Data Mining 1 - Project

Midterm Draft

Bertocci Alberto [477945]
Jornea Ion [637765]
Macchia Alessandro [636679]
Stortoni Cristian [639184]

University of Pisa
A.Y. 2021/2022

Contents

1 Introduction

Mining activity has always been linked to seismic hazard, and mining risk assessment represents one of the main issues in the working field, both for the safety of workers and the proficiency of the ore extraction. Despite the evolution of more and more advanced seismic and seismoacoustic monitoring systems, seismic hazard remains one of the hardest detectable and predictable of the natural hazards. Accuracy of so far created methods is however far from perfection, since the complexity of the seismic processes and due to the big disproportion between the number of low-energy seismic events and the number of high-energy phenomena. In fact, the latter causes statistical techniques to be insufficient to predict seismic hazard. The task of seismic prediction can be defined in different ways, but the main aim of all seismic hazard assessment methods is to predict an increased seismic activity which can cause a rockburst. In this project, the problem of high energy seismic bumps forecasting is being discussed, with the use of the seismic bumps data set[1], which contains records from two longwalls located in a Polish coal mine.

In order to perform the analysis, a variety of programming languages (Python, R) and software (Jupyter Notebook, RStudio, Tableau, Orange) has been used.

2 Data Understanding and Preparation

2.1 Data semantics

The dataset has 2584 instances and a total of 19 attributes, which define a summary statement about seismic activity in the rock mass within one shift (8 hours). There are 4 categorical features, and 15 numerical features (9 discrete and 5 continuous), described in Table 1. The data is recorded by geophones, devices that convert ground oscillations into recorded signals, being able to collect both the energy magnitude of a pulse and the number of pulses in a given time interval. In particular, the reference for the measures is usually the most active geophone monitoring the longwall, called GMax. The hazard assessment is then made upon different methods (seismic, seismoacoustic) and it is expressed in a scale ranging from a to d, where a is lack of hazard, b low hazard, c high hazard and d danger state. All variables refer to observations made in the previous 8-hour shift, except the "class" attribute, which refers to the current shift. Furthermore, a high-energy seismic bump is defined as one having an energy level of at least 10^4 J.

In the domain column, we indicated first the observed domain and in parentheses the theoretical domain of each variable.

Table 1: Variables description

Name	Type	Domain	Description
seismic	Categorical (Ordinal)	$\{a, b\}$ $(\{a, b, c, d\})$	Result of the hazard assessment using the seismic method.
seismoacoustic	Categorical (Ordinal)	$\{a, b, c\}$ $(\{a, b, c, d\})$	Result of the hazard assessment using the seismoacoustic method.
shift	Categorical (Binary)	$\{W, N\}$	Type of shift (W = working, N = preparation).
genergy	Numerical (Continuous)	$x \in [100, 2595650]$ $(\mathbb{R}^+)$	Seismic energy recorded by the Gmax.

