figure 4. Decision Tree and Ensemble Method

Table 1. Performance Evaluation of Algorithm Models for Classifying Potential Earthquakes and Tsunamis

Model	AUC	CA	F1	Precision	Recall
Tree	0.693	0.726	0.729	0.731	0.726
Random Forest	0.719	0.709	0.701	0.695	0.709
AdaBoost	0.629	0.691	0.691	0.690	0.691

Of the three algorithm models, namely Tree, Random Forest, and AdaBoost, the best algorithm model is Random Forest because, seen from its AUC, it is 0.872 or 87%. The Tree algorithm model has a result of 0.801 or 80%, and the smallest is the Ada Boost model, which is 0.794 or 79%. The results of the Data Table, which is connected to the Confusion Matrix, show that there are 244 data in the kNN prediction of potential earthquakes with Actual potential earthquakes. The Data Table, which is connected to the Confusion Matrix, shows a prediction error of 40 data because the kNN predicts no potential; the Actual should have Potential.The data table connected to the confusion matrix shows 31 prediction errors because kNN predicts Potential, but the actual should not have Potential. The Data Table connected to the Confusion Matrix shows that the kNN prediction of earthquakes has no potential, with Actual earthquakes having no potential of 83 data.

Meanwhile, in the Tree Viewer above, the potential or non-potential for a tsunami due to an earthquake is divided based on the continent. The first benchmark is the Potential for a tsunami to occur, with a percentage level of 71.4%. In the scatter plot, Depth and Magnitude variables influence the Tsunami, with the highest point being Depth = 598,100 and Magnitude = 8.3 Potential. The Depth and Magnitude variables affect the Tsunami, with the lowest point being Depth = 2,700 and Magnitude = 6.5, entering No Potential. For the scatter plot above, it is a Tsunami distribution with Depth and Magnitude variables. Look at the No Potential and Potential predictions. You can see the distribution; on average, a high Dept will have the Potential for a Tsunami (blue), and for the No Potential distribution, it is only spread over a higher Depth. From 400km (red). Don't look at the Magnitude, but the magnitude range 6.5 - 7.9 SR has a high tsunami potential, and 6.5 - 6.9 SR is at a depth of more than 400km. For Variable Random Forest influencing the Tsunami, Random Forest predictions for Potential predict Potential for 265 data and Not Potential for 45 data. Apart from that, Random Forest predictions with no potential show potential predictions of 19 data and no potential of 60 data.

3.1.4 Linear Regression

From Figure 5, the correlations can be seen for magnitude, cdi, mmi, sig, nst, dmin, gap, depth, latitude, and longitude. We chose the target tsunami. In Correlations, the highest-ranking variables are Magnitude and Depth. If you use several folds of 2, R2, or the correlation between Magnitude and Depth data, the tsunami becomes 4.4%, which means an increase of 0.6%.

