UNIVERSITÀ DI PISA

Data Mining Project

Analysis of the "seismic-bumps" dataset

Data Mining
AY 2021/2022

Cosimo Faeti 636812
Riccardo Galarducci 637763
Lorenzo Pieri 578814
Rocco Tiesi 636678

1 Data Understanding and Preparation

Mining activity is connected with the occurrence of dangers, a special case of such a threat is the seismic hazard which frequently occurs in many underground mines.
Seismic hazard is the hardest detectable and predictable of natural hazards and in this respect it is comparable to an earthquake.
With this regard seismic-bumps Dataset tries to describe the problem of high energy seismic bumps forecasting in a coal mine; its data come from detections carried out by two longwalls in a Polish coal mine.
The main goal of this report is to predict in which conditions it is more likely that seismic activity could cause a rockburst using machine learning methods.

1.1 Data Semantics

The dataset is composed of 2584 instances (rows), each of them represents an eight-hour work shift in the coal mine, and from 19 attributes (columns) that summarize the features detected for every shift.
Since the work shifts are in chronological order, the dataset can be also considered as a time series.
We proceed to describe the attributes:

Categorical Attribute

Shift There are two possible types of work shift in the coal mine:

- W - Coal-getting Shift
- N - Preparation Shift

Seismic is the result of shift seismic hazard assessment in the mine working obtained by the seismic method.
Indeed the work shift can be marked by the following letters, corresponding to an increasing scale of predicted risk: a - lack of hazard, b - low hazard, c - high hazard, d - danger state.
The seismic method takes into consideration tremors caused by seismic waves that spread in the rock mass and can be primary waves (P-waves), that propagate in a longitudinal way, or secondary waves (S-waves), which instead are transversal.
The seismic hazard assessment is based on the intensity of seismic tremors occurrence and can be a qualitative assessment for low seismic activity or a quantitative assessment for high seismic activity.
The level of seismic activity is determined on the basis of the number andenergy of tremors recorded in the observed longwall in a certain timeinterval (a shift).

SeismoAcoustic is the result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method.
As for the seismic assessment, it can assume the following values: a - lack of hazard, b - low hazard, c - high hazard, d - danger state.
The seismoacoustic method analyzes acoustic waves (compressions and rarefactions) which propagate in the air and move on the longitudinal axis. The essence of this method is recording and analysis of seismoacoustic emissions occurring within a given longwall.
Seismoacoustic emissions are described by their intensity that is the result of the activity (number of registered events) and by their energy, therefore the following factors are crucial for seismic hazard assessment:

- registration of the seismoacoustic emission,
- the number of pulses recorded by geophones, which is then converted by an appropriate formula for so-called conventional seismic energy.

The main criteria for assessment are changes in registered seismoacoustic activity and energy. Moreover, deviations of values calculated during successive time intervals also influence defining one of the four states a, b, c, d of seismic hazard.

Hazard is a variation of the seismoacoustic assessment method, in this case the recording of the number of pulses comes only from Gmax, which is the most active geophone in the longwall.
The seismic hazard assessment states according to

the hazard method are the same as for seismic method and seismoacoustic method: (a - lack of hazard, b - low hazard, c - high hazard, d - danger state).

Numerical Attribute

GEnergy (continuous): seismic energy recorded during the shift by GMax;

GPuls (continuous): total number of pulses recorded by GMax during the shift;

GdEnergy (continuous): deviation of seismic energy recorded by GMax from the average seismic energy recorded during eight previous shifts;

GdPuls (continuous): deviation of a number of pulses recorded by GMax from the average number of pulses recorded during eight previous shifts;

Energy (continuous): sum of the energy released by all the seismic bumps occurred during the shift recorded by geophones;

MaxEnergy (continuous): maximum value of energy released by a seismic bumps occurred during the shift recorded by geophones;

NBumps (discrete): total number of seismic bumps recorded during the shift;

NBumps n (discrete): number of seismic bumps. For each range $[10^n, 10^n+1]$ recorded during the shift ($n = 2, 3, 4, 5, 6, 7, 8$–9)

Target Variable

Class (dummy variable), it assumes value:
1 - if high energy seismic bumps (higher than 10^4 J) occurred in the next shift (hazardous state);
0 - if no high energy seismic bumps (higher than 10^4 J) occurred in the next shift (non-hazardous state).

1.2 Distribution of the variables and statistics & Variables Transormations

Class. The distribution of the target variable is highly unbalanced. As we can see in Figure 1, high energy seismic bumps ($> 10^4$ J) occurred only in 170 out of 2578 shifts (6.6%). It's in this big disproportion between the number of low-energy seismic events and the number of high-energy phenomena that it is so difficult to fairly predict seismic hazards.

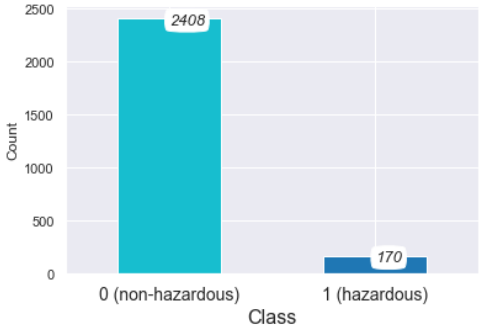

Figure 1: Countplot of *class*

In the dataset are present four categorical attributes:

Shift. We can see that there is a higher proportion of "W: Coal getting" (1662 out of 2578) than "N: Preparation" (916 out of 2578).

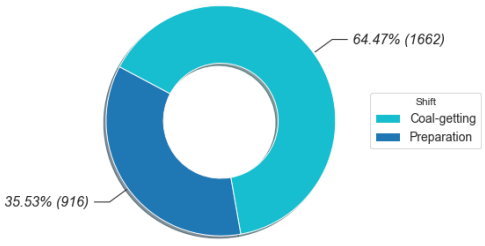

Figure 2: Donut chart of *shift*

We analysed the relationship with the target variable

through a crosstab (Figure 3). There is a higher presence of records with "Class = 1" in the "coal getting" shift. It means that a hazardous state is more likely (7.4% more) to happen in this kind of shift.

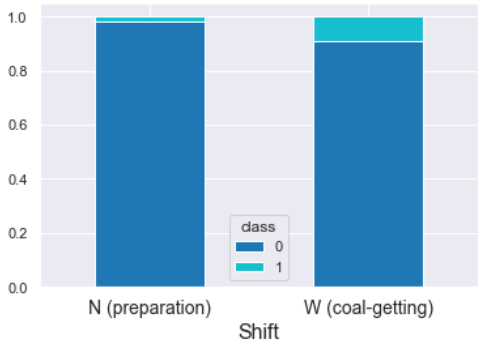

Figure 3: Class distribution in *shift*

Seismic. The variable seismic can assume values "a"," b"," c" or "d" that denote increasing states of seismic risk in the mine, however in the 2587 instances of the dataset the variable assumes only values "a" and "b". Therefore, according to the seismic method, none of the shifts is evaluated as high risk or danger state. In the figure below we can see that there is a higher number of records marked by "a - lack of hazard" (1676) than the number of records assessed as "b - low hazard" (902).

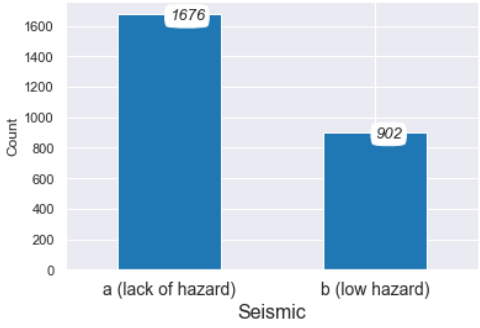

Figure 4: Countplot of *seismic*

The relationship between seismic and class can be explored with a crosstab as we have done before.

Figure 5 shows that, as can be expected, the percentage of shifts of "class = 1" is higher (9.7%) within the shift assessed as low hazard (seismic = b) rather than in the shift with lack of hazard (seismic = a), where the percentage reaches the 4.9% of the total.