[1] `https://archive.ics.uci.edu/ml/datasets/seismic-bumps`

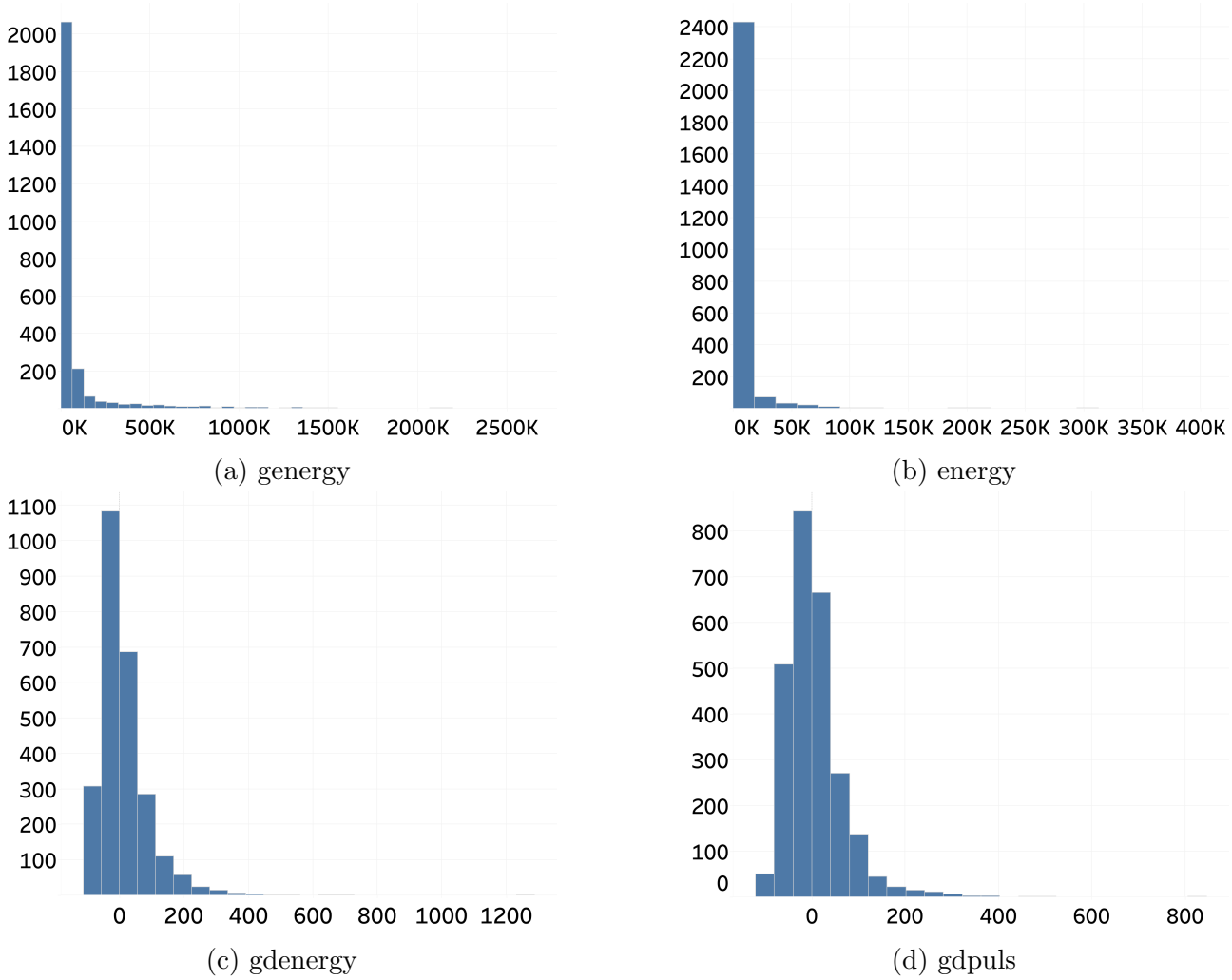
gpuls	Numerical (Discrete)	$x \in [2, 4518]$ ($\mathbb{N}$)	Number of pulses recorded by the GMax.
gdenergy	Numerical (Continuous)	$x \in [-96, 1245]$ ($\mathbb{R}$)	Deviation of "genergy" from average of previous 8 shifts.
gdpuls	Numerical (Continuous)	$x \in [-96, 838]$ ($\mathbb{R}$)	Deviation of "gpuls" from average of previous 8 shifts.
hazard	Categorical (Ordinal)	$\{a, b, c\}$ ({a, b, c, d})	Result of the hazard assessment using the seismoacoustic method, relying only on GMax.
nbumps	Numerical (Discrete)	$n \in [0, 9]$ ($\mathbb{N}$)	Number of seismic bumps recorded.
7 variables of the type "nbumps**x**"	Numerical (Discrete)	$n \in [0, 8]$ ($\mathbb{N}$)	Number of seismic bumps recorded in the previous shift with energy in the range $[10^x, 10^{x+1})$
energy	Numerical (Continuous)	$x \in [0, 402000]$ ($\mathbb{R}$)	Total energy of the seismic bumps recorded.
maxenergy	Numerical (Continuous)	$x \in [0, 400000]$ ($\mathbb{R}$)	Maximum energy of the seismic bumps recorded.
class	Categorical (Binary)	$\{0, 1\}$	Decision attribute (0 = no high-energy seismic bump recorded in the next shift, 1 = high-energy seismic bump recorded in the next shift).

2.2 Distribution of the variables

In this specific section we will analyse the most significant variables in terms of distributions. The latter will be visualised thanks to eight different plots, regarding specifically: four continuous variables and four categorical and discrete ones. As it is possible to see in Figure 4, the first four attributes present a very positively skewed distribution and an evident long tail to the right, most likely due to the large range of the variables. An important factor that can be noted, is the high number of 0 values that occur in all the distributions, specifically in the distributions of "genergy" and "energy".

In Figure 2 are represented some of the categorical and discrete variables. Regarding these plots there are few aspects which are important to be mentioned. First of all, in Figure 2a is clear how the number of class "0" (absence of high energy seismic bumps) is disproportionate with respect to the number of class "1" (presence of high energy seismic bumps). Almost the same situation is present in Figure 2c, where the number of a is far greater than the number of b and c. A similar situation occur in Figure 2d, where it is clear that the numbers of the shift "W" and the numbers of the shift "N" are unbalanced. In Figure 2b it can be seen how once again the frequency of "0 nbumps" is greater respect all the others.

Further discussion regarding the distribution of the variables is provided in the section dedicated to the analysis of variables correlation.



