

Seismic trend analysis: a data mining approach for pattern prediction

Laberiano Andrade-Arenas¹, Cesar Yactayo-Arias²

¹Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú

²Departamento de Estudios Generales, Universidad Continental, Lima, Perú

Article Info

Article history:

Received Jan 24, 2024

Revised Feb 13, 2024

Accepted Feb 28, 2024

Keywords:

Data mining

Geophysical variability

Grouping

Prevention

Seismic events

ABSTRACT

In the global context, seismic movements represent a constant for the population due to geophysical variability and other factors that make them possible, carrying with them the risk of losing innocent lives. The main purpose of our research is to apply data mining techniques to prevent seismic events of any magnitude to anticipate and mitigate future events. In the development of the research, we applied knowledge discovery database methodology. The clustering analysis results revealed the following: cluster 0 encompassed 45 items, with average magnitude of 0.230, representing 15.5% of the total events. Cluster 1 comprised 56 items with average magnitude of 0.156, equivalent to 19.2% of the total. Cluster 2, the largest, consisted of 94 items with average magnitude of 0.156, constituting 32.3% of the total seismic events. Cluster 3 was composed of 54 items, with average magnitude of 0.155, representing 18.3% of the total. Lastly, cluster 4 included 42 items, with average magnitude of 0.155, representing 14.5% of the total. In conclusion, cluster 3 emerged as the most significant, with 94 events and average magnitude of 0.141, equivalent to 32.3% of the total seismic events. This discovery underscores the need to utilize data mining techniques for earthquake prediction, enabling proactive measures against potential events, which are frequent in various geographic areas.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Cesar Yactayo-Arias

Departamento de Estudios Generales, Universidad Continental

Lima, Perú

Email: cyactayo@continental.edu.pe

1. INTRODUCTION

In the current international context, the study and analysis of seismic activity have become crucial to understanding and mitigating the risks associated with seismic events. The increasing complexity of seismic patterns and the variability in the magnitude and intensity of earthquakes pose significant challenges for seismology experts and for communities exposed to such events. The need for advanced strategies to anticipate and understand the evolution of seismic patterns has led to a growing interest in the application of data mining techniques [1], [2].

Seismic activity, an inherently unpredictable phenomenon, presents considerable challenges to seismic scientists and practitioners. As seismic events of varying magnitudes continue to affect regions around the world, the need to understand their behavioral patterns becomes more urgent [3], [4]. Manual analysis of extensive seismic data sets over decades proves to be a monumental task and often insufficient to reveal the underlying complexities of seismic activity. Geographic variability, temporal evolution, and the interrelationship of multiple factors make the task of identifying meaningful patterns challenging. The need for advanced methods of analysis becomes evident, and it is in this context that data mining emerges as a

powerful tool to unravel the secrets buried in seismic data [5], [6]. In this scenario, the application of data mining concepts presents itself as a promising solution. Data mining, a discipline that combines statistical, artificial intelligence, and machine learning techniques, offers the ability to explore large data sets for patterns, correlations, and trends. The complexity of seismic activity, which is characterized by multiple interrelated variables, makes data mining tools especially relevant. By using advanced algorithms, such as clustering and regression, we can identify patterns not evident to the naked eye and reveal relationships that might escape conventional analysis [7], [8].

The rationale for this study is based on the critical importance of improving our abilities to understand and forecast seismic activity. Seismic events can have devastating consequences, affecting both human populations and critical infrastructure. Traditional analysis often focuses on understanding past events, but the ability to forecast and mitigate the impact of future events is essential. Data mining offers a promising avenue for uncovering hidden patterns and meaningful correlations in seismic data over time [9], [10]. By improving our predictive capabilities, we can advance infrastructure planning, community preparedness, and response strategies, thereby contributing to the safety and resilience of seismically prone areas. Ultimately, this research seeks not only to understand the complexity of past seismic activity but also to provide practical tools to address future challenges associated with seismic events [11]. Data mining, by exploring hidden patterns and relationships in seismic data, allows for a deeper understanding of the factors influencing seismic activity. Identifying early indicators, predicting emerging patterns, and assessing the probability of significant seismic events become possible through the application of data mining algorithms. By improving our predictive capabilities, we can advance infrastructure planning, community preparedness, and response strategies, thus contributing to the safety and resilience of seismic-prone areas [12], [13]. Ultimately, this research seeks not only to understand the complexity of past seismic activity but also to provide practical tools to address future challenges associated with seismic events.

The lack of detailed attention to this specific topic in previous studies highlights the importance of our research proposal. By focusing on filling these knowledge gaps, we hope to contribute significantly to the global understanding of the effects of seismic events in specific areas. Our rigorous methodological approach and the use of advanced data analysis techniques will allow us to obtain a more complete and accurate view of this natural phenomenon. In this study, the objective is to apply data mining techniques to seismic events that occurred in Peru during the period from 1960 to 2021. The purpose of this research is to identify significant patterns of information and relationships that can contribute to the prediction of future seismic events. It is anticipated that the developed model will provide valuable information, allowing informed decision-making in the face of maximum-magnitude seismic events. This approach seeks to advise the population, guiding to avoid panic and adopt appropriate measures in the event of such events.

2. REVIEW LITERATURE

In this literature review section, the topic of data mining applied to earthquake prediction was explored. In this regard, it is important to analyze this issue from a scientific perspective, highlighting the valuable contributions of seismic experts, and at the same time, identifying limitations and opportunities to advance in the development of new solutions to address this challenge. The combination of data mining technologies and seismology emerges as a promising approach to prevent or take more cautious measures against seismic movements, as neglecting these factors could be tragic for people's lives.

In an approach focused on data analysis to identify geographical areas with a higher density of seismic events through prediction, spatial clustering algorithms based on density were implemented. The application of density-based spatial clustering of applications with noise (DBSCAN) proved effective in detecting groups in specific geographic coordinates. Over a specific period, the identified clusters allowed the construction of a spatial model of earthquake distribution, highlighting areas with higher density on the map. Overall, the study's results were successfully compared with the general seismic zoning map of the Republic of Kazakhstan, validating the reliability of density-based clustering [14]. In another context, within the disciplines of earth sciences and geology, machine learning tools have been employed to identify specific patterns associated with extreme terrains. In this study, a deep learning model was proposed whose main function is to extract spatiotemporal patterns from data to predict extreme earthquakes. This model utilizes spatial grids and synthetic deep-learning neural networks.