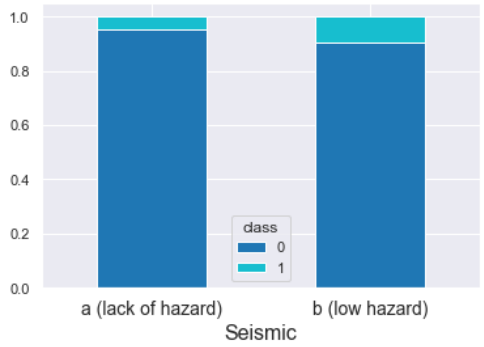

Figure 5: Class distribution in *seismic*

We have also analyzed the seismic variable with respect to the shift type. The Figure 6 points out that for the subset of records with seismic = "a", the ratio between the two types of shift is almost balanced, instead for the records where seismic = "b" more than the 80% of the subset is composed of coal-getting shifts.

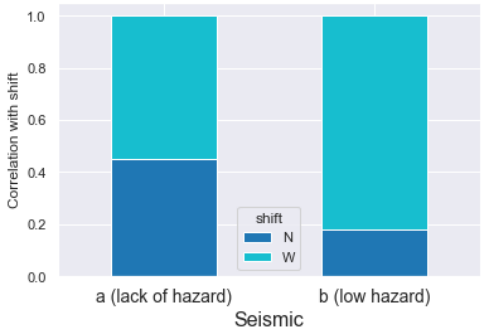

Figure 6: Shift distribution in *seismic*

Seismoacoustic. The distribution of the variable is represented by the bar plot below (Figure 7), where on the x-axis we have the state of seismic hazard pre-

dicted by the seismoacoustic method and on the y-axis the relative frequencies: 61% of the records were predicted as "lack of hazard", 37% as "low hazard" and only 2% as "high hazard". Note that no "danger state" was predicted.

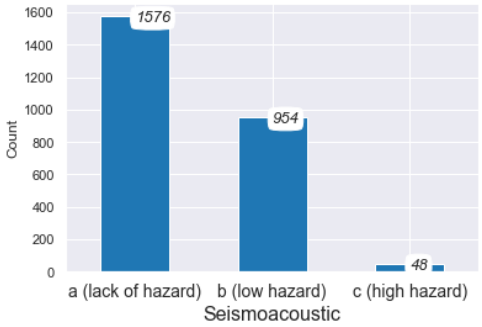

Figure 7: Countplot of *seismoacoustic*

If we analyse the relationship with the target variable we can state that the probability of a high energy seismic bump occurring in the next shift is the same for all the three categories ($\sim 6-7\%$).

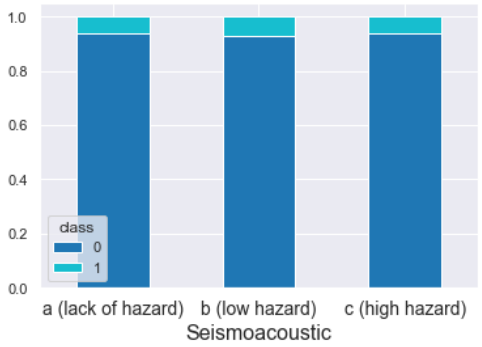

Figure 8: Class distribution in *seismoacoustic*

Analysing it with the variable "Shift" we obtained important results. As shown in the figure below, the fraction of W is larger than N, but the interesting part here is that the frequency of W (coal getting) increases with the increasing state of seismic hazard assessment. In fact, in "high hazard" the ratio between the two types of shifts increases by almost 20% compared to "lack of hazard". In other words, the seismoacoustic method suggests that there could be a strong relationship between the type of shift and the result of seismic hazard assessment according to the seismoacoustic method.

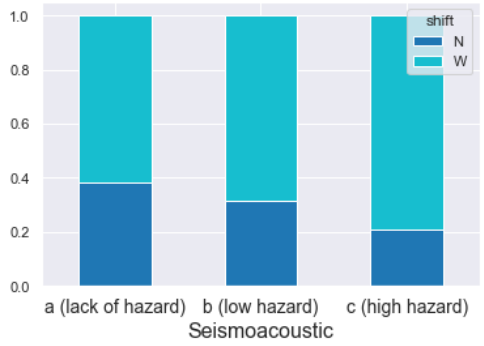

Figure 9: Shift distribution in *seismoacoustic*

Hazard. We can see from the bar plot (Figure 10) a very high distribution of the value "a" (2236) which indicates a state of "no hazard", while there are few instances with values "b" (212) and "c" (30), and none with value "d".

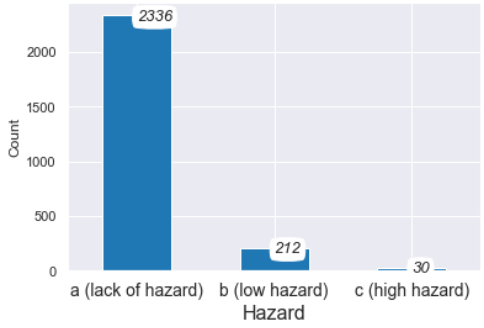

Figure 10: Countplot of *hazard*

Using a crosstab (Figure 11) to analyze the relationship between the attribute and the target variable class, we can see that the probability of high energy seismic bumps occurred in the next shift is the same for no hazard and low hazard, about $6,67\%$, while for the high hazard risk (c) no "hazardous state" occurs in the next shift.

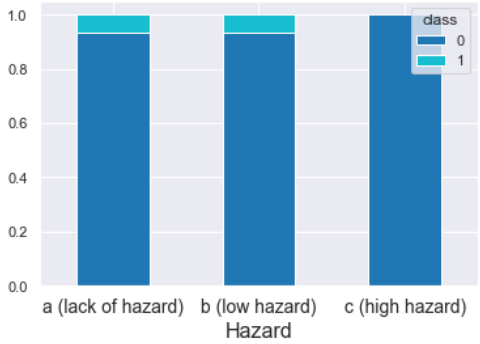

Figure 11: Class distribution in *hazard*

We think this result is unusual when considering the difference between no risk and low risk because we should expect some seismic bumps activities in proximity of shifts with high risk measurements. However, the number of high risk instances is almost non-existent considering the size of the entire dataset so it may not be a reliable sample. As for the relationship between Hazard and Shift, we can notice a similar behavior that occurs with the seismoacoustic attribute: a higher frequency of coal-getting shift is observed during low hazard states than during lack of hazard instances.

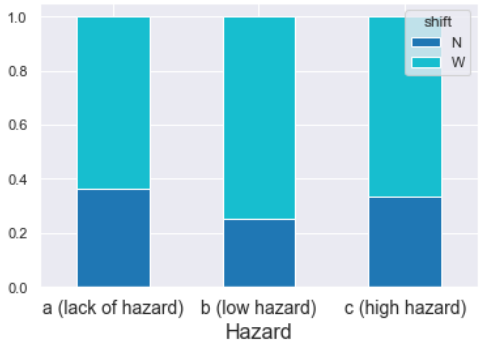

Figure 12: Shift distribution in *hazard*

This might suggest a link between shifts and seismic hazard but the scenario changes when considering also high risk instances where the frequency of coal-getting shifts returns to the level recorded during lack of hazard shifts. However, also in this case, the reason for this inconsistent trend could be identified in the small number of high risk instances.

In addition, we transformed the categorical attributes so that we could make better use of the classification models.
In particular, for "Seismic", "Seismoacoustic" and "Hazard" we decided to use the ordinal encoding because the values ("a","b","c" and "d") represent an order of hazardous. However we used the label encoding for "Shift" since there is no order between the records.

In seismic-bumps dataset are also present 14 numerical attributes: 6 continuous and 8 discrete.
Note that for each continuous attribute we used Sturges's rule to calculate the number of bins for the histograms (12 bins).

Nbumps. The domain of the Nbumps ranges between 0 and 9. As shown in the figure below we can note that it's a right-skewed distribution with the majority of records having "zero" as values. Since it's a counting variable, it only tells us how many seismic bumps were recorded during a shift, but it does not tell us anything about the energy that was released.