(a) genergy

(b) energy

(c) gdenergy

(d) gdpuls

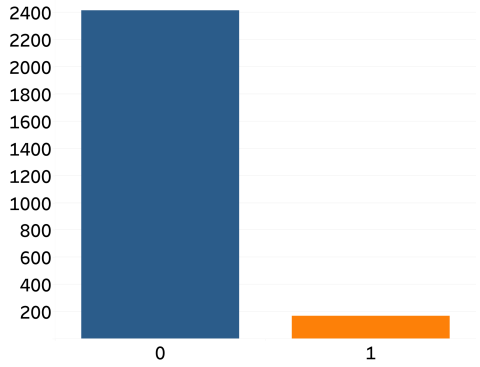
Figure 1: Frequency distribution of selected numerical variables

2.3 Assessing data quality

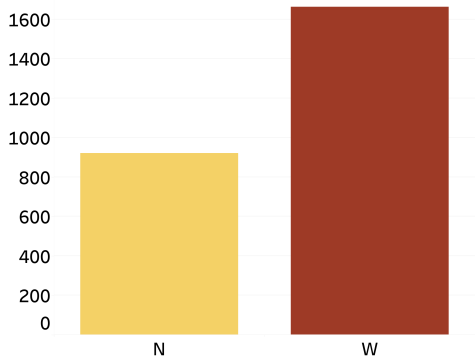
None of the variables in the data set presents any missing values, so no adjustments in this sense are necessary. It is to be remarked again that the data set is very unbalanced with respect to the attribute "class": there are only 170 records of class 1, against 2414 records of class 0.

With concern to timeliness, the data set has been published in 2013 so it could be slightly out of date. Findings of the current work may not be applicable to current situations.

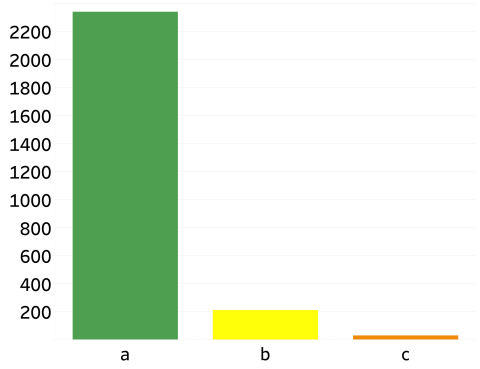
The attributes "nbumps6", "nbumps7" and "nbumps89" all contain exclusively zero values. This is most likely due to a lack of seismic bumps of sufficient intensity to fall in those categories. We do not and cannot know however whether possible malfunctions of the detection systems may have had an impact, giving origin to some hidden missing values. The only contradiction we can effectively check is whether for records with "nbumps" equal to 1, "energy" and "max-energy" are equal. No such contradictions have been found, thus increasing our confidence that there should be no significant occult quality issues. Furthermore, there are no contradictions between the variables "energy" and "maxenergy" overall. The continuous variables present a huge number of outliers: by applying a Z-score standardisation and looking at their boxplots (Figure 3a), we can observe many points above the upper whisker, confirming the right skewness of the data. Due to the very large number of points, we cannot really consider them as outliers, but rather as an actual part of the data. This justifies our choice not to eliminate any of them, as they represent useful information for our analysis. Clearly, they need to be



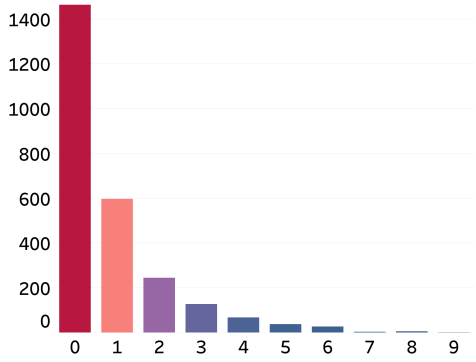
(a) class



(b) shift



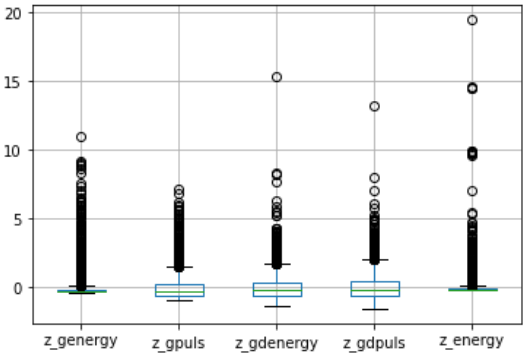
(c) hazard



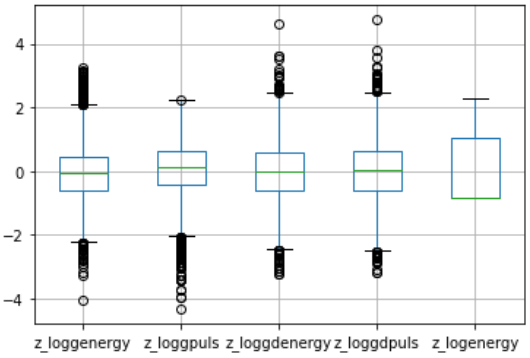
(d) nbumps

Figure 2: Frequency distribution of selected categorical and discrete variables

transformed in some way in order to be managed more easily, which is the focus of the next section.