The obtained results revealed that the proposed model shows a strong correlation in predictions regarding the location and magnitude of earthquakes in Southern California [15]. It is crucial to address the challenges faced by urban infrastructure, especially their vulnerability to significant damage during high-intensity earthquakes. With this purpose, the research proposed a methodological model that evaluates specific characteristics of urban objects to determine their seismic resistance. K-means and hkmeans clustering algorithms were employed, using Euclidean distance as a proximity measure in prediction. The

elbow method facilitated the identification of prominent variables with a greater impact on seismic resistance. The results indicated that the obtained clustering coincided with expert estimates, demonstrating that the characteristics of urban objects can be effectively determined through data modeling using clustering algorithms [16], [17]. In another instance, the use of big data focused on earthquakes in Lombok, a structure located in three active layers in Indonesia designed to optimize seismic damage mitigation, is of great relevance. The conic multivariate adaptive regression splines (CMARS) algorithm, based on a chronic quadratic programming (CQP) framework, was employed for this analysis. This model highlighted independent variables such as epicenter distance (100%), magnitude (31.08%), and depth (3.53%) about peak ground acceleration (PGA) value [18].

In another study, the application of mathematical algorithms to seismic events aimed to identify significant patterns in high-magnitude seismic activities. This investigation proposes a comparison between predictive models focused on seismic magnitudes before and after applying clustering techniques. Three prediction models were used: decision trees, support vector machines (SVM), and k-nearest neighbors (KNN). The results revealed that the implementation of clustering improves the accuracy of predictive models. The maximum prediction accuracy and homogeneity of seismic sources are achieved by clustering earthquakes according to non-spatial attributes. Of the three models tested, the decision tree shows the highest accuracy [19].

In another investigation, devastating high-magnitude earthquakes have caused significant human losses in various communities. This research addressed the detection of areas prone to experiencing deadly seismic movements by employing the k-means clustering algorithm, implemented through the rapid miner tool. The collected data corresponds to 34 provinces in Indonesia, a region prone to such disasters. The results were obtained by classifying provinces into four groups based on earthquake magnitudes. Fourteen provinces with high magnitudes and a clustering center of 528.25 were identified, along with 14 provinces with medium magnitudes and a clustering center of 96.071, and finally, six provinces with low magnitudes and a clustering center of 57.604 [20]. Another focus in seismological research centers on the analysis of historical tremors. This study aims to perform data analysis using deep learning to detect tremors from seismic data over 50 years old, employing the ResNet architecture to extract patterns in seismic waveforms. The results indicate that the implemented proposal has great potential for tremor detection, which could be crucial for preventive actions and a deeper understanding of the relationships between tremors and earthquakes [21]. In a final investigation focused on earthquake prediction from historical seismic data in local areas, an approach implementing a deep learning model called electrical properties tomography (EPT) was employed. This model uses global feature extraction blocks (GFEB) to identify potential movement patterns in the crust and tectonic plates, using data collected from the global historical seismic catalog. Finally, the validation of the EPT model was carried out using five sets of historical data, resulting in a model accuracy of 90% [22].

After reviewing studies conducted by experts in the field, data mining algorithms were investigated for earthquake prediction, and the most relevant characteristics for analysis were conceptualized. However, unresolved gaps were identified, such as the lack of application of an agile methodology to build data mining models. On the other hand, the studies provide valuable insights but also present limitations that must be considered in future research, such as the dependence on historical data and the complexity of deep learning models, which may hinder their interpretation. Additionally, the implementation of complex algorithms could require computational resources and technical expertise, limiting their adoption in resource-limited environments. These identified gaps offer opportunities for future research and improvements in the application of analytical methods and algorithms in seismic prediction. This research aims to address these limitations and contribute to advancing solutions to the identified challenges.

3. METODOLOGY

3.1. Definition of knowledge discovery database methodology

The knowledge discovery database (KDD) methodology is fundamentally integrated into the process of creating predictive models based on data mining. This methodological approach follows an iterative and interactive structure that combines various traditional data analysis techniques with machine learning technologies. The underlying motivation for applying the KDD methodology is the identification of relevant information, enabling the extraction of crucial results for the researcher. This, in turn, facilitates strategic decision-making to address specific issues. In the development of the KDD methodology, various stages are implemented, each contributing progressively to the exploration, selection, and refinement of data, with the ultimate goal of extracting valuable and applicable knowledge [23], [24]. The stages applied in this process are described, as shown in Figure 1.

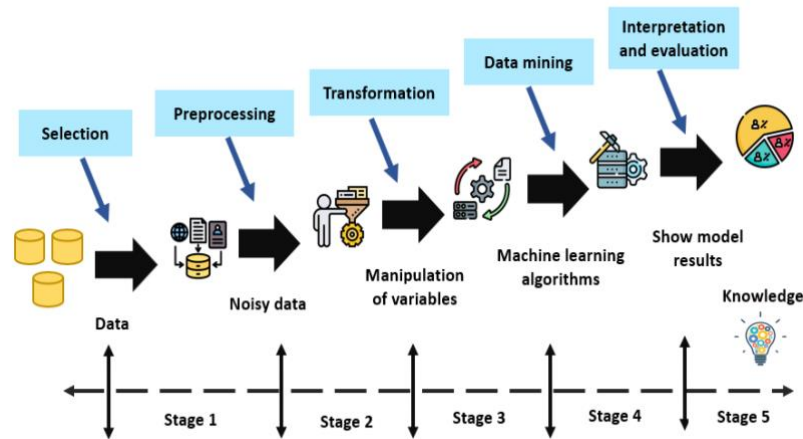


Figure 1. KDD methodology process

3.2. Phases of the knowledge discovery database process

In this section, the consolidation of the KDD methodology structure will be carried out, taking into account the various stages proposed by this approach. It is crucial to recognize that these phases form the foundation of the model for the development of solid and reliable results. The KDD methodology, focusing on extracting useful knowledge from large datasets, spans from the identification and selection of relevant data to the application of advanced analysis and modeling techniques. Each phase plays an essential role in the iterative and interactive process, contributing to the progressive construction of a robust analytical framework [25], [26]. The proper application of these stages not only supports the acquisition of valuable information but also lays the groundwork for informed decision-making and effective resolution of specific issues.