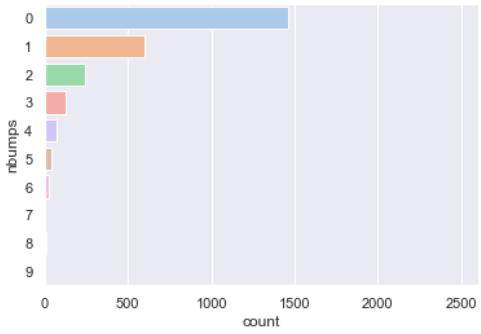

Figure 13: Countplot of *nbumps*

Nbumps n. Also in this case attributes have right-skewed distributions and Nbumps 6,7,8-9 assume all

values zero, this means that no seismic bumps of energy greater than 10^6 occurred in the dataset.

Genergy Figure 14 exhibits the distribution of Genergy. Similarly as what we have seen in some of the previous attributes, there is a right long tail. Hence, the majority of the distribution is located on the left side. The variable ranges from a minimum of 0 to a maximum of 2,595,650.

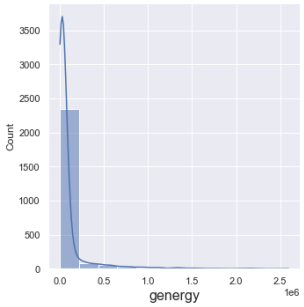

Figure 14: Distribution of *genergy*

With the aim of making Genergy's distribution approximately a Gaussian's distribution, we have applied the logarithmic transformation. In the Figure 15, it is possible to see that the peak of the distribution is in correspondence of log10-Genergy equals to 10. Thus, the majority of the distribution has a seismic bump (although in the previous figure we can see that the maximum distribution is located near zero).

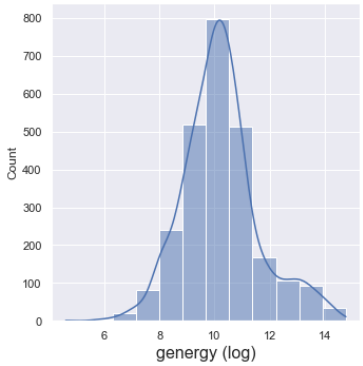

Figure 15: Transformed distribution of *genergy*

We analysed the transformed distribution of Genergy with the shift type (see Figure 16). It is curious to observe that hidden under the log Genergy distribution there are two very different distributions belonging to the two types of shift. The first (N - Preparation) has a lower density and on average a lower energy. The second distribution (W: Coal getting) has a higher density and on average a greater energy.

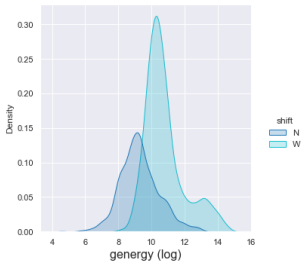

Figure 16: Shift distribution in *genergy*

Gpuls The distribution of the attribute Gpuls is a right-skewed distribution, in fact, as we can see in Figure 17, the majority of the distribution is located near the value zero. It is between the values 2 and 4518.

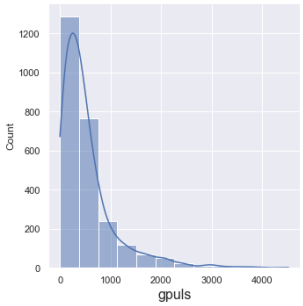

Figure 17: Distribution of *gpuls*

We applied a logarithmic transformation with the goal to have a more accurate Gaussian distribution, as shown in Figure 18.
Finally, we divided the logarithmic distribution according to the type of shifts. The result shown in

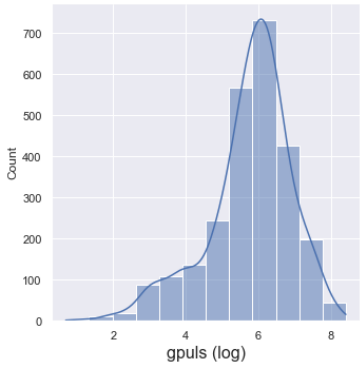

Figure 18: Transformed distribution of *gpuls*

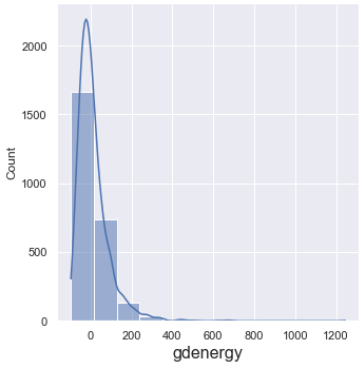

Figure 20: Distribution of *gdenergy*

Figure 19 indicates a peak shifted on the right in case of Coal-getting shift (W), because on average a bigger number of pulses occur during the working shifts than during the preparation shifts. That shows how the activity of coal-getting produces a greater seismic energy and a greater number of seismic pulses.

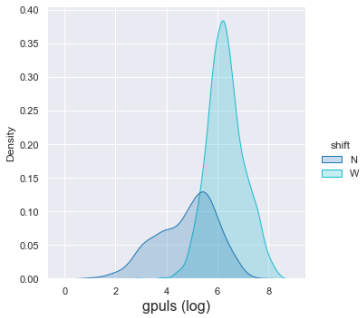

Figure 19: Shift distribution in *gpuls*

Gdenergy As shown in the Figure 20, the distribution of Gdenergy is again a right-skewed one. It's range is between -96 and 1245. Since this attribute refers to the deviation of seismic energy from the average seismic energy recorded during eight previous shifts, it's valuable to notice that the majority of records are slightly below zero.
As we did before we performed a variable transformation by adding the absolute value of the minimum record plus one to all values, due to the negative records. Then we applied the logarithmic transformation in order to have a distribution that is as close to a Gaussian as possible.

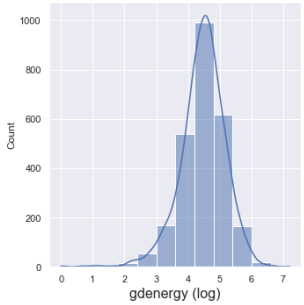

Figure 21: Transformed distribution of *gdenergy*

Gdpuls As shown in the Figure 22, the distribution of Gdpuls is again a right-skewed one. It's range is between -96 and 838. Since this attribute refers to the deviation of pulses from the average number of pulses recorded during eight previous shifts, it's valuable to notice that the majority of records are slightly below zero as before.
We followed the same process we performed for the transformation of Gdenergy. As shown in the figure below, now the distribution assumes approximately a Gaussian shape.

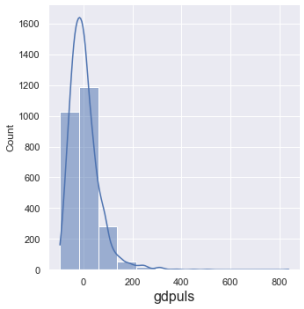

Figure 22: Distribution of *gdpuls*

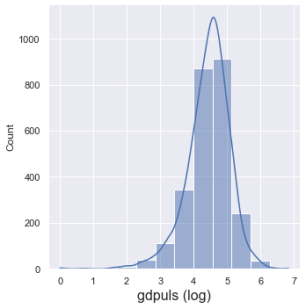

Figure 23: Transformed distribution of *gdpuls*

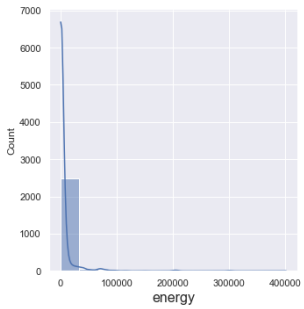

Figure 24: Distribution of *energy*

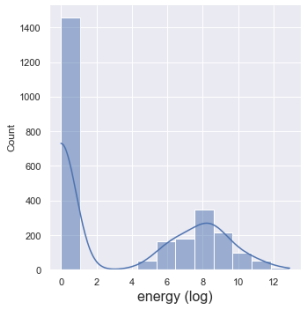

Figure 25: Transformed distribution of *energy*

Energy Figure 24 shows the distribution of Energy, also in this case the right tail is longer (positive skew) and the mass of the distribution is concentrated on the left of the figure. That is because in most of the records no seismic bumps occured and so the value of the total energy recorded is 0. The range of the variable is between 0 and 402000.