(a) Boxplots of selected standardised variables



(b) Boxplots of selected standardised transformed variables

Figure 3: Boxplots

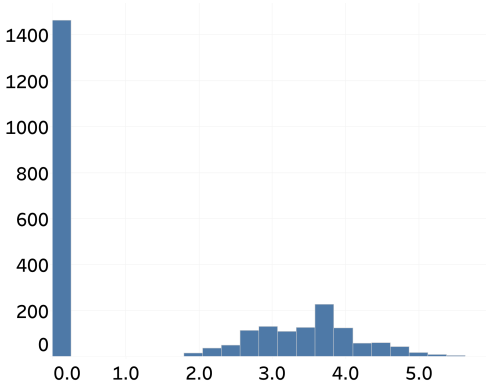
2.4 Variable transformation

For the highly right skewed variables, we decided to apply a log transformation to the data to stabilise the variance and possibly reduce the number of outliers. For variables containing only positive variables, we directly applied a log transformation (with base 10); for variables presenting null values, we added the constant 1 to each record and then applied a log transfor-

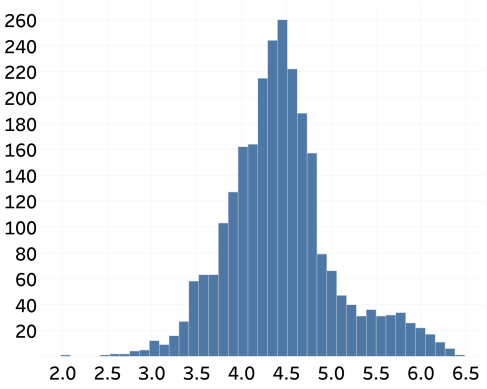
mation. For variables presenting negative values, we had to add an arbitrary constant, which was chosen in such a way to make the distribution of the data as close to a normal distribution as possible (this was done by minimising the test statistic of the Jarque-Bera test on the data). The procedure led to the following transformations (where X indicates the original variable, and Y the transformed variable):

- "genergy": $X \to Y = log(X)$
- "gpuls": $X \to Y = log(X)$
- "gdenergy": $X \to Y = log(X + 115)$
- "gdpuls": $X \to Y = log(X + 125)$
- "energy": $X \to Y = log(X + 1)$
- "maxenergy": $X \to Y = log(X + 1)$

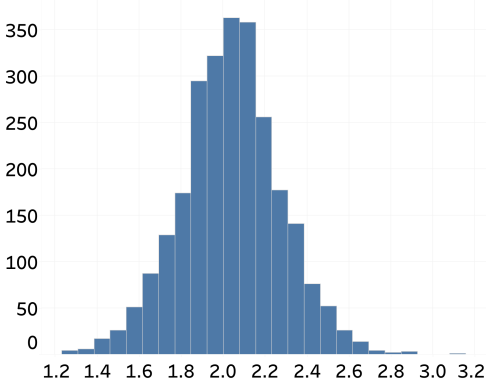
Subsequently, the log-transformed variables have been normalised to see whether the problem of outliers has diminished. As evident from Figure 3b, the data are now concentrated mostly within 4 standard deviations from the mean, confirming a big improvement from the original situation, although a huge number of outliers is still present. The distributions (Figure 4) are now also much less skewed and resemble more a normal distribution, although the Jarque-Bera test rejects the hypothesis of normality in all cases. It is to be remarked how the variable energy (Figure 4a) presents a very unbalanced structure: this is caused by the fact that many of the records present an energy level equal to 0, but among those with positive energy, we can recognise a distribution similar to the other log-transformed variables.



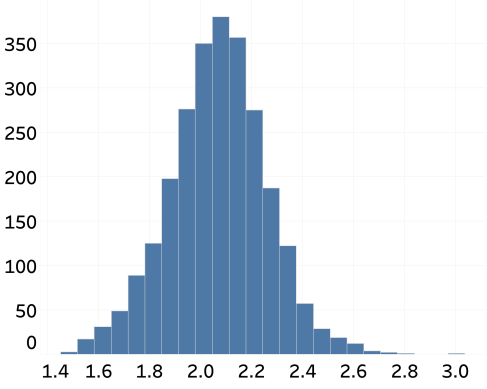
(a) log energy



(b) log genergy



(c) log gdenergy



(d) log gdpuls

Figure 4: Frequency distribution of selected log-transformed continuous variables

2.5 Pairwise correlation and elimination of variables

With regard to our quantitative attributes we decided to investigate the correlation between the variables using the correlation matrix (Figure 5), which can show how all variables are related to the others. To assess the strength of the correlation between characters, it was decided to use the correlation coefficient of Bravais-Pearson, which can take values between -1 and 1. As far as the six variables treated in the previous section are concerned, we used their log-transformed version.

The most important correlation is found between "maxenergy" and "energy" (correlation coefficient equal to 1). Note that both measure the same identical event, that is the energy of recorded seismic shocks: the difference is that the former records the maximum energy, while the latter records the total energy. For this reason we decided to eliminate the variable "maxenergy" from future analysis as all its information is already contained in "energy".

We can note, therefore, that there is a strong positive correlation, that is a condition of concordance, between different groups of variables: the first is the relationship between "gdenergy" and "gdpuls" in which we can see a high value of 0.82. This makes us understand that probably the behaviour of these attributes could be very similar. However, we do not consider the value so strong in order to delete one of them. We are also aware of the presence of several medium-high correlation coefficients (say greater than 0.6). None of them is strong enough to make us immediately want to delete some variables, but we need to take this information into account when going into future tasks, where some adjustments may be needed.