3.2.1. Data selection

In this phase, data collection has been carried out for use in subsequent stages, addressing the research topic. In total, around 10,000 records have been gathered, each composed of 8 fields that function as important features for the model analysis, as shown in Table 1. The data selection was based on information available on the portal of the Geophysical Institute of Peru (IGP), the entity responsible for monitoring seismic events such as earthquakes, volcanic eruptions, and torrential rains [27], [28]. In this context, the selected data provides extremely valuable information that will allow the detection of significant patterns in the chosen records. This approach ensures the quality and relevance of the data for the construction of a robust analytical model in the later stages of the KDD process.

Table 1. Collected database

Id	Date_utc	Time_utc	Latitude	Length	Depth	Magnitude	Cut_date
0	19600113	154034	-16.145	-72.144	60	7.500	20223006
1	19600115	93024	-15	-75	70	7	20223006
2	19600117	25758	-14.500	-74.5	150	6.400	20223006
3	19600123	33732	-12.500	-68.5	300	5.800	20223006
4	19600130	50724	-5.500	-77.5	100	5.700	20223006
5	19600208	190616	-8.500	-74.500	136	5.300	20223006
6	19600213	204006	-17.500	-70	150	5.900	20223006

3.2.2. Data preprocessing

The chosen data undergoes various cleaning and transformation techniques to address potential issues such as outliers, missing data, or redundancies. This process is carried out using key variables such as latitude, longitude, depth, and magnitude [29]. The identification and removal of noise in the data are performed, including records with empty spaces or strange characters that could pose complications when consolidating the results in the final phase of the model. Data cleaning is essential to ensure the integrity and quality of the information that will be used in later stages of the process [30], as mentioned in Figure 2.

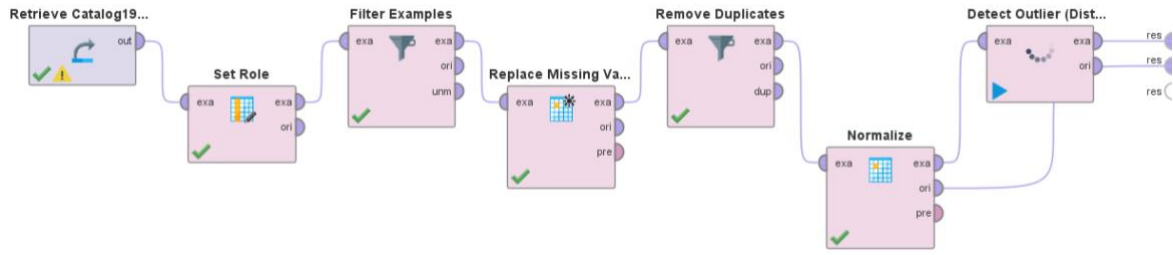


Figure 2. Data preprocessing stage

3.2.3. Data transformation

At this crucial point in the process, the data previously preprocessed in earlier stages undergoes a significant transformation to obtain a more accurate representation suitable for the analysis sought according to the objectives outlined in the research [31]. This step involves two fundamental aspects: dimensionality reduction and the creation of new features that will enrich the established model. Dimensionality reduction is essential for handling datasets with many variables [32]. By applying techniques such as principal component analysis (PCA) or feature selection methods, it is possible to preserve essential information while reducing the complexity of the dataset. This not only improves computational efficiency but also helps avoid overfitting to noise in the data, as mentioned in Figure 3.

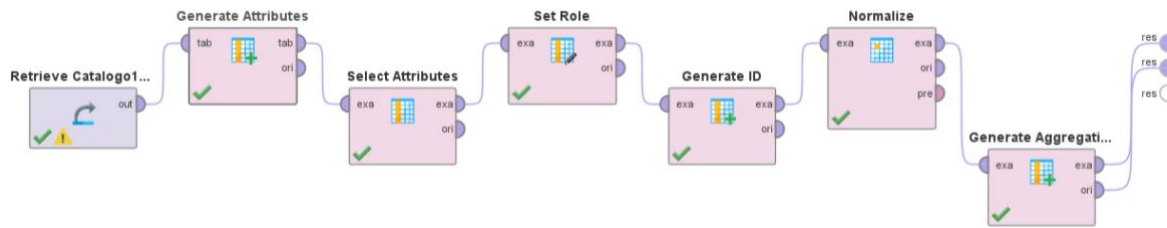


Figure 3. Data transformation stage

3.2.4. Data mining

In the data mining phase, we have employed specific mathematical algorithms within the RapidMiner Studio tool. In our strategy, we have opted for the use of the clustering algorithm, complemented with specific components in the model. Our proposal aims to integrate fundamental concepts of statistics and mathematics, reflected in the selected algorithm, with the clear objective of extracting relevant information that allows us to anticipate seismic events over time [28], [33].

– K-means algorithms

In the context of data mining and pattern analysis, a cluster refers to a group or set of elements that share certain similarities or common properties. The formation of clusters involves grouping data in such a way that elements within the same group are more similar to each other than to elements in other groups. Essentially, a cluster is a collection of objects or data points that exhibit affinities with each other based on specific criteria [34], [35]. The underlying idea is that elements within the same cluster share closer or similar characteristics, while those in different clusters exhibit more pronounced differences.

a) Euclidean distance

Euclidean distance is a fundamental measure of the distance between two points in Euclidean space. This concept, crucial in both geometry and data analysis, is expressed through a specific formula. In particular, the Euclidean distance between two points, p and q , in an n -dimensional space is shown in (1):

$$d(p, q) = \sqrt{\sum_{i=0}^n (p_i - q_i)^2} \quad (1)$$

b) Centroid of a cluster

In the context of clustering, centroids are representative points that summarize the information of a group or cluster of data. Each centroid is the central or average point of a set of points in a multidimensional

space. These central points are crucial in clustering algorithms such as K-Means, where the goal is to partition the group of data, called clusters, and assign each point to the cluster whose centroid is closest to that point, as mentioned in (2). The centroid of a dataset, denoted as C for a set of points, is calculated as the point whose i -th coordinate is the average of the corresponding coordinates of all points in the set. If C is a set of points, the centroid is calculated as in (2):

$$c_i = \frac{1}{|c|} \sum p \in c P_i \quad (2)$$

In RapidMiner studio, the previously mentioned criteria are described through specific operators that support the clustering algorithm, leveraging the unique features of the tool. In this context, the clustering algorithm is grounded in the concepts previously discussed, such as Euclidean distance and centroids. The configuration of the model is done using specific operators that allow for setting the necessary parameters to obtain coherent results. This process is visualized in Figure 4, representing the graphical interface of RapidMiner Studio during the construction and execution of the workflow for clustering analysis.