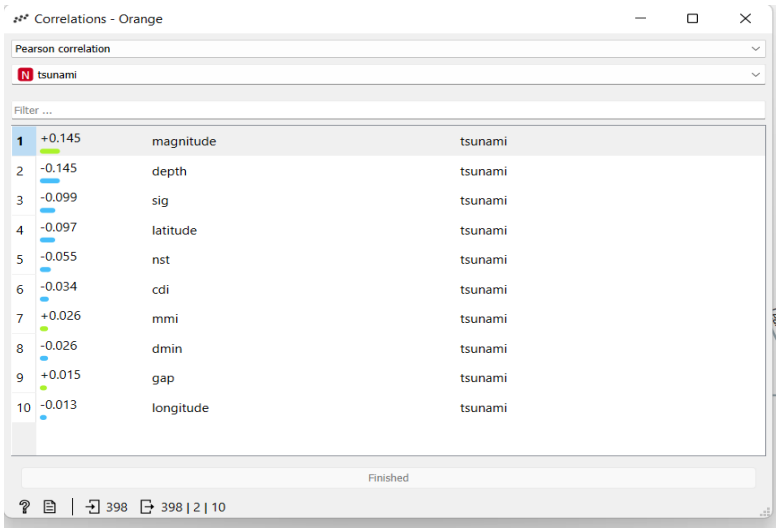

Figure 5. Linear Regression

Table 2. Linear regression

Model	MSE	RMSE	MAE	R2
Linear Regression	0.195	0.442	0.390	0.044

In Table 2 of Linear Regression, it can be seen that R2 is 0.038 or 3.8%, where with this data, we calculate the level of Regression strength, and we get a Regression strength level of 3.8%. If you use several folds of 2, R2, or the correlation between Magnitude and Depth data, the tsunami becomes 4.4%, meaning an increase of 0.6%. The X-axis, depth, and the Y-axis, magnitude, show the magnitude and depth variables influencing the tsunami. It can also be interpreted that a depth below 200 km and a magnitude of 6.5 - 7.8 SR has a more significant potential for a tsunami. These two variables can influence whether a tsunami will occur. Analysis of this dataset looking at the potential for tsunamis to happen and not to occur was carried out in this project using 4 (four) methods, namely kNN (k-Nearest Neighbor), Naïve Bayes, Tree Decision and Ensemble Method, and Linear Regression. From comparing the results of all methods, the kNN method is considered very good. We are judging from the AUC and Precision values of kNN, namely 0.944 and 0.828. The prediction results from Precision shows a figure of 83%, which means that the prediction results are considered almost accurate compared to other methods. These values prove that the Magnitude and Depth variables influence the occurrence or non-occurrence of a potential tsunami due to an earthquake.

3.2 Discussion

The analysis conducted in this project aimed to assess the potential occurrence of tsunamis following earthquakes using four different methods: k-Nearest Neighbor (kNN), Naïve Bayes, Decision Tree and Ensemble Method, and Linear Regression. Each method was evaluated based on several performance metrics, including Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, and Recall. The kNN method demonstrated notable performance, as indicated by its high AUC of 94.4% and precision of 82.8%. This suggests that the kNN model's predictions were robust and accurate, with a precision value above 80%. Furthermore, the Confusion Matrix revealed that the kNN model correctly predicted 244 instances of potential tsunami events out of 244 actual occurrences, indicating a robust predictive capability. However, there were some misclassifications, with 40 cases falsely predicted as non-potential tsunamis and 31 instances falsely predicted as potential tsunamis.

Similarly, the Naïve Bayes method exhibited promising results, with an AUC of 84.5% and Precision of 78.6%. The Confusion Matrix showed that the model correctly classified 232 instances of potential tsunamis and 37 cases of non-potential tsunamis out of 269 actual occurrences. However, 52 instances were falsely

predicted as non-potential tsunamis and 77 instances were falsely predicted as potential tsunamis, indicating some misclassifications.

The Decision Tree and Ensemble Method, represented by Random Forest and AdaBoost models, also showed reasonable performance. Among the three models evaluated (Tree, Random Forest, and AdaBoost), Random Forest exhibited the highest AUC of 71.9%, followed by Tree with 69.3% and AdaBoost with 62.9%. While Random Forest has the highest AUC, further analysis is required to determine the most suitable model for predicting potential tsunamis accurately. Additionally, Linear Regression was employed to explore the relationship between earthquake attributes (such as magnitude and depth) and the occurrence of tsunamis. The analysis revealed a correlation coefficient (R2) of 3.8%, suggesting a weak relationship between earthquake attributes and tsunami occurrence. However, the correlation increased to 4.4% when considering specific folds, indicating a slight improvement in predictive capability. The analysis suggests that the kNN method performed the best among the evaluated models in predicting potential tsunamis following earthquakes. The study also highlights the importance of considering various machine learning techniques and assessing their performance metrics to improve the accuracy of tsunami prediction models. Further research incorporating more sophisticated modeling approaches and additional earthquake attributes may enhance the predictive capability of tsunami risk assessment models.

4. Related Work

The research conducted in this project aimed to evaluate the potential occurrence of tsunamis following earthquakes using four distinct methodologies: k-Nearest Neighbor (kNN), Naïve Bayes, Decision Tree and Ensemble Method, and Linear Regression. Each method undergoes assessment based on several performance metrics, including Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, and Recall. The kNN method emerged with notable performance, characterized by its high AUC of 94.4% and Precision of 82.8%, suggesting robust and accurate predictions. The Confusion Matrix indicated that the kNN model correctly predicted 244 instances of potential tsunami events out of 244 actual occurrences, reflecting its predictive solid capability. However, there were instances of misclassification, with 40 cases falsely predicted as non-potential tsunamis and 31 instances falsely predicted as potential tsunamis.

Similarly, the Naïve Bayes method exhibited promising results, with an AUC of 84.5% and Precision of 78.6%. Yet, the Confusion Matrix showed misclassifications of 52 instances falsely predicted as non-potential tsunamis and 77 instances falsely predicted as potential tsunamis. The Decision Tree and Ensemble Method, exemplified by models like Random Forest and AdaBoost, demonstrated reasonable performance, with Random Forest displaying the highest AUC of 71.9%. However, further analysis is needed to identify the most suitable model for accurately predicting potential tsunamis. Additionally, Linear Regression was utilized to explore the relationship between earthquake attributes and tsunami occurrence, revealing a correlation coefficient (R2) of 3.8%. While this indicates a weak relationship, considering specific folds improved the correlation to 4.4%, hinting at a slight improvement in predictive capability. The study underscores the significance of employing diverse machine learning techniques and evaluating their performance metrics to refine the accuracy of tsunami prediction models, advocating for further research integrating advanced modeling approaches and additional earthquake attributes to bolster tsunami risk assessment models.