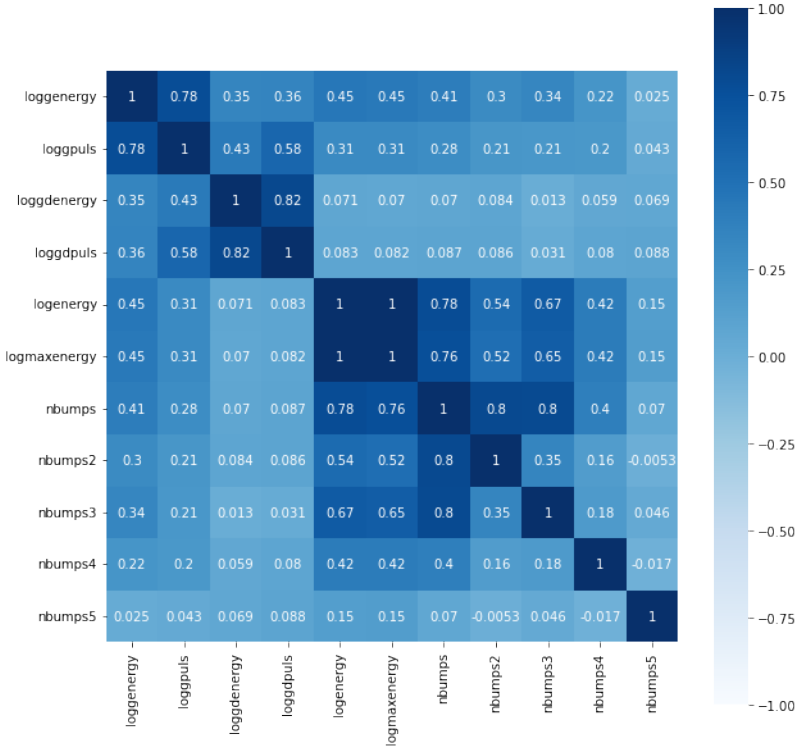


Figure 5: Correlation matrix between numerical attributes

Looking at the categorical data, we can investigate about their independence. In particular, it could be interesting to look at how the attribute "class" is distributed with respect to other categorical variables. This is summarised by the contingency tables in Figure 6. Using a Chi square test for independence, we can confirm that "class" is independent with respect to the "seismoacoustic" and "hazard" variables, whereas it is not independent from "seismic" and, even more strongly, from "shift". As evident, the presence of class 1 increases (in relative

seismic \ class	0	1	Total
a	1599.0	83.0	**1682.0**
b	815.0	87.0	**902.0**
Total	**2414.0**	**170.0**	**2584.0**

(a) against seismic

seismoacoustic \ class	0	1	Total
a	1479.0	101.0	**1580.0**
b	890.0	66.0	**956.0**
c	45.0	3.0	**48.0**
Total	**2414.0**	**170.0**	**2584.0**

(b) against seismoacoustic

shift \ class	0	1	Total
N	904.0	17.0	**921.0**
W	1510.0	153.0	**1663.0**
Total	**2414.0**	**170.0**	**2584.0**

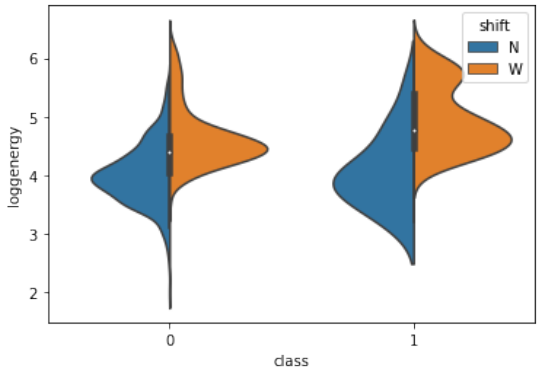
(c) against shift

hazard \ class	0	1	Total
a	2186.0	156.0	**2342.0**
b	198.0	14.0	**212.0**
c	30.0	0.0	**30.0**
Total	**2414.0**	**170.0**	**2584.0**

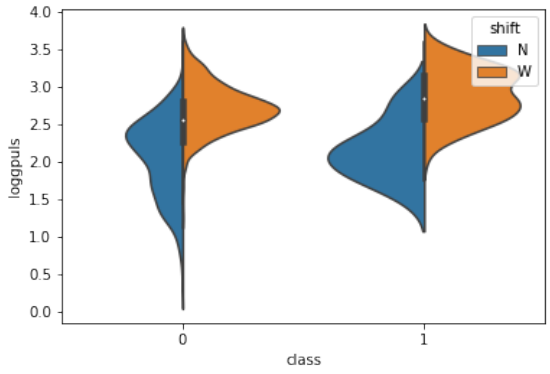
(d) against hazard

Figure 6: Contingency tables of variable class against other categorical variables

terms) as the seismic assessment goes from a to b (from 5.19% to 10.67%). Similarly, class 1 records are more likely to appear during a working shift (W), rather than a preparation one (N), with a percentage of 10.13% in the first case and 1.88% in the second. Interesting insights may come from analysing the distribution of continuous variables with respect to class and shift.