In order to make the distribution of energy close to a Gaussian we've added 1 to the energy value for each record, and then we've applied the logarithmic transformation. In the Figure 25 below the new distribution of log-energy. As we can see there are two peaks; the first one corresponds to the records where no seismic bumps occurred, instead the second peak matches the distribution of records where one or more seismic bumps occurred. Figure 26 divided the distribution of log-energy according to the shift type. It's interesting to observe that while the first peak is homogeneously composed by the two types of shift, the second peak presents a much higher density of coal-getting shift and so most of the energy is recorded during this kind of shift. This is due to the fact that the 64.5% of the records consist of coal getting shift but also because it is more likely that one or more seismic bumps occurred in that type of shift.

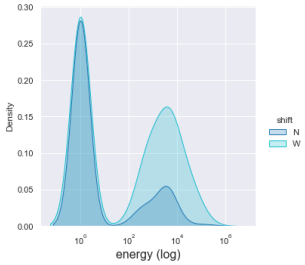

Figure 26: Shift distribution in *energy*

Max Energy Since max energy is a fraction of energy, the two variables are highly correlated. Its distribution is similar to the distribution of energy, indeed as we can see is positively skewed. The variable ranges from a min of 0 to a max of 400000. The same transformations performed to Energy have been applied to Max Energy.

1.3 Assessing Data Quality

We started the assessing data quality phase by checking the semantic quality of the dataset. There is no need to data casting any attribute.
We also performed an analysis for the presence of missing values, which was negative.
Since a malfunction in the measuring system is a real risk, we checked if there were any zero values in Genergy. Again, the result was negative, so the geophones had no problems.
We found that there were 6 duplicates, specifically there were 3 pairs of rows with identical values. It is unlikely that there are one after the other records or like in our case pairs of identical records. In addition, it is mathematically impossible to have the same values in the deviation attributes (gdenergy and gdpuls) in pairs of subsequent rows. So, we decided to remove them.
We performed a further step by checking the relationship between Nbumps(n) and Energy. In particular, we checked, for each record, if when Nbumps is equal to zero also the value of Energy is zero.
About the presence in the dataset of outliers we have to make a premise; seismic events are characterized by big disproportion between the number of low-energy seismic events and the number of high-energy phenomena, this lead to a situation where the most interesting instances of the dataset occur in the form of outliers compared to the others. This fact has become evident in the distribution of the variables above, where, most of them, present a long right tail. Since there is no way to verify the accuracy of the measurements in the dataset and since the records with high energy values are the most interesting to investigate, no instances were removed from the dataset.

1.4 Pairwise correlations and elimination of variables

Analyzing the correlation matrix (considering only the continuous variables) it is clear that there are some variables with a high value of correlation. Some results confirmed our expectations, because they take into consideration two aspects of the same phenomenon but still following the same trend, such as Genergy-Gpuls and Gdenergy-Gdpuls, respectively with pearson coefficients of 0.75 and 0.81.
Different is the case of Energy and Maxenergy where the latter is directly included as a factor in the computation of the former, and therefore we expected a high coefficient of correlation as actually happened (0.99). We decided to drop only "Maxenergy" in order to avoid redundancy and bias in clustering, and to keep the other variables. We made the decision to keep "Energy" as it contains the value of Maxenergy and also captures the energy released by the seismic phenomenon in its entirety and not only the maximum peak.
The other two pairs of variables are highly correlated but they have different meanings and capture different aspects of the seismic phenomenon which are interesting to include in the clustering analysis.

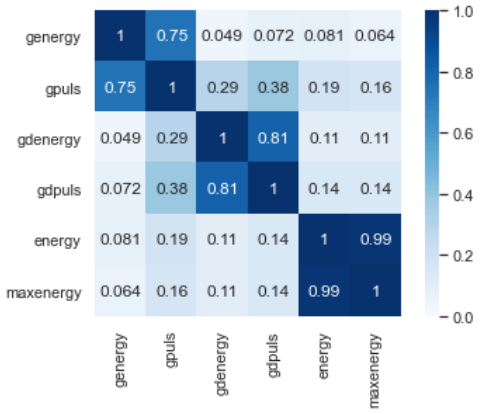

Figure 27: Correlation Matrix of the attributes

2 Clustering

Clustering analysis belongs to unsupervised machine learning techniques. The main goal of the task is to discover intrinsic and hidden structure in the data by finding some commonality in the attributes.
Therefore, the output of clustering algorithms are clusters, that are groups of elements which share among them common features and at the same time are different from the elements belonging to other clusters.
In the following section we analyze one by one the results we have obtained with the three most common clustering algorithms: K-means, Hierarchical clustering and DBscan (density based clustering).
The section ends with a final discussion in which we compare the obtained results in order to choose, among the three algorithms, the one that provides the better result.
There are two pre-processing steps before initializing the clustering algorithms. Firstly, we have to decide which variables we need to use. We know that the three techniques we are going to apply require only continuous attributes. Thus, in our dataset; genergy, gpuls, gdenergy, gdpuls and energy.
In order to decide which of these variables we need to utilize, we use the correlation matrix; variables that are highly correlated provide similar information about the objects and lead them in similar positions in the space. As we have already seen in the previous section, the pairs genergy - gpuls and gdenergy - gdpuls are strongly correlated.
We decided to pick genergy and gdpuls so as to maintain information both on the energy of the seismic activity and the pulse.
The three chosen variables are genergy, gdpuls and energy.
The second step consists in normalizing the variables. Variables with a wider range bias the clustering results and dominate in the determination of cluster composition, this is a situation that should be avoided.
As our dataset is rich in outliers that have an important meaning and can not be removed, we decided to normalize the data using the robust scaler which can handle their presence.

2.1 K-Means Clustering

There are two key points to set, that are fundamental to get good results, before the initialization of the k-means algorithm:

- determine the correct number of clusters;
- choose the proper initial centroids.

In order to find the best number of clusters (k), we performed the Elbow method. It consists of plotting the SSE for k in range [2,51], where 51 is the square root of the number of the objects.

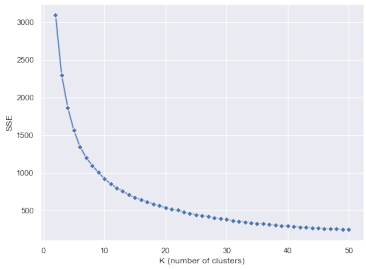

Figure 28: *SSE* for k number of cluster

As we can observe in the figure above, the most suitable k are in the range between 3 and 6 (the points at the elbow).
In order to obtain a more precise k, we also computed the average Silhouette coefficient. In the figure below, we can see that k equal to three has the highest value, so combining the two methods we decided to initialize the k-means algorithm with three clusters.

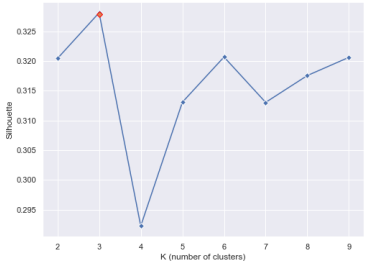

Figure 29: *Silhouette* for k number of cluster

The selection of initial centroids was performed using

the K-means++ approach . This procedure provides significantly better clustering results in terms of lower SSE.
Figure 30 shows the result of clustering: the division of the sample according to the three clusters generated, red points identify the clusters' centroids.

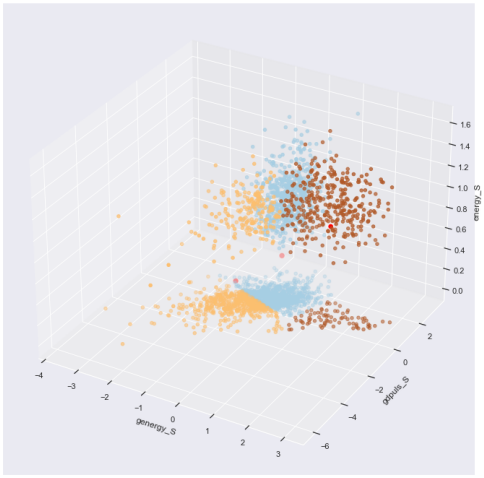

Figure 30: 3D visualization of K-Means

- cluster A correspond to the light blue cloud;
- cluster B corresponds to the orange cloud;
- cluster C corresponds to the brown cloud.

We can see that the sample has assumed two well-defined distributions in space, depending on whether the value of energy is zero or whether seismic bumps occurred.
Therefore, the value of energy is positive.
The instances are not well distributed among the three clusters, as we can observe that their size is quite different.