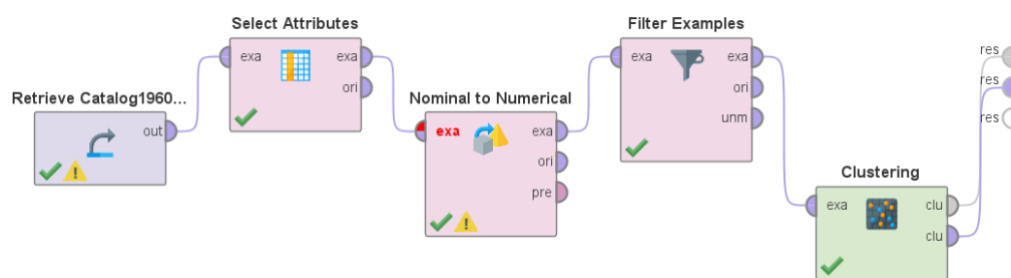


Figure 4. Data mining stage

4. RESULTS

4.1. Evaluation of result

In the application of the K-Means algorithm, the obtained result reveals certain patterns in the magnitude of the most devastating earthquakes at periodic intervals. These patterns are grouped into five categories, with data-sharing similarities based on Euclidean measurement between them. The clustering of this data using RapidMiner shows an output with characteristics that are akin to each other.

From the results of the clustering process, five data groups were formed, each with a different number of records. Specifically, the first cluster contains 45 records (15.5%), the second cluster has 56 records (19.2%), the third cluster presents 94 records (32.3%), the fourth cluster contains 54 records (18.5%), and the fifth cluster exhibits 42 records (14.5%). These figures represent the distribution of records in each cluster. The outcome of the clustering process confirms that the planning has been successful according to the research objectives. The number of records is shown in Table 2.

Table 2. Cluster model-number of items

Classification of clusters according to their items		
Clúster 0	45 items	15.5 %
Clúster 1	56 items	19.2 %
Clúster 2	94 items	32.3 %
Clúster 3	54 items	18.5 %
Clúster 4	42 items	14.5 %
Total, number of items	291 items	100%

Initially, it is necessary to assign a value based on the provided records, ordering them in descending order based on Euclidean measurement. These values are grouped at the center of the initial cluster, which will then undergo the K-Means process. Centroid calculations were performed through 10 iterations to determine the final grouping of a total set of 291 objects. From this process, Table 3 with the corresponding results was generated.

- Clúster 0. The initial group consists of 45 elements and stands out for having the highest mean, normalized in a range from 0 to 1 with a value of 0.230, surpassing the other sets. Additionally, this

cluster presents data with earlier dates, spanning the period between 1960 and 1976. Characteristically, these data are more dispersed, indicating a greater Euclidean distance between the elements that compose them. In summary, the first cluster is distinguished by having the highest average magnitude of earthquakes compared to the other clusters.

- Clúster 1. Regarding the second set with 56 elements, its mean has a normalized value of 0.156, considered an average value compared to the other groups. This cluster is notable for grouping cases with more recent dates, spanning the period between 1998 and 2021, and exhibits the highest magnitude peak. Consequently, the second cluster stands out for containing the maximum magnitude value recorded among all sets.
- Clúster 2. About the third group, which has 94 elements and is the largest of all, the magnitude mean, with a normalized value of 0.141, is the lowest compared to the other clusters, equaling only the fifth cluster. The date range extends from 1980 to 2001, placing this cluster in a more central position compared to the others. In summary, the third cluster presents a smaller magnitude of earthquakes compared to the other groups, highlighting, however, its amplitude and frequency.
- Clúster 3. As for the fourth set with 53 elements, its mean magnitude, with a normalized value of 0.155, is considered an average value compared to the other clusters. This cluster covers the most recent date range, from 2001 to 2021. Consequently, the fourth cluster is positioned as an average value compared to the other groups, as it does not present extremely high or low magnitudes in the mean.
- Clúster 4. The fifth set, the smallest with 42 elements, is characterized by having a mean normalized magnitude of 0.141, one of the lowest along with the third cluster. This cluster also aggregates older records, dating back from 1960 to 1982. In summary, this cluster is classified with the fewest occurrences and presents lower magnitudes compared to the other sets.

Table 3. Centroid content

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Latitude	0.613	0.489	0.531	0.402	0.576
Length	0.495	0.553	0.531	0.535	0.451
Depth	0.305	0.201	0.188	0.165	0.142
Magnitude	0.230	0.156	0.141	0.155	0.141
Id	552.911	13615.250	190.543	15440.130	1073.071
Utc_date	19657779.111	20087.495	19917133.202	20116631.537	19701892.381
Local_time	178369.289	19363.9286	124014.617	65346.537	539442.952

In the bar chart in Figure 5, the percentage distribution of records in each of the clusters generated through the K-means process is visually observed. Each bar represents a specific cluster, and the height of the bar reflects the percentage of records belonging to that cluster about the total analyzed data. This type of representation facilitates the interpretation of the prevalence of each group and provides a clear insight into the distribution of seismic events in the different categories identified by the clustering algorithm. On the other hand, Figure 6 shows a representation of the groupings based on clusters, relating two of the most important variables, such as the date of seismic events on the y-axis and magnitudes on the x-axis, and referencing each cluster with a specific color.