(a) Distribution of genergy for class and shift



(b) Distribution of gpuls for class and shift

Figure 7: Violin plots

Considering the elements identified in this analysis, let's see how our target class behaves, that is the risk class during the phases of work and preparation. Through a couple of violin plots (Figure 7) we can see that the distribution of "genergy" and "gpuls" for the shift of work (W) contains a more pronounced peak of values compared to the preparation shift, which contains a tail of lower values. We can remark how for both variables we notice an interesting behaviour when recording energy in the risk class, where there is a double bell in the distribution: this means a second higher seismic behaviour even for weaker values.

3 Clustering Analysis

Before entering the clustering task, a bit of data pre-processing is necessary to reach better results. In particular, not all variables will be considered for this analysis: all categorical attributes have been dropped, as most clustering algorithms require a measure of distance, which is not easily computable for this type of attributes. Furthermore, the fields of the type "nbumps**x**" have also been ignored, in order to reduce the dimensionality of the data. The information contained in those variables is somehow embedded in the variables "nbumps" (which is the sum of the previous variables) and "energy" (which gives an idea of the energy level of the seimic bumps), so that no significant information loss occurred.

Thus, the variables considered are the following: "genergy", "gpuls", "gdenergy", "gdpuls", "energy", and "nbumps" (all numerical, with the first 5 being log-transformed as described earlier). Given the massive presence of outliers, the data have been standardised using the robust scaler, which is less susceptible to the presence of outliers. This also solves the issue of the variables having different scales, so that they were not directly comparable. As distance metric, we have adopted the Euclidean distance, which was deemed as the most appropriate for the kind of variables at hand.

3.1 Analysis by hierarchical clustering

We decided to start approaching the clustering task by performing an agglomerative hierarchical clustering. This could allow us to have a hint about the number of clusters to use in later techniques (like K-means). We explored four linkage methods (single, complete, average, Ward), whose results can be seen in the four dendrograms in Figure 8. We take advantage of the default colour threshold of the scipy Python library to have a first idea of the possible most appropriate number of clusters.

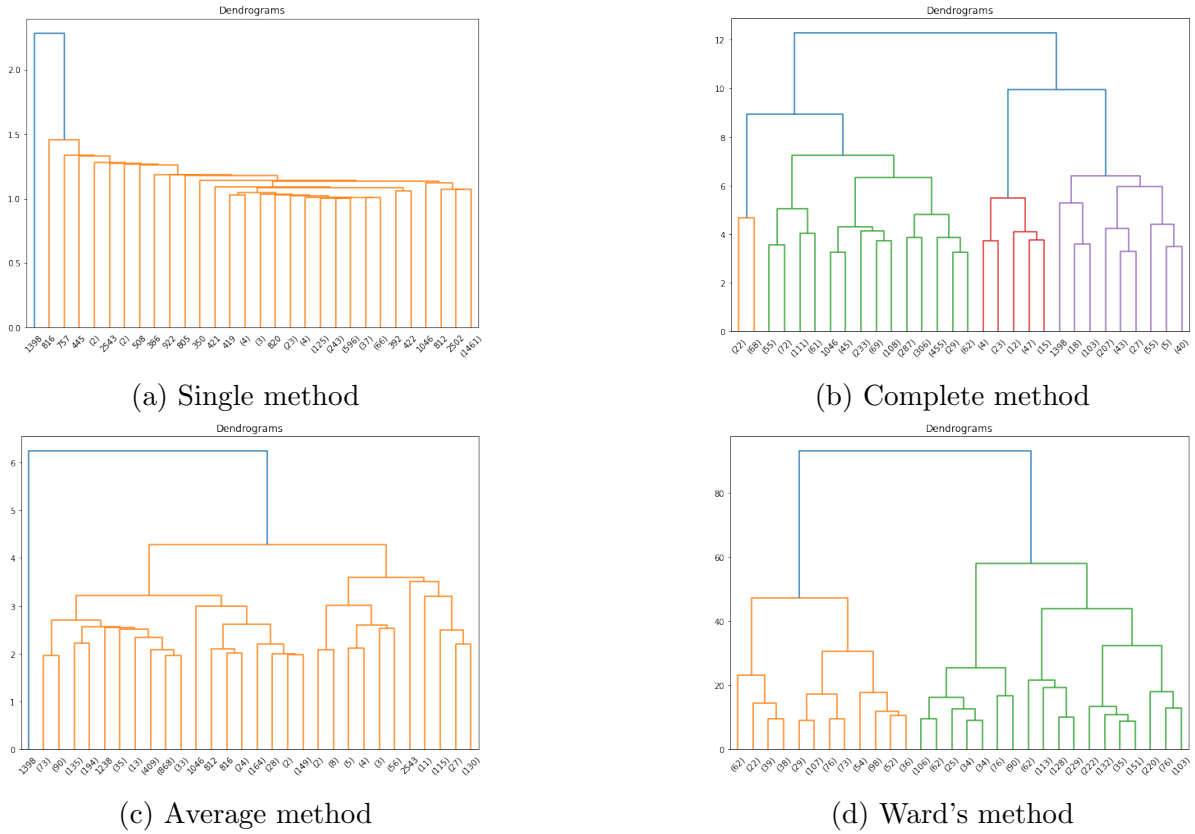
As evident from the images, the single method (Figure 8a) and average method (Figure 8c) return very unbalanced results, with 2 clusters, one of which is a singleton. This suggests that they are not suitable to treat the data at hand. On the other side, the complete (Figure 8b) and Ward (Figure 8d) methods return more balanced and elegant results, suggesting 4 and 2 clusters respectively. Adopting the suggested number of clusters, we calculated the Silhouette score for the two clusterings and we obtained a score of 0.158 for the complete method with K=4, and a score of 0.350 for K=2. In order to double check, we calculated the Silhouette score for K ranging from 2 to 18 (Figure 9), and we found that Ward's method consistently performed better than the complete method. In both cases the Silhouette score was highest for a value of K=2. Looking at the dendrograms for both methods, it is not unreasonable to think that 2 may be an appropriate number of clusters. As summarised in Table 2, the complete method generates clusters of much more variable size, hinting that perhaps some of them may be merged (which graphically would translate into cutting the dendrogram in Figure 8b at a higher level, say 10, than what proposed by default).