We have analyzed the centroids through a parallel coordinate plot Figure 31. This method helped us to visualize the values assumed by the centroids for the three attributes used.

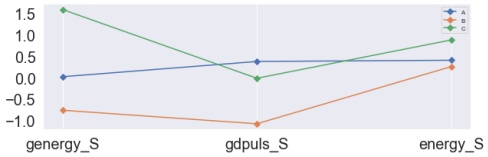

Figure 31: Centroids attribute values

The interquartile range for both genergy and gdpuls is approximately $[-0.5, 0.5]$. The values assumed by the three centroids for these two attributes are well enough separated. Howerever, the interquartile range of energy is quite different $\sim [0, 1]$. Even though from the graph the values taken by the centroids seem quite close, they are not that close, especially the cluster "C".
Eventually, we analyzed the composition of the clusters with respect to the target variable to see if there is a greater presence of instances with class equal to 1 - "hazardous state" - in one of the three clusters.

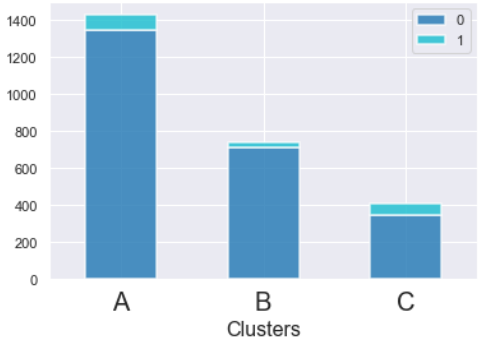

Figure 32: Distribution of *class* in clusters

As Figure 32 shows, there is a higher frequency percentage about instances of class 1 in cluster A and C. However, cluster C has a lower size, this could mean that it is able to "capture" more "hazardous state" instances.

2.2 DBSCAN

The second clustering algorithm we have applied is DBSCAN, which requires to tuned two parameters: MinPts and Eps.
We set MinPts equal to 6, i.e. the number of attributes used in the algorithm multiplied by two.
In order to determine the most suitable Eps for MinPts equal to 6, we plot the point sorted according to the distance of the 6th nearest neighbor versus the 6th nearest neighbor distance.

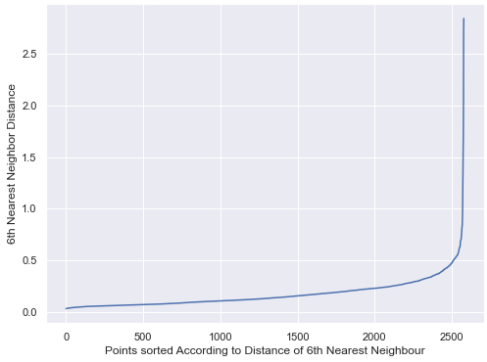

Figure 33: Elbow Curve

We picked the value of where the curve presents a sharp change. According to Figure 33, we set $Eps = 0.5$.
Figure 34 shows the 3D plot of the clusters obtained with DBSCAN.
As we can see the algorithm found 2 clusters that groups the data in a quite balanced way. Cluster 0, which corresponds to the orange cloud, counts 1452 instances. Instead Cluster 1 (the brown one) counts 1105 instances.
The algorithm also found 21 noise points, colored in light blue.

Table 1 shows the median values for the two clusters for the 3 variables.

Figure 35 shows that there is a higher presence of "hazardous state" instances in cluster 1 than in cluster 0.

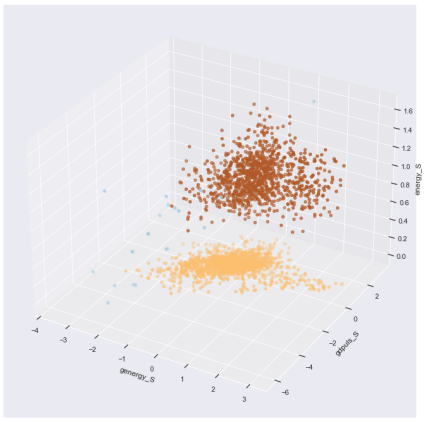

Figure 34: 3D visualization of DBSCAN

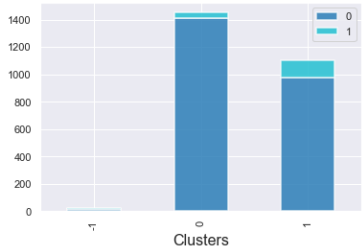

Figure 35: Distribution of *class* in clusters

Attribute	**Cluster 0** (Median)	**Cluster 1** (Median)
genergy	16490	44010
gdpuls	-11	0
energy	0	3400

Table 1: Median values of attributes

2.3 Hierarchical Clustering

The last clustering algorithm we have applied is hierarchical clustering. We set as a metric to compute the linkage between clusters the euclidean distance.
The key operation of hierarchical clustering is defining the proximity between clusters. In order to understand which proximity measure gave us the best results, we analyzed the dendrograms for 4 different proximity approaches: Single Link, Complete Link, Group Average, Ward's Method.

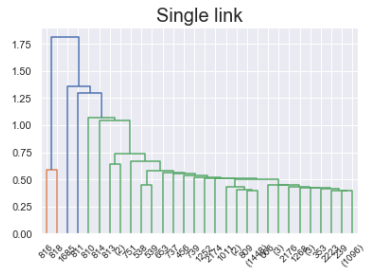

Figure 36: Single Link

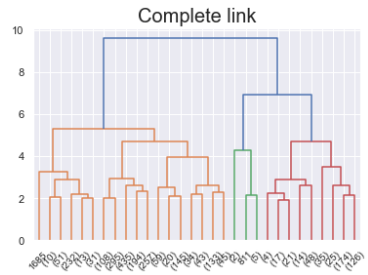

Figure 37: Complete Link

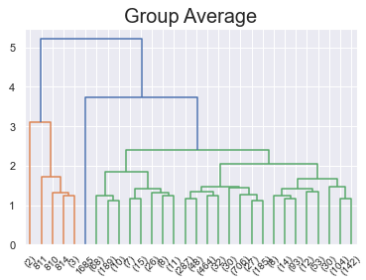

Figure 38: Group-Average

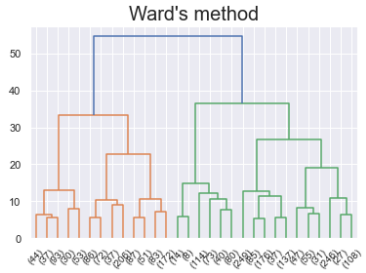

Figure 39: Ward's Method

Among the 4 dendrograms, the most balanced, regarding the composition of the clusters, is the one corresponding to the ward method. Therefore, we decided to focus on this one.
In addition, the dendrogram seems to suggest that the number of clusters suitable for our dataset is two. We verified this hypothesis starting from a qualitative analysis (studying the dendrograms) through the silhouette score (quantitative analysis).
Figure 40 represents the silhouette score (y-axis) and the number "k" of the cluster (x-axis) with k up to 10, for the 4 different types of proximity measures. Ward method silhouette score decreases with k greater than two. The same trend follows the other proximity approaches, except the complete link that decreases with k greater than 3. Therefore, the number of clusters equal to 2, suggested by the dendrogram, is confirmed by the silhouette score.

Setting the metric (Euclidean distance), the proximity measure (ward method) and the number of clusters (2), we have initialised the hierarchical algorithm. The 3D graph (Figure 41) shows the shapes of the clusters.
As we did before, we reported in Table 2 the median values of the two clusters for the three variables, as well in this case Cluster 1 reports higher values.

Attribute	**Cluster 0** (Median)	**Cluster 1** (Median)
genergy	16320	53390
gdpuls	-29	38
energy	0	400

Table 2: Median values of attributes

However, the instances of class 1 are not unbalanced towards cluster 1, as it is possible to see in Figure 42.