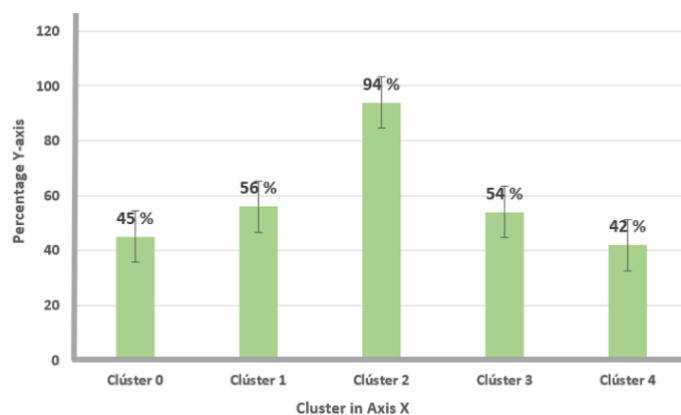


Figure 5. Bar graph representation

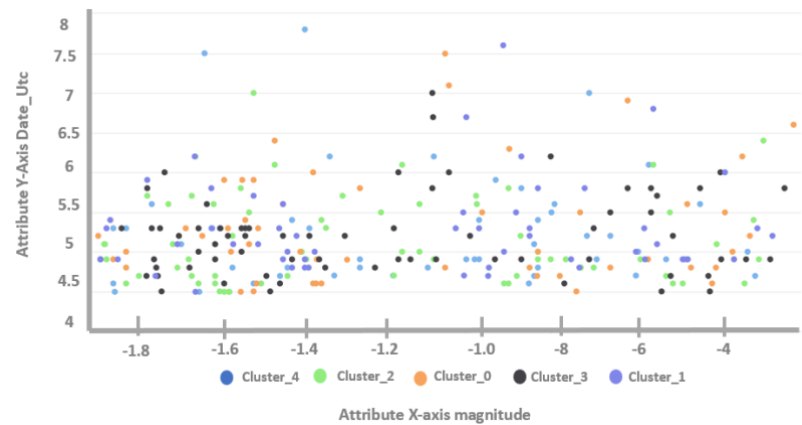


Figure 6. A graphic representation of the clusters

In Figure 7, a detailed representation of the correlation matrix obtained through the use of various tools in RapidMiner Studio can be appreciated. In this process, specific operators play a crucial role, such as the selection of important attributes and the use of the correlation matrix operator. In contrast, Figure 8 provides a more detailed visualization of the correlation matrix results, presenting the information in a structured way in a heat map. This map intuitively reflects the relationships and dependencies between specific fields of the selected database. The graphical representation facilitates the identification of patterns and trends, which can be invaluable for informed decision-making in data analysis and the exploration of relationships between variables.

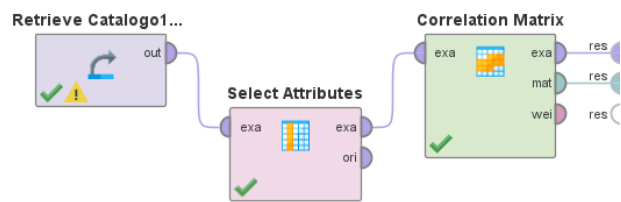


Figure 7. Correlation matrix operators

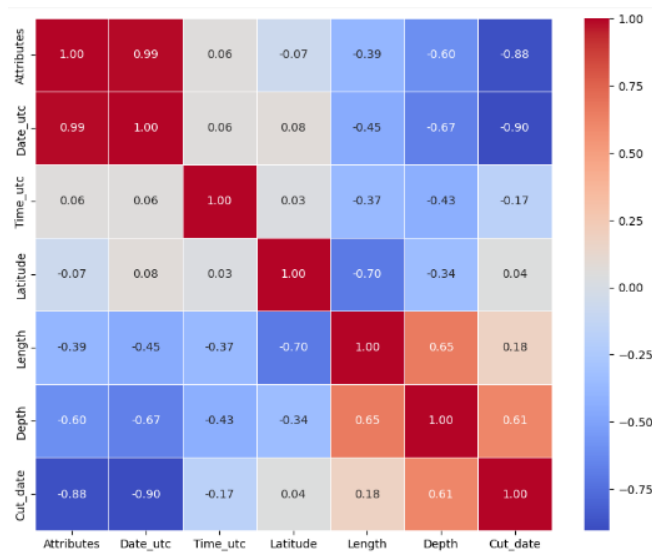


Figure 8. Correlation matrix scheme

4.2. Model comparison

In this section, a comparison of the algorithms used in data mining is carried out to evaluate their effectiveness in solving problems related to this discipline. To achieve this, a cartesian coordinate system has been chosen as the most distinctive for distinguishing the differences between the algorithms. Figure 9 presents a comparison between models such as naive Bayes, decision tree, and rule induction. With a score of 1.0, rule induction stands out as the most outstanding algorithm, while the others yield lower results. This analysis contributes to a better understanding of the various algorithms that the tool can construct.

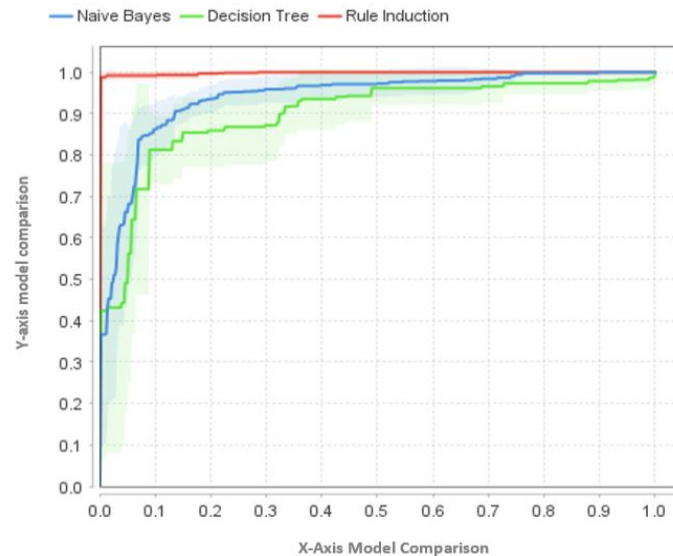


Figure 9. Comparison of algorithmic models

4.3. Comparison of methodologies

The choice of the KDD methodology over SEMMA and CRISP-DM was based on its suitability for the proposed data mining project. The KDD methodology stands out for its comprehensive approach, covering all stages of the data mining process, from data selection and preparation to model evaluation. Moreover, KDD has proven to be highly effective in identifying patterns and valuable insights in extensive datasets, a crucial aspect of our project. KDD surpassed SEMMA and CRISP-DM in versatility and the ability to more effectively and comprehensively address the specific challenges of our data mining project, as evidenced in Table 4.