Furthermore, the study emphasizes the need for further research integrating advanced modeling approaches and additional earthquake attributes to enhance tsunami risk assessment models. This aligns with the broader research landscape of improving tsunami predictions through machine learning techniques. Recent studies have explored the use of Convolutional Neural Networks for tsunami early warning systems [21], machine learning-based tsunami inundation prediction from offshore observations [22], and estimation of tsunami bore forces using Extreme Learning Machines [23]. These advances underscore the continuous efforts to leverage sophisticated algorithms and data sources to refine tsunami prediction models and effectively mitigate the impact of earthquake-generated tsunamis.

5. Conclusion

For analysis using the Orange application, the requirement is to check that the data is complete (no missing data) because missing or incomplete data will affect the analysis results. If there is missing data, it will take longer to analyze the results. Several methods need to be added and developed so that the results of data mining analysis can be compared with other, more accurate methods. Machine learning techniques

and evaluating their performance metrics to improve the accuracy of tsunami prediction models. In addition, further research integrating more sophisticated modeling approaches and additional earthquake attributes could improve the predictive capabilities of tsunami risk assessment models.

In conclusion, the analysis carried out in this project aims to assess the potential for a tsunami after an earthquake using various machine learning algorithms, including k-Nearest Neighbor (kNN), Naïve Bayes, Decision Tree and Ensemble Method, and Linear Regression. The kNN method showed the best performance among the evaluated models, with a high AUC of 94.4% and precision of 82.8%, indicating robust and accurate predictions. However, several misclassifications highlight the need for further improvements. Other methods, such as Naïve Bayes, Decision Tree, and Ensemble Methods, show promising results but require additional improvements. Linear Regression was used to explore the relationship between earthquake attributes and tsunami events, which showed a weak correlation. Further research incorporating more sophisticated modeling approaches and additional earthquake attributes could improve the predictive capabilities of tsunami risk assessment models.

References

[1] Opilah, B. S., Karyadi, B., Johan, H., & Mayub, A. (2023). Model Integrasi Mitigasi Bencana Gempa Bumi pada Konsep Gelombang. *Jurnal Pendidikan Fisika*, *11*(1), 28-39. DOI: http://dx.doi.org/10.24127/jpf.v11i1.6754.

[2] Syafitri, Y., Bahtiar, B., & Didik, L. A. (2019). Analisis pergeseran lempeng bumi yang meningkatkan potensi terjadinya gempa bumi di pulau Lombok. *Konstan-Jurnal Fisika Dan Pendidikan Fisika*, *4*(2), 139-146. DOI: https://doi.org/10.20414/konstan.v4i2.43.

[3] Hadi, H., Agustina, S., & Subhani, A. (2019). Penguatan kesiapsiagaan stakeholder dalam pengurangan risiko bencana alam gempabumi. *Geodika: Jurnal Kajian Ilmu dan Pendidikan Geografi*, *3*(1), 30-40. DOI: https://doi.org/10.29408/geodika.v3i1.1476.

[4] Deswita, D., Yuliharni, S., & Efniyati, N. N. (2023). STUDI KASUS: GAMBARAN KESIAPSIAGAAN REMAJA MENGHADAPI GEMPA BUMI DAN TSUNAMI. *Jurnal'Aisyiyah Medika*, *8*(2). DOI: https://doi.org/10.36729/jam.v8i2.1112.

[5] Puspitasari, L., Prastistho, B., & Prasetya, J. D. (2023). MODEL KESIAPSIAGAAN KELUARGA TERHADAP ANCAMAN BAHAYA BENCANA GEMPA BUMI DESA CONDONGCATUR, KAPANEWON DEPOK, KABUPATEN SLEMAN, DI YOGYAKARTA. *Indonesian Journal of Environment and Disaster*, *2*(1), 1-9. DOI: https://doi.org/10.20961/ijed.v2i1.479

[6] Prathivi, R. (2020). Optimasi Algoritme Naive Bayes Untuk Klasifikasi Data Gempa Bumi di Indonesia Berdasarkan Hiposentrum. *Telematika*, *13*(1), 36-43.

[7] Marutho, D. (2019). Perbandingan Metode Naive Bayes, KNN, Decision Tree Pada Laporan Water Level Jakarta. *Jurnal Ilmiah Infokam*, *15*(2). DOI: https://doi.org/10.53845/infokam.v15i2.175.