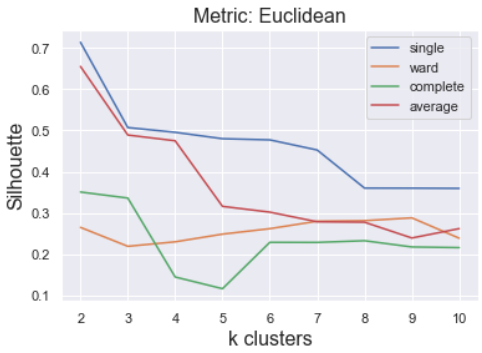

Figure 40: *Silhouette* of hierarchical methods

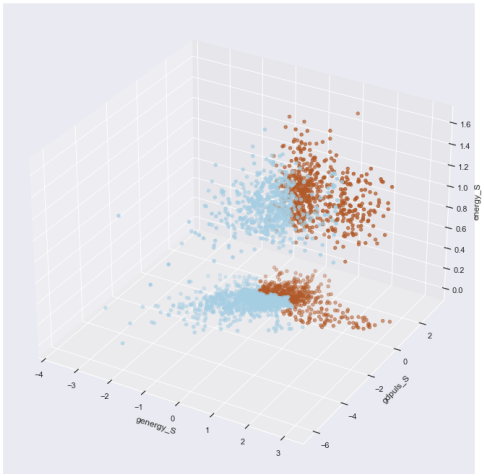

Figure 41: 3D visualization of Hierarchical

2.4 Final Discussion

To compare the results obtained by the three clustering algorithms, we computed the silhouette score (Table 3)

Clustering technique	Silhouette Score
K-means	0.33
DBSCAN	0.26
Hierarchical	0.27

Table 3: Clustering final score

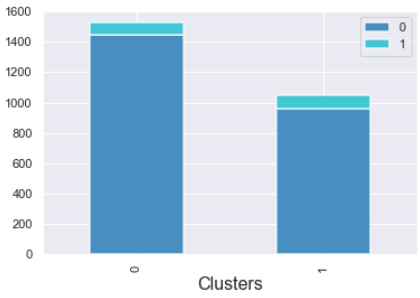

Figure 42: Distribution of *class* in clusters

As we can see, the k-means algorithm presents a higher silhouette score. Consequently, silhouette is not suitable to work with DBSCAN and in particular with density based clustering techniques.
We proceed the analysis of the results from a qualitative point of view.

K-Means. Clusters have very different sizes. However, the centroids are well distributed and cluster "C", which presents high values of energy and genergy, has a relatively high presence of class 1 elements within it.

DBSCAN. Clusters are biased by energy attributes, as it is possible to see in Figure 34. The clusters follow the division between instances in which seismic bumps occurred ($energy > 0$) and instances without the occurrence of seismic bumps ($energy = 0$). However, if we look at the distribution of $class = 1$ within the two clusters, there is a clear bias towards cluster 1.

Hierarchical. Clusters are quite well distributed but the silhouette score is lower than k-means. The two clusters have a homogeneous distribution of instances of class 1, this means that the two clusters do not succeed in picking the division between hazardous state and non hazardous state.

3 Classification

In this section we performed the classification by using three different algorithms: Decision Tree, KNN and Random Forest.
The aim of this task is to build models to predict the target variable on unknown instances. Class gives us information about whether or not a rockburst will occur in the next shift.
The three classifiers used the following attributes:
Numerical: Genergy, Gpuls, GDenergy, GDpuls, Energy, Nbumps;
Categorical: Shift.

We discarded Nbumps 2-3-4-5, Seismic, Seismoacoustic and Hazard, because they are irrelevant and redundant attributes. Furthermore, after performing classification tests (see Figure 44), we removed those categorical attributes because their "feature importance" ended up to be poor. This measure is calculated as the decrease in node impurity weighted by the probability of reaching that node.
Although "Shift" has low values, we have decided to keep it because in the data understanding section it was interesting in combination with other attributes as well as with the target variable.

We used the Holdout method to divide the dataset. Therefore, we assigned 70% of the instances to the training set and 30% to the test set.
Moreover, the dataset is strongly imbalanced according to the target variable, consequently we performed two sampling techniques on the training set:
Oversampling: duplicates samples from the minority class;
SMOTE: (Synthetic Minority Oversampling Technique) generates new objects of the minority class by observing the most similar pre-existing objects.
The goal is to obtain a balanced distribution of the target variable, in order to improve the result of the classifier.
We show the results of the Decision tree and Random forest classifiers for both the sampling techniques, and for the original dataset, named as "no sampling".

3.1 Classification by Decision Tree

We began the process by tuning the hyperparameters, i.e. applying the Random Search algorithm to the training set.
Note that the algorithm gives the best configuration of parameters for the models while trying to maximize the "recall". We chose this scoring because we want to minimize the false negative by predicting the positive class, i.e. we want to prevent scenarios that might be dangerous for the workers!
The algorithm trained different classifiers, working with a different combination of parameters, such as:
Max depth: the maximum depth of the tree;
Min sample split: the minimum number of samples required to split an internal node;
Min sample leaf: the minimum number of samples required to be at a leaf node;
Criterion: Gini or Entropy.

Figure 44 shows the models, its parameters and the results of the performance evaluation metrics for both training and test set:

	Modello 1	**Modello 2**	**Modello 3**
	No Sampling	*Random Oversampling*	*SMOTE*
Criterion	Gini	Entropy	Gini
Max Depth	12	None	6
Min Sample Split	18	5	3
Min Sample Leaf	1	3	3
Training Accuracy	0,9484	0,9929	0,8267
Training F1	0,7046	0,9900	0,8200
Test Accuracy	0,9121	0,9000	0,7041
Test F1	0,5295	0,6000	0,5100
Test AUC-ROC	0,5896	0,6010	0,6480
Test AUC-Precision Recall	0,1611	0,2490	0,1150

Figure 43: Hyperparameters Tuning Results

Regarding the training set, the three models produce a very high accuracy, with SMOTE technique a bit lower.
For the test set, we can notice a decrease in terms of accuracy for all three models, that is in line with our expectations.
In order to compare the performance between the models, we performed the analysis using the

Modello 1 *with categorical*	**Modello 2**	**Modello 3** *with categorical*	**Modello 4**	**Modello 5** *with categorical*	**Modello 6**
No Sampling	*No Sampling*	*Random Oversampling*	*Random Oversampling*	*SMOTE*	*SMOTE*
Gpuls : 0.21	Genergy : 0.20	Genergy : 0.26	Genergy : 0.24	Energy : 0.25	Energy : 0.56
Genergy : 0.16	Nbumps : 0.20	Energy : 0.21	Energy : 0.22	Nbumps : 0.23	Genergy : 0.20
Gdenergy : 0.15	Gpuls : 0.16	Gdenergy : 0.16	Gdenergy : 0.17	Seismic : 0.16	Gdenergy : 0.07
Nbumps : 0.14	Energy : 0.15	Gpuls : 0.15	Gdpuls : 0.17	Genergy : 0.14	Nbumps : 0.07
Gdpuls : 0.14	Gdenergy : 0.14	Gdpuls : 0.14	Gpuls : 0.14	Gpuls : 0.06	Gpuls : 0.04
Energy : 0.14	Gdpuls : 0.14	Nbumps : 0.03	Nbumps : 0.04	Gdenergy : 0.06	Gdpuls : 0.03
Seismic : 0.04	Shift : 0.02	Seismic : 0.02	Shift : 0.09	Gdpuls : 0.05	Shift : 0.02
Seismoacoustic : 0.01		Seismoacoustic : 0.01		Seismoacoustic : 0.03	
Shift : 0.0		Shift : 0.01		Shift : 0.01	
Hazard : 0.0		Hazard : 0.0		Hazard : 0.0	

Figure 44: Feature Importance

Precision-Recall curve instead of the ROC curve. The reasons that led us to choose the PR curve were that it is more suitable when the data is imbalanced. For an unbalanced dataset with a positive rate up to 10%, the optimal value for the AUC Precision-Recall Curve is around 0.50.
In our case, we have a positive rate of 6%, as we can see in figure n, we chose the model with the higher value of AUC Precision-recall, i.e. Random Oversampling.
This model has an AUC Precision-Recall curve of almost 0,25; it is shown in the Figure 45.