Table 4. Comparison of methodologies

Attribute comparison	Methodology KDD	Methodology CRIPS-DM	Methodology SEMMA
Structure and sequence	Otras metodologías incluyen la selección, limpieza, transformación y extracción de datos, así como la evaluación y aplicación de conocimientos, pero su estructura no es tan rigurosa [36].	This methodology consists of six steps: business understanding, data understanding, data preparation, modeling, evaluation, and implementation [37].	The five steps of the SEMMA methodology are sampling, exploration, modification, modeling, and assessment [38].
Business orientation	It recognizes the importance of the company's business objectives and seeks to learn to gain a competitive advantage.	It understands the business objectives from the outset and ensures that the results are actionable and valuable for decision-making.	It performs information analysis considering the company's objectives and how the results are utilized.
Flexibility	Through a broader and less structured approach, it provides a general framework for knowledge discovery.	It is adaptable to a variety of contexts and projects and can be scaled for commercial use.	Although it follows a predetermined sequence of steps, it is adaptable to various projects.
Interaction	It requires iteration. However, it lacks an evident structure, similar to the Semma or Crisp-DM methodology.	It learns to use an iterative results review method. It adapts to projects in constant evolution.	It is a procedure that can be carried out in stages as needed. Adjustments can be made throughout the process.

5. DISCUSSION

In the collected research, the objective is to identify geographical areas with a higher density of seismic phenomena through prediction using machine learning algorithms, specifically the K-means algorithm. However, the results obtained in this research are not detailed in depth, and when comparing them with our model, we structure the K-means cluster model at each stage, obtaining valuable results for each particular cluster [14]. In the field of earth sciences and geology disciplines, other research also employs machine learning-based tools to identify patterns in the detection of extreme earthquakes. Unlike our research, the results of these studies only mention the correlation of predictions in relation to location and magnitude [15], which may limit the specificity of the findings. However, it is important to note that for the study by the authors [16], [17], which addresses high-density earthquakes affecting urban structures, there is a certain coincidence with our research, as both propose the prediction of extreme seismic movements using the K-means clustering algorithm. In another study, the CMARS algorithm was applied using quadratic programming for analysis [18]. This research aligns with the selected variables in our model, as the results related to the epicenter, magnitude, and depth are specified, which were also objectives in our research.

Research by Hashemi and Karimi [19], which emphasizes the identification of significant patterns for seismic activities, the result was that the applied models had greater accuracy in the clustering theme. This coincidence is relevant since our results are also accurate according to the established objectives. Another coincidence is found in the research proposed in [20], which addresses areas most prone to experiencing deadly seismic movements using the K-means algorithm in the RapidMiner studio tool. Although the objectives differ, both studies apply K-means, highlighting the stark differences in results due to different approaches. In the field of seismology, the aim was to analyze earthquakes with over 50 years of history using deep learning [21]. These results align with our proposed model, as both cases seek to predict earthquakes considering features such as magnitude and depth, considering the existing relationships between tremors and earthquakes. Finally, Purnomo [22] proposed historical seismic data for local areas was considered using a deep learning technique called EPT. Although the results do not match the implementation of our model, they focused on verifying the model's validation compared to our results, which focus on prediction patterns regarding tremors and earthquakes.

6. CONCLUSION

In conclusion, our research, which aimed to apply data mining techniques to seismic events, successfully identified the sought-after patterns, thus fulfilling the established purpose at the beginning of the study. The obtained results provide valuable information that will be crucial for taking measures in the face of possible seismic events in the future. The implementation of the KDD methodology was fundamental to structuring the appropriate process for the consolidation of the model and the achievement of the desired results. This methodology allowed for managing changes and understanding the data structure while maintaining its reliability and quality. Consequently, we conclude that the KDD methodology facilitated obtaining reassuring results and laid solid foundations for future research on the subject. The results of our implementation demonstrated the utility of the K-means algorithms used in the RapidMiner studio tool for seismic data analysis. The perspective provided by the results of each cluster, considering concepts such as Euclidean distance and centroids, offers a deeper understanding of the magnitude of seismic events. Our research suggests that the performance of our model will benefit many geographical areas with a history of experiencing seismic events. The discovery generated confidence in certain institutions dedicated to the detection of catastrophic events, as the impact of the results obtained is considered an important solution that allows for effective measures to prevent panic among the population. Regarding the limitations of the research, it is concluded that no significant obstacles were encountered during the consolidation of the model. Our model effectively aligns with the established objective, overcoming certain difficulties that arise in the different stages of the KDD methodology. However, it is important to consider additional research that allows for a deeper understanding of the proposed model to confirm its validation, which provides accurate predictions. Our model effectively aligns with the established objective, overcoming certain difficulties that arose in the different stages of the KDD methodology. Therefore, this study not only contributes to the field of seismology but also opens doors to new possibilities in the technological and scientific realms, including institutions dedicated to making predictions in this scientific field. Likewise, it is essential to highlight that our research is not limited to the academic sphere but also has practical and social implications. The results obtained can be used by government authorities, disaster management organizations, and local communities to enhance preparedness and response to seismic events. This practical application demonstrates the positive impact that scientific research can have on society and underscores the importance of investing in science and technology. In conclusion, we maintain the vision that our work lays the groundwork for future research in seismology and other disciplines such as economics, health, or education. Additionally, we highlight the

possibility of applying other technologies, such as developing applications based on this model, creating mobile applications that utilize these concepts, or even exploring big data to generate valuable information in massive databases. The observations also refer to our findings providing compelling evidence that these seismic events are associated with the collision of moving tectonic plates, releasing energy during a sudden reorganization of materials in the earth's crust.