Method	**Number of clusters**	**Composition of clusters**	**Silhouette score**
Ward	2	{1898, 686}	0.350
Complete	4	{1508, 753, 222, 101}	0.158

Table 2: Results of hierarchical clustering using the complete and Ward's method

So, considering the Silhouette score, the composition of clusters, and the fact that there is no clear theoretical justification, we lean towards adopting the results of the Ward's method, because this method works out which observations to group based on reducing the sum of

squared distances of each observation from the average observation in a cluster. Thus, we deemed K=2 as a reasonable choice for the number of clusters.



(a) Single method

(b) Complete method

(c) Average method

(d) Ward's method

Figure 8: Dendrograms of various clustering methods

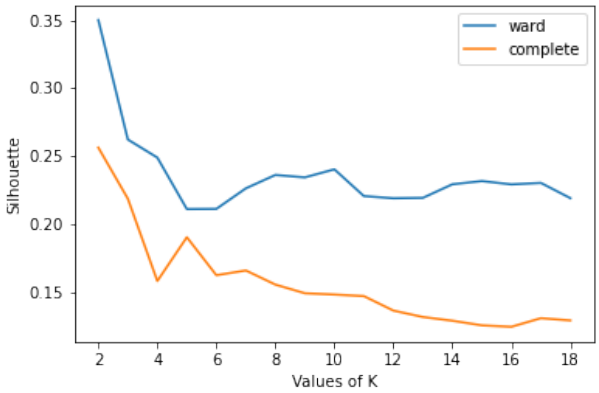
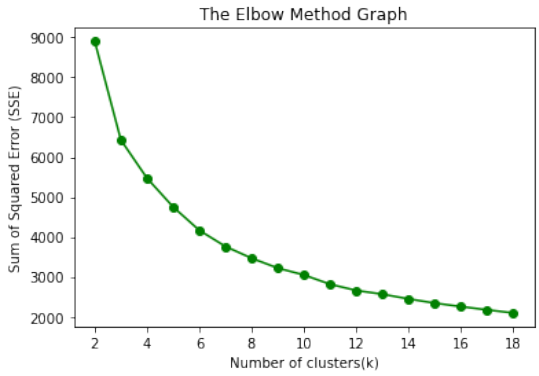


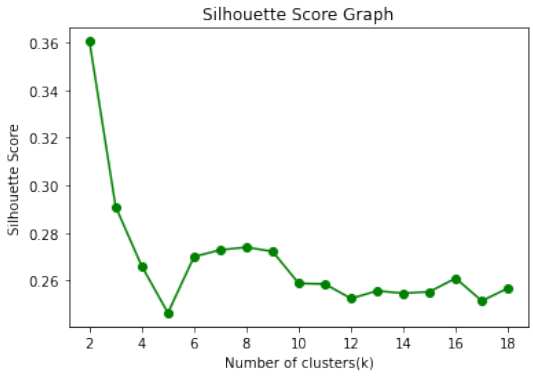
Figure 9: Silhouette score for different values of K

3.2 Analysis by K-means

To choose the best value of K we relied on the use of the elbow method and on the use of Silhouette score. As it can been seen in Figure 10 both plots have been delineated considering a hypothetical value of K ranging from 2 to 18. For each value of K, we ran the K-means algorithm choosing the centroids at random, for a total of 10 initiations each with a maximum number of iterations equal to 300. Given the results of the two graphs we decided to confirm the previous intuition of K=2 as in the hierarchical algorithm. In fact, even if setting two as number of clusters meant to obtain the highest SSE, at the same time it also meant to get the highest silhouette score. Nevertheless, we decided to consider also other options, but we realised that even with a slight increase of the number of clusters, it would have meant a non-significant change in terms of clearness, but rather a noticeable decrease of the Silhouette score.