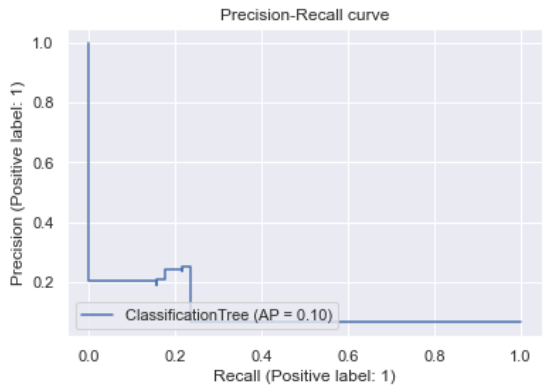

Figure 45: Precision-Recall Curve

The representation of the decision tree is shown in Figure 47. We can see that Energy is the first attribute used as a test condition for the root node, while entropy is the chosen measure of impurity.
The left child node has 1530 instances of which 1120 are negative classes, thus the node is labeled as class 0.
Although we cut the representation of the decision tree early, it is interesting to note the presence of two leaf nodes which classify the unknown instances as negative class.

Below the confusion matrix obtained from the Random Oversampling Model; among the 51 actual positives the model correctly predicts only 13 of them, thus occurring 38 false negatives, which as said previously, is the situation we absolutely want to avoid.

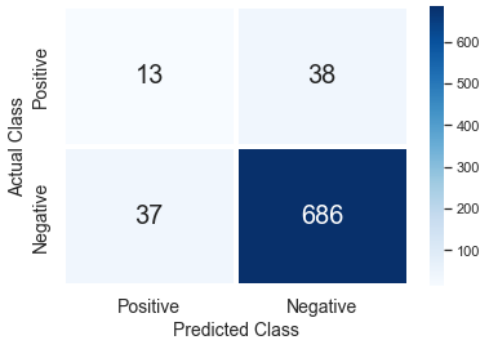

Figure 46: Confusion Matrix - Test Set

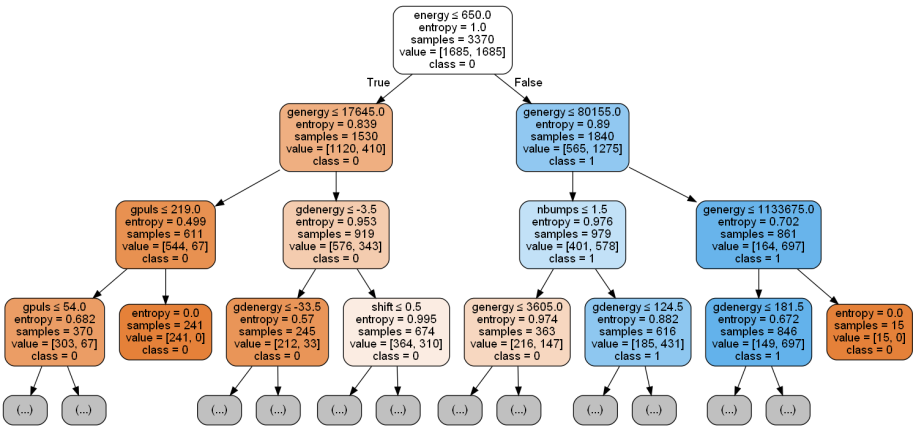

Figure 47: Decision Tree visualization

The performance of the Random Oversampling model on the test set is shown in the table below.

Target	Precision	Recall	F1-Score
Non-hazardous state [0]	0,95	0,95	0,95
Hazardous state [1]	0,26	0,26	0,26
Macro average	0,6	0,6	0,6

Figure 48: Performance evaluation metrics

3.2 Classification by KNN and Random Forest

The KNN requires to set the value of k, that is the number of nearest neighbors. A general practice is to set k as the square root of the number of instances in the training set, i.e 48. This is because we used the 10-fold cross validation that, for every run, assigned 90% of the instances to the training set and the remaining 10% to the test set.
We exploited the default measure, i.e the Minkovski distance, and performed the model both with a classification function that weighted the vote according to the distance of the nearest neighbors and uniform. Below, the results in terms of overall accuracy and F1-score.
We performed the hyperparameter tuning of the Random forest by using the random search algorithm.

Classification function	Accuracy	F1-Score
Weight	0,9317	0,4823
Uniform	0,9341	0,483

Figure 49: KNN Results

The table shows the result for No Sampling, Random Oversampling and SMOTE.

	Modello 1	Modello 2	Modello 3
	No Sampling	*Random Oversampling*	*SMOTE*
Number of estimators	10	48	72
Min Sample Split	25	10	8
Min Sample Leaf	5	2	2
Max Features	sqrt	sqrt	auto
Max Depth	10	12	10
Training Accuracy	0,936	0,993	0,951
Test Accuracy	0,933	0,922	0,846
Test AUC-ROC	0,7394	0,7235	0,7267
Test AUC-Precision Recall	0,1821	0,2519	0,2526

Figure 50: RandomForest Results

According to the AUC Precision-Recall curve the best model is the Random Oversampling.

3.3 Final Discussion

In general, our dataset is strongly unbalanced. Using Random Oversampling on the training set, the results of the different models showed a slight improvement over the results of the original training set.

As shown in the previous tables, the Decision Tree and Random Forest results are very similar in terms of AUC Precision-Recall curve. According to this measure the Random Forest is slightly better than the Decision Tree.

4 Pattern Mining

The last task of the report is to mine frequent sequential patterns, that is to say finding statistically relevant patterns within data. In order to do this we used the Apriori algorithm.

4.1 Pre-processing

There is one fundamental pre-processing step that is the discretization of the continuous attributes. In our case; genergy, gpuls, gdenergy, gdpuls and energy. We used a quantile based discretization function in order to obtain, for each continuous attribute, four bins of equal size. It is important to underline that energy is divided only in two bins due to its distribution composed mainly by zeroes values, i.e. Energy (0.0,2675], (2675,402000].

4.2 Frequent Itemset

In order to obtain frequent itemsets we applied the Apriori algorithm with zmin (minimum number of items per itemset) equal to 3, and a minsup (minimum support value) equal to 3%. We have obtained the value of the minsup by analysing the graphs in Figures 51 and 52.
As we can see, the number of frequent itemsets collapses with minsup values exceeding $10 - 12\%$. We performed a further analysis in order to obtain frequent itemsets with class equal to 1. From the graph we can see that the number of frequent itemsets with class equal to 1 collapses with minsup values greater than or equal to 4%. Thus, we performed our analysis with minsup equal to 3%.
Figure 53 shows the first nine frequent (and closed) itemsets with the highest support value.
The analysis of the frequent itemset shows that most working shifts present a lack of hazard status in the seismic, seismoacoustic and hazard methods. Where no seismic bumps or low energy seismic bumps occurred, they have a non-hazardous state.

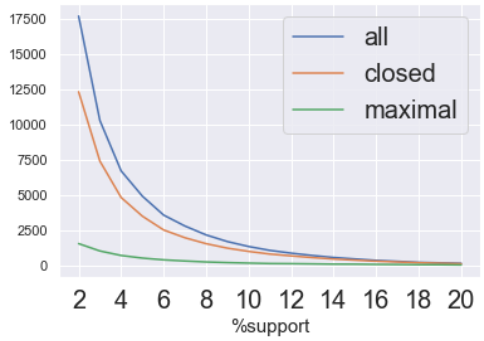

Figure 51: Number of Frenquent Itemsets for different *support*

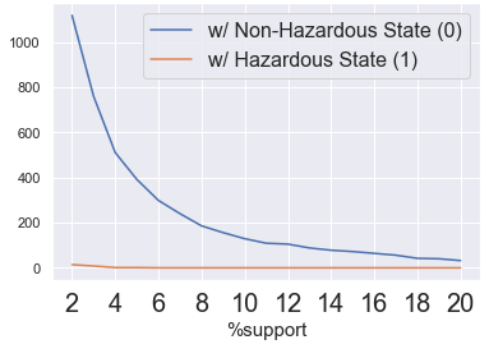

Figure 52: Frequent Itemsets for each *class*

Concerning the maximal frequent itemset, the table below shows the first five maximal frequent itemsets with the highest support value.

4.3 Association Rules

The main goal of this section is to discuss the most interesting association rules obtained with the Apriori algorithm.
From the frequent itemsets previously obtained, we have extracted 39,705 rules by setting minconf (minimum confidence threshold) equals 60%. In Figure 55, we can see as the minconf increases, the number of rules decreases linearly.

