PAPER

Volcanic Explosivity Index (VEI) prediction - classification model prediction

Cristina Porta[1],*
[1] IT Academy
* cristinadelarge@gmail.com

Abstract

The world is woefully underprepared for a massive volcanic eruption and the likely repercussions on global supply chains, climate and food. There is a broad misconception that risks of major eruptions are low but data gathered from ice cores on the frequency of eruptions over deep time suggests there is a one-in-six chance of a magnitude seven explosion in the next one hundred years.

This paper tries to analyze the relations between characteristics of the volcanoes to predict possible very large explosions. The results of the study will support further investigations and prevention of potential catastrophes caused by these eruptions.

Key words: *Volcanoes, VEI, Extra Trees Classifier, Supervised Learning, Classification*

I. Introduction

The word "volcano" comes from the Roman name "Vulcan". "But who was Vulcan?" you might ask. He was the Roman god of fire!

A volcano is a rupture in the crust of a planetary-mass object, such as Earth, that allows hot lava, volcanic ash, and gases to escape from a magma chamber below the surface. On Earth, volcanoes are most often found where tectonic plates are diverging or converging, and most are found underwater. For example, a mid-ocean ridge, such as the Mid-Atlantic Ridge, has volcanoes caused by divergent tectonic plates whereas the Pacific Ring of Fire has volcanoes caused by convergent tectonic plates.

Volcanoes can also form where there is stretching and thinning of the crust's plates, such as in the East African Rift and the Wells Gray-Clearwater volcanic field and Rio Grande rift in North America. Volcanism away from plate boundaries has been postulated to arise from upwelling diapirs from the core–mantle boundary, 3,000 kilometers (1,900 mi) deep in the Earth. This results in hotspot volcanism, of which the Hawaiian hotspot is an example. Volcanoes are usually not created where two tectonic plates slide past one another.

Large eruptions can affect atmospheric temperature as ash and droplets of sulfuric acid obscure the Sun and cool the Earth's troposphere. Historically, large volcanic eruptions have been followed by volcanic winters which have caused catastrophic famines. [1]

Motivation

The purpose of this analysis is to apply a supervised classification model to a dataset with information about volcanoes around the world and their eruptions to have a better insight on the volcanic explosivity index in order to be able to predict very large explosions and, thus, prevent potential catastrophes.

II. Methods

Dataset

The data collected in these datasets comes from The Smithsonian Institution. Founded on August 10, 1846, it is a group of museums and education and research centers, the largest such complex in the world, created by the U.S. Government "for the increase and diffusion of knowledge". It has a "Global volcanism program" with a lot of information of volcanoes around the world. Although complete updates are done about every 6-8 weeks, these particular datasets include information until 2020. [2]

Data preparation

Dealing with missing values and non-explanatory features:

To perform all of the following, Pandas [3] was used.

We began by exploring the data by analyzing the percentage of missing values on the two datasets used. We found that the percentage of missing values on ten columns of the volcano dataset (the ones referring to type of rocks) was too high and had to be deleted. As we merged both datasets and they contained some identical data, we deleted the duplicates. We also deleted some rows with a lot of missing values such as the end year, month and day. This issue comes from the fact that some of the eruptions occurred in the era before Christ and it is very difficult to have information and certainty about it. Finally, we decided that the evidence method for dating volcanoes, the evidence category and the eruption category don't give us a lot of information since most of the volcanoes are dated using historical observations, most of the eruptions were observed and there were just 5 uncertain eruptions (the other were confirmed) and we deleted these features too.

Transformation of features:

Then, we transformed some categorical data as the type of volcano or the type of tectonic settings to numerical data in order to use it in our prediction model.

Since our targets were so imbalanced (one class had more than 1700 values and other just 1) and in order to have less class types (we had 8) we decide to summarize them into 5: [4]

VEI 0 - non explosive = 0
VEI 1 - small = 1
VEI 2 and 3 -moderate = 2
VEI 4 and 5 - large = 3
VEI 6 and 7 - very large = 4

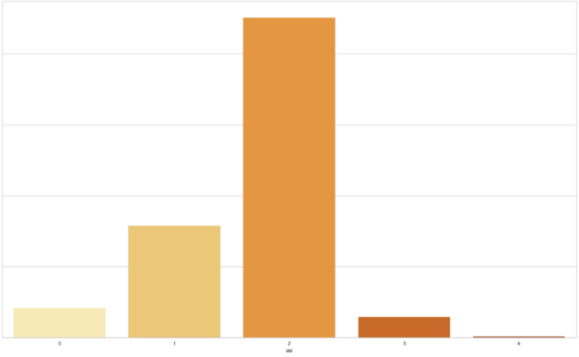


Figure 1. Distribution of the target

Addition of new features:

We decided to add a new column with the total duration (in days) of each eruption because we think it can be useful for the analysis of volcano types and it will help us to build a good model. [5]

After all this data cleaning, we can divide the explanatory variables as follows:

Volcano	Geographical
Volcano number	Latitude
volcano name	Longitude
eruption number	country
primary volcano type	elevation
start year	tectonic settings
start month	population_within_5_km
start day	population_within_10_km
end year	population_within_30_km
end month	population_within_100_km
end day	
last eruption year	
total days of eruption	

Table 1. Variable categories

The correlation matrix heatmap built with the matplotlib library [6] allows us to cover three objectives of the following steps of the exploratory analysis: analyze the existence or not of multicollinearity, evaluate the consistency of the correlations between the explanatory variables and in relation to the target and obtain a first approximation to the predictors of the classification model.

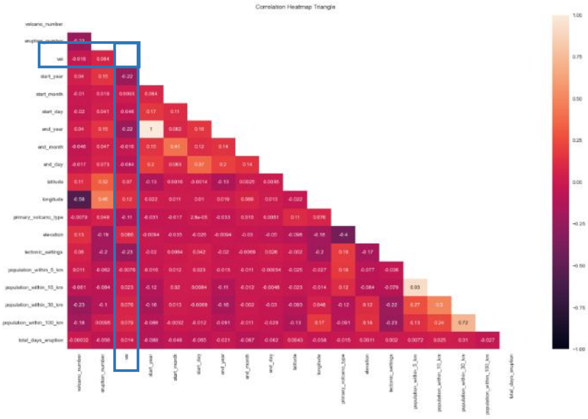


Figure 2. Correlation heatmap

As shown in Figure 2, there is no significant multicollinearity between the explanatory variables, except in the case of start year and end year and population within 5 and 10 km, but we have kept the two variables since they are interesting to obtain a model. We expected a greater correlation between the primary volcano type and the tectonic settings, or latitude/longitude and tectonic settings too.

Using plotly[7] we have analyzed the target variable, the volcanic explosivity index, compared to all the numerical explanatory variables. We found that there are many outliers in