REFERENCES

- [1] J. Fayaz, R. Astroza, C. Angione, and M. Medalla, "Data-driven analysis of crustal and subduction seismic environments using interpretation of deep learning-based generalized ground motion models," *Expert Systems with Applications*, vol. 238, 2024, doi: 10.1016/j.eswa.2023.121731.
- [2] J. Yang, L. Z. Li, X. Wang, and S. X. Hu, "Experimental study on the seismic performance of concrete shear walls partially replaced by MRPC," *Case Studies in Construction Materials*, vol. 20, 2024, doi: 10.1016/j.cscm.2023.e02695.
- [3] E. Pirot, C. Hibert, and A. Mangeney, "Enhanced glacial earthquake catalogues with supervised machine learning for more comprehensive analysis," *Geophysical Journal International*, vol. 236, no. 2, pp. 849–871, 2024, doi: 10.1093/gji/ggad402.
- [4] D. D'Angela, G. Magliulo, C. D. Salvatore, and M. Zito, "Seismic assessment and qualification of acceleration-sensitive nonstructural elements through shake table testing: reliability of testing protocols and reliability-targeted safety factors," *Engineering Structures*, vol. 301, 2024, doi: 10.1016/j.engstruct.2023.117271.
- [5] M. Chinello, E. Bersan, M. Fondriest, T. Tesei, R. Gomila, and G. Di Toro, "Seismic cycle in bituminous dolostones (Monte camicia thrust zone, central apennines, Italy)," *Geochemistry, Geophysics, Geosystems*, vol. 24, no. 12, 2023, doi: 10.1029/2023GC011063.
- [6] L. Lu, J. Zhang, G. Zhang, H. Peng, B. Liu, and H. Hao, "The influence of box-strengthened panel zone on steel frame seismic performance," *Buildings*, vol. 13, no. 12, 2023, doi: 10.3390/buildings13123042.
- [7] C. Drooff and J. T. Freymueller, "New insights into the active tectonics of the Northern Canadian cordillera from an enhanced earthquake catalog," *Journal of Geophysical Research: Solid Earth*, vol. 128, no. 12, 2023, doi: 10.1029/2023JB026793.
- [8] N. Barrera, D. M. Ruiz, J. C. Reyes, Y. A. Alvarado, and D. C. Beltrán, "Seismic performance of a 1:4 scale two-story rammed earth model reinforced with steel plates tested on a bi-axial shaking table," *Buildings*, vol. 13, no. 12, 2023, doi: 10.3390/buildings13122950.
- [9] J. Hu, T. S. Pham, and H. Tkalčić, "Seismic moment tensor inversion with theory errors from 2-D Earth structure: implications for the 2009–2017 DPRK nuclear blasts," *Geophysical Journal International*, vol. 235, no. 3, pp. 2035–2054, 2023, doi: 10.1093/gji/ggad348.
- [10] S. Takemura *et al.*, "A review of shallow slow earthquakes along the Nankai trough," *Earth, Planets and Space*, vol. 75, no. 1, 2023, doi: 10.1186/s40623-023-01920-6.
- [11] H. B. Yang *et al.*, "Probabilistic seismic hazard assessments for Myanmar and its metropolitan areas," *Geoscience Letters*, vol. 10, no. 1, 2023, doi: 10.1186/s40562-023-00301-x.
- [12] S. M. Mousavi and G. C. Beroza, "Machine learning in earthquake seismology," *Annual Review of Earth and Planetary Sciences*, vol. 51, pp. 105–129, 2023, doi: 10.1146/annurev-earth-071822-100323.
- [13] H. Kikuchi, "Data mining method in seismology by applying cellular automaton equivalence of ground vibration fluctuations recorded near the epicenter of the 2011 Mw 9 East Japan earthquake," *Earth Science Informatics*, vol. 16, no. 3, pp. 2615–2633, 2023, doi: 10.1007/s12145-023-01054-z.
- [14] M. Karmenova *et al.*, "An approach for clustering of seismic events using unsupervised machine learning," *Acta Polytechnica Hungarica*, vol. 19, no. 5, pp. 7–22, 2022, doi: 10.12700/APH.19.5.2022.5.1.
- [15] B. Feng and G. C. Fox, "Spatiotemporal pattern mining for nowcasting extreme earthquakes in Southern California," *Proceedings - IEEE 17th International Conference on eScience, eScience 2021*, pp. 99–107, 2021, doi: 10.1109/eScience51609.2021.00020.
- [16] R. Kaneko, H. Nagao, S. I. Ito, K. Obara, and H. Tsuruoka, "Convolutional neural network to detect deep low-frequency tremors from seismic waveform images," *Lecture Notes in Computer Science*, vol. 12705, pp. 31–43, 2021, doi: 10.1007/978-3-030-75015-2_4.
- [17] W. Wojcik, M. Karmenova, S. Smailova, A. Tlebalidnova, and A. Belbeubaev, "Development of data-mining technique for seismic vulnerability assessment," *International Journal of Electronics and Telecommunications*, vol. 67, no. 2, pp. 261–266, 2021, doi: 10.24425/ijet.2021.135974.
- [18] D. Priyanto, M. Zarlis, H. Mawengkang, and S. Efendi, "Analysis of seismic hazard prediction using non parametric conic multivariate adaptive regression splines (C-Mars) methods," *Journal of Physics: Conference Series*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012057.
- [19] M. Hashemi and H. A. Karimi, "Seismic source modeling by clustering earthquakes and predicting earthquake magnitudes," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICTST*, vol. 166, pp. 468–478, 2016, doi: 10.1007/978-3-319-33681-7_39.
- [20] C. Hu, Z. Cui, J. Lin, M. Huang, and S. Liu, "Application research of clustering algorithm in earthquake disaster prediction system of sanhe," *Journal of Physics: Conference Series*, vol. 1237, no. 2, 2019, doi: 10.1088/1742-6596/1237/2/022013.
- [21] S. Karimi, M. Heydari, J. Mirzaei, O. Karami, B. Heung, and A. Mosavi, "Assessment of post-fire phenological changes using MODIS-derived vegetative indices in the semiarid oak forests," *Forests*, vol. 14, no. 3, 2023, doi: 10.3390/f14030590.
- [22] M. R. A. Purnomo, "A Bayesian reasoning for earthquake prediction based on IoT system," *Journal of Physics: Conference Series*, vol. 1471, no. 1, 2020, doi: 10.1088/1742-6596/1471/1/012022.
- [23] M. Donauer, P. Peças, and A. Azevedo, "Nonconformity root causes analysis through a pattern identification approach," *Lecture Notes in Mechanical Engineering*, vol. 7, pp. 851–863, 2013, doi: 10.1007/978-3-319-00557-7_70.
- [24] A. Dekhtyar and J. H. Hayes, "Automating requirements traceability: Two decades of learning from KDD," *2018 1st International Workshop on Learning from other Disciplines for Requirements Engineering*, pp. 12–15, 2018, doi: 10.1109/D4RE.2018.00009.
- [25] L. Aguagallo, F. S. -Fierro, J. G. -Santillán, M. P. -Yépez, P. L. -López, and I. G. -Santillán, "Analysis of student performance applying data mining techniques in a virtual learning environment," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 11, pp. 175–195, 2023, doi: 10.3991/ijet.v18i11.37309.
- [26] L. Al-Alawi, J. Al Shaqsi, A. Tarhini, and A. S. Al-Busaidi, "Using machine learning to predict factors affecting academic performance: the case of college students on academic probation," *Education and Information Technologies*, vol. 28, no. 10, pp. 12407–12432, 2023, doi: 10.1007/s10639-023-11700-0.