Support	Frequent-Closed Itemsets
60% - 70%	{Energy:(0.0,2675], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]} (support = 0.64)
50% - 60%	{Seismic: Lack of hazard [a], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]} (support = 0.57)
	{Seismoacoustic: Lack of hazard [a], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]} (support = 0.56)
	{Nbumps: 0, Energy:(0.0,2675], Class: Non-hazardous state [0]} (support = 0.55)
	{Shift: Coal-getting [W], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]} (support = 0.52)
	{Nbumps: 0, Energy:(0.0,2675], Hazard: Lack of hazard [a]} (support = 0.50)
40% - 50%	{Nbumps: 0, Energy:(0.0,2675], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]} (support = 0.49)
	{Seismic: Lack of hazard [a], Energy:(0.0,2675], Class: Non-hazardous state [0]} (support = 0.49)
	{Seismic: Lack of hazard [a], Energy:(0.0,2675], Hazard: Lack of hazard [a]} (support = 0.47)

Figure 53: Frequent-Closed Itemsets

Support	Maximal Itemsets
0,06	**1)** {Seismic: Low hazard [b], Nbumps: 0, Seismoacoustic: Lack of hazard [a], Shift: Coal-getting [W], Energy:(0.0,2675], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]}
0,05	**2)** {Energy:(2675, 402000], Seismoacoustic: Lack of hazard [a], Shift: Coal-getting [W], Seismic: Lack of hazard [a], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]}
	3) {Genergy: (11692,25500], Nbumps: 0, Shift: Coal-getting [W], Seismic: Lack of hazard [a], Energy:(0.0,2675], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]}
	4) {Seismoacoustic: Low hazard [b], Nbumps: 0, Shift: Coal-getting [W], Seismic: Lack of hazard [a], Energy:(0.0,2675], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]}
	5) {Gdpuls: (-6, 30], Nbumps: 0, Seismoacoustic: Lack of hazard [a], Seismic: Lack of hazard [a], Energy:(0.0,2675], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]}

Figure 54: Maximal Itemsets

However, the histogram below presents a particular trend for rules equal to 60%. In fact, the number of rules is roughly constant for values of confidence between 60% and 95%, while we have a peak of rules for values equal to 100%. This is due to the fact that the values of the attributes are concentrated around a value, or a narrow range of values, as we have observed in the data understanding section.

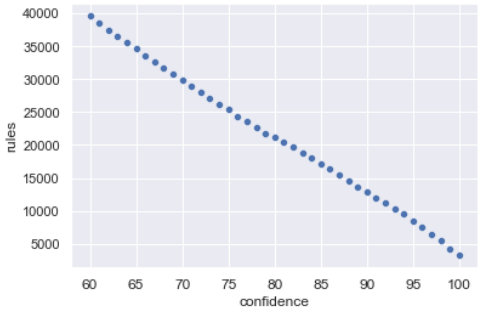

Figure 55: Number of rules for different *minconf*

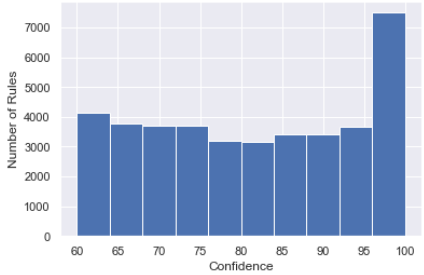

Figure 56: Number of rules for their confidence

The following graph shows as many rules present a $lift = 1$, which means that antecedent and consequent are statistically independent. In addition, there is an interesting number of events that are positively correlated.

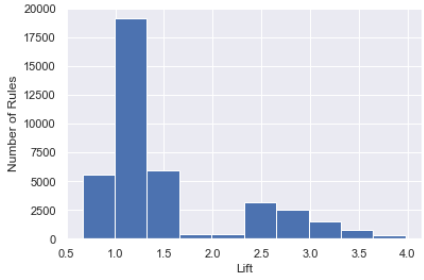

Figure 57: Distribution of Lift

Figure 58 shows the most significant association rules, in terms of Lift, with minsup 3% and minconf 60% ($zmin = 3$):

Association Rules (minsup = 3% , minconf = 60%)
{Genergy: (100, 11692], Gdpuls: (-96, -36], Shift: Preparation [N], Seismic: Lack of hazard [a], Hazard: Lack of hazard [a]} => Gpuls: (2, 191] (conf = 0.99) (lift = 3.97)
{Genergy: (100, 11692], Gdpuls: (-96, -36], Shift: Preparation [N], Seismic: Lack of hazard [a]} => Gpuls: (2, 191] (conf = 0.99) (lift = 3.97)
{Genergy: (100, 11692], Gdpuls: (-96, -36], Shift: Preparation [N], Seismic: Lack of hazard [a], Hazard: Lack of hazard [a], Class: Non-hazardous state [0]} => Gpuls: (2, 191] (conf = 0.99) (lift = 3.97)

Figure 58: Association Rules

As we can observe, the shifts in a coal mine with a low number of pulses are usually preparation shifts with a lack of hazard status in terms of seismic and hazard methods, and a low seismic energy.

Association rules with a "Non hazardous state" as a consequent are showed in Figure 59:

Association Rules (minsup = 3% , minconf = 60%)
{Nbumps: 1, Genergy: (25500, 52757], Energy: (2675, 402000]} => Class: Non-hazardous state [0] (conf = 1.00) (lift = 1.07)
{Nbumps: 1, Gpuls: (380, 669], Seismic: Lack of hazard [a]} => Class: Non-hazardous state [0] (conf = 1.00) (lift = 1.07)
{Gdpuls: (-36, -6], Genergy: (100, 1693], Shift: Preparation [N], Nbumps: 0, Seismic: Lack of hazard [a], Energy: (0, 2675], Hazard: Lack of hazard [a]} => Class: Non-hazardous state [0] (conf = 1.00) (lift = 1.07)

Figure 59: Association Rules as consequent *Negative class*

In order to find association rules with "Hazardous state" as a consequent, we had to decrease the minimum support a the level of 30%:

Association Rules (minsup = 3% , minconf = 30%)
{Gdpuls: (-36, -6], Gpuls: (669, 4518], Genergy: (52758, 2595650], Seismoacoustic: Lack of hazard [a], Shift: Coal-getting [W], Hazard: Lack of hazard [a]} => Class: Hazardous state [1] (conf = 0.30) (lift = 4.66)
{Gdpuls: (-36, -6], Gpuls: (669, 4518], Genergy: (52758, 2595650], Seismoacoustic: Lack of hazard [a], Shift: Coal-getting [W]} => Class: Hazardous state [1] (conf = 0.30) (lift = 4.66)
{Gdpuls: (-36, -6], Gpuls: (669, 4518], Genergy: (52758, 2595650], Seismoacoustic: Lack of hazard [a], Hazard: Lack of hazard [a]} => Class: Hazardous state [1] (conf = 0.30) (lift = 4.60)

Figure 60: Association Rules as consequent *Positive class*

As illustrated, the shifts in a coal mine with an hazardous state are working shifts with the highest value of pulses and seismic energy.

4.4 Assiciation Rules for predicting the target variable

We used the extracted rules for predicting the target variable. With minconf equal to 60% the model has no positive classes as a consequence, we decided to set the threshold to 30% to find such classes. We used the four extracted rules to predict the target variable of the test set instances used in the classification task. The results are shown in the confusion matrix below. Out of a total of 51 shifts with a positive class, only 7 were predicted correctly, while 44 were predicted negatively. In other words, the workers are in danger in 44 shifts.
It means that this predictive method is not very reliable.

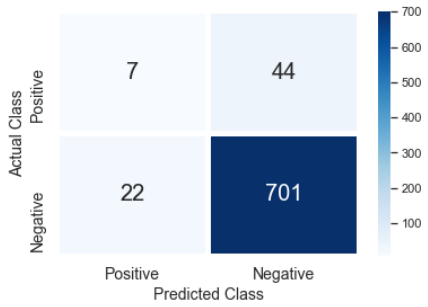

Figure 61: Confusion Matrix - Predicted Rules