(a) SSE (y axis) for different values of K (x axis)



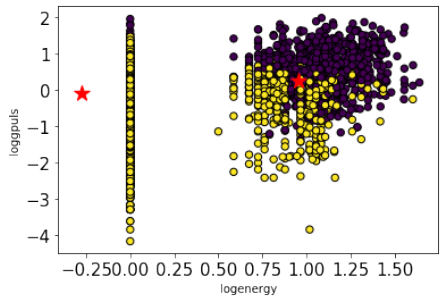
(b) Silhouette score (y axis) for different values of K (x axis)

Figure 10: Choice of K

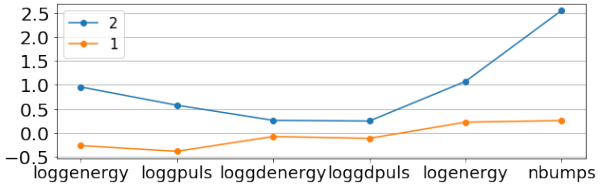
Figure 11a shows the graphical result of K-means clustering where K=2. As it can be seen, in this case two variables were represented, specifically "energy" and "gpuls". It is important to notice how almost all the points seem to be divided in two cluster by a fictional downward-sloping oblique line. Furthermore, to better understand the choice of K=2, it is possible to have a close monitoring to the Figure 11b. Here it is clear that the trend of the two clusters is quite well-separated for the first two variables, very similar for the third and fourth variables but completely different for the last two. This evolution can be explained since in Figure 11a cluster 1 (in yellow) is characterised by a lot of 0 values for almost all the variables, while cluster 2 (in blue), despite having a similar trend to the previous one, it presents a marked increase of the values in terms of "energy" and "nbumps". When performing the K-means clustering with higher values of K, we always obtained much less well-defined clusters, with some of them having almost completely overlapping centroids as the value of K increased. Our strategy was to maximise clearness and interpretability, so we sticked with the choice of 2 clusters.

Table 3 shows the distribution of the data among the two clusters. The highly unbalanced occurrence shows a more frequent lower risk scenario in cluster 1 (only 3.19%), while in cluster 2 we see a concentration of class 1 cases (i.e., the presence of a hazardous seismic bump), at 18.79%. The clusters are of quite different sizes, with the first containing more than twice as much points as the second one, but this reflects the nature of the data, which is more concentrated around lower values for many of the variables.

To better visualise the clusters we then plotted the scaled and log-transformed attributes for each of them, with the Kernel Density Estimation (KDE). In Figure 12c the "energy" variable shows distinctively the difference between cluster 1 and cluster 2, with the profile of cluster 2



(a) Clusters and their respective centroids



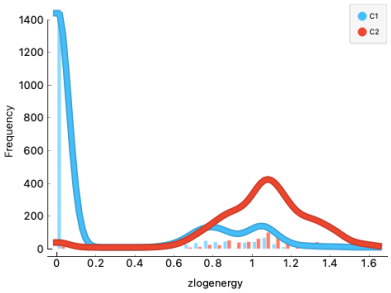
(b) Parallel coordinates plot

Figure 11: Cluster centroids

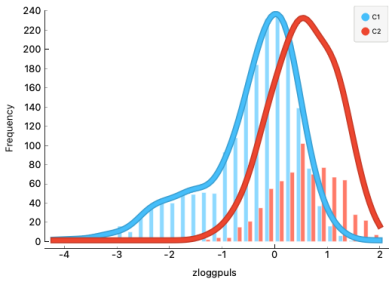
Table 3: Distribution of the attribute class across the two clusters

	Class 0	**Class 1**	**Total**
Cluster 1	1,818	58	1,876
Cluster 2	596	112	708
Total	2,414	170	2,584

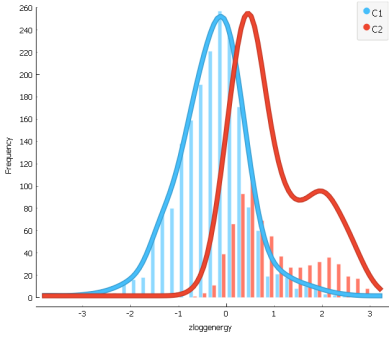
peaking far from the lower values of cluster 1. In Figure 12b the "gpuls" variable shows a similar behaviour between the two curves, with an abundance of lower-valued cases in cluster 1 and a shifted peak position on the right side, which confirms again the nature of cluster 2. Figure 12c then depicts again shifted peaks, but it is the presence of a second local peak on high values of "genergy" that this time characterises cluster 2. We can summarise this information by stating that we have a cluster characterised by low values of energy and pulsations, which results in less risky events (in terms of class 1 records), whereas the second cluster is characterised by higher values of energy and pulsations, with a consequently higher risk in terms of hazardous seismic bumps. The variations in energy ("gdenergy") and pulsations ("gdpuls") do not seem to play a significant role, although cluster 2 presents slightly larger for these variables.



(a) Distribution of scaled log energy



(b) Distribution of scaled log gpuls



(c) Distribution of scaled log genergy

Figure 12: Distribution of selected variables inside the clusters, with KDE superimposed