- [27] C. E. A. Guajardo, X. A. L. -Cortes, and S. H. Alvarez, "Deep learning algorithm applied to bacteria recognition," *2022 IEEE International Conference on Automation/25th Congress of the Chilean Association of Automatic Control: For the Development of Sustainable Agricultural Systems, ICA-ACCA 2022*, 2022, pp. 1-6, doi: 10.1109/ICA-ACCA56767.2022.10005945.
- [28] B. K. F. Aquino, Á. E. C. Baquijano, and C. Ovalle, "Algorithm based on deep learning to improve the logistics management of a company that distributes reading material," *Frontiers in Artificial Intelligence and Applications*, vol. 363, pp. 281–286, 2022, doi: 10.3233/FAIA220544.
- [29] A. H. Azizan *et al.*, "A machine learning approach for improving the performance of network intrusion detection systems," *Annals of Emerging Technologies in Computing*, vol. 5, no. 5, pp. 201–208, 2021, doi: 10.33166/AETiC.2021.05.025.
- [30] P. T. -Carrión, C. V. -Tene, Y. Jiménez, and D. Castillo, "Machine learning model for the prediction of emotions in a mobile application," *Communications in Computer and Information Science*, vol. 1388 CCIS, pp. 263–271, 2021, doi: 10.1007/978-3-030-71503-8_20.
- [31] N. Akhtar, M. R. Talib, and N. Kanwal, "Data mining techniques to construct a model: Cardiac diseases," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, pp. 532–536, 2018, doi: 10.14569/IJACSA.2018.090173.
- [32] S. Kurniawan, W. Gata, D. A. Puspitawati, I. K. S. Parthama, H. Setiawan, and S. Hartini, "Text mining pre-processing using gata framework and rapidminer for Indonesian sentiment analysis," *IOP Conference Series: Materials Science and Engineering*, vol. 835, no. 1, 2020, doi: 10.1088/1757-899X/835/1/012057.
- [33] D. R. -Albarrán, L. G. Torrealba, S. A. Aguirre, and O. Alexander, "Support system to predict student dropout in universities," *Smart Innovation, Systems and Technologies*, vol. 318, pp. 3–12, 2023, doi: 10.1007/978-981-19-6347-6_1.
- [34] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Computer Science*, vol. 6, pp. 1–43, 2020, doi: 10.7717/PEERJ-CS.267.
- [35] Y. M. Cheang and T. C. Cheah, "Predicting movie box-office success and the main determinants of movie box office sales in Malaysia using machine learning approach," *ACM International Conference Proceeding Series*, pp. 57–62, 2021, doi: 10.1145/3457784.3457793.
- [36] H. J. G. Palacios, R. A. J. Toledo, G. A. H. Pantoja, and Á. A. M. Navarro, "A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change," *Advances in Science, Technology and Engineering Systems*, vol. 2, no. 3, pp. 598–604, 2017, doi: 10.25046/aj020376.
- [37] J. Bokrantz, M. Subramaniyan, and A. Skoogh, "Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM," *Production Planning and Control*, pp. 1–21, 2023, doi: 10.1080/09537287.2023.2234882.
- [38] S. L. -Torres *et al.*, "IoT monitoring of water consumption for irrigation systems using SEMMA methodology," *Lecture Notes in Computer Science*, vol. 11886 LNCS, pp. 222–234, 2020, doi: 10.1007/978-3-030-44689-5_20.

BIOGRAPHIES OF AUTHORS



Laberiano Andrade-Arenas    is Doctor in Systems and Computer Engineering and Master in Systems Engineering. He graduated with a Master's Degree in University Teaching. He also graduated with a Master's degree in accreditation and evaluation of educational quality. He is a systems engineer, scrum fundamentals certified, a research professor with publications in Scopus-indexed journals. He can be contacted at email: landrade@uch.edu.pe.



Cesar Yactayo-Arias    obtained a bachelor's degree in administration from Universidad Inca Garcilazo de la Vega and a master's degree in education from Universidad Nacional de Educación Enrique Guzmán y Valle, he is a doctoral candidate in administration at Universidad Nacional Federico Villarreal. Since 2016 he has been teaching administration and mathematics subjects at the Universidad de Ciencias y Humanidades and since 2021 at the Universidad Continental. Currently, he also works as an administrator of educational services at the higher level, he is the author and co-author of several refereed articles in journals, and his research focuses on TIC applications to education, as well as management using computer science and the internet. He can be contacted at email: yactayocesar@gmail.com.