[8] Giri, G. A. V. M. (2018). Klasifikasi Musik Berdasarkan Genre dengan Metode K-Nearest Neighbor. *Jurnal Ilmu Komputer*, *11*(2), 104-108.

[9] Rahmansyah, N., Muliyani, D., Mardiani, E., & Rahman, A. (2022). Perancangan Sistem Transaksi Berbasis Web pada UKM Pangkas Rambut Tasik. *Jurnal Sistem Informasi Bisnis (JUNSIBI)*, *3*(1), 22-31. DOI: https://doi.org/10.55122/junsibi.v3i1.412

[10] Mardiani, E., Rahmansyah, N., & Kurniati, I. (2023). Website Design at SDN Cipete Utara 07. *SITEKIN: Jurnal Sains, Teknologi Dan Industri*, *20*(2), 891-898. DOI: http://dx.doi.org/10.24014/sitekin.v20i2.22438.

[11] Mardiani, E., & Ramadhan, F. A. (2023). Design Information System Sales of Nuts and Bolts at PT. Catur Naga Steelindo. *SITEKIN: Jurnal Sains, Teknologi dan Industri*, *20*(2), 729-735. DOI: http://dx.doi.org/10.24014/sitekin.v20i2.21948.

[12] Matondang, N., Mardiani, E., Wahyudi, P., & Saebani, A. (2019). Aplikasi Komputer. *Jakarta: Mitra Wacana Media*.

[13] Indriyawati, H., & Khoirudin. (2019). Penerapan Metode Regresi Linier dalam Koherensi Pengolahan Data Bahan Baku Tiandra Store Guna Meningkatkan Mutu Produksi. *Sintak*, *3*(2).

[14] Pratama, F. H., Triayudi, A., & Mardiani, E. (2022). Data mining k-medoids dan k-means untuk pengelompokan potensi produksi kelapa sawit di indonesia. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *7*(4), 1294-1310.

[15] Hozairi, H., Anwari, A., & Alim, S. (2021). Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes. *Netw. Eng. Res. Oper*, *6*(2), 133.

[16] Djamaludin, M. A., Triayudi, A., & Mardiani, E. (2022). Analisis Sentimen Tweet KRI Nanggala 402 di Twitter menggunakan Metode Naïve Bayes Classifier. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *6*(2), 161–166. https://doi.org/10.35870/jtik.v6i2.398.

[17] Mardiani, E., Rahmansyah, N., Ningsih, S., Lantana, D. A., Wirawan, A. S. P., Wijaya, S. A., & Putri, D. N. (2023). Komparasi Metode Knn, Naive Bayes, Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar Sma. *Innovative: Journal Of Social Science Research*, *3*(2), 13880-13892. DOI: https://doi.org/10.31004/innovative.v3i2.1949.

[18] Indriyanti, I., Ichsan, N., Fatah, H., Wahyuni, T., & Ermawati, E. (2022). Implementasi Orange Data Mining Untuk Prediksi Harga Bitcoin. *Jurnal Responsif: Riset Sains dan Informatika*, *4*(2), 118-125. https://doi.org/10.51977/jti.v4i2.762.

[19] Basuki, B., & Rejeki, S. (2021). PENDEKATAN DAN METODA PENELITIAN FENOMENA GEMPA BUMI. *Jurnal Teknik Sipil Giratory UPGRIS*, *2*(2), 52-65. DOI: https://doi.org/10.26877/giratory.v2i2.10343.

[20] Sarofi, M. A. A., Irhamah, I., & Mukarromah, A. (2020). Identifikasi Genre Musik dengan Menggunakan Metode Random Forest. *Jurnal Sains dan Seni ITS*, *9*(1), D79-D86. http://dx.doi.org/10.12962/j23373520.v9i1.51311.

[21] Rim, D., Baraldi, R., Liu, C. M., LeVeque, R. J., & Terada, K. (2022). Tsunami early warning from global navigation satellite system data using convolutional neural networks. *Geophysical Research Letters*, *49*(20). https://doi.org/10.1029/2022gl099511

[22] Mulia, I. E., Ueda, N., Miyoshi, T., Gusman, A. R., & Satake, K. (2022). Machine learning-based tsunami inundation prediction derived from offshore observations. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-33253-5

[23] Mazinani, I., Ismail, Z., Shamshirband, S., Hashim, A. M., Mansourvar, M., & Zalnezhad, E. (2016). Estimation of tsunami bore forces on a coastal bridge using an extreme learning machine. *Entropy*, *18*(5), 167. https://doi.org/10.3390/e18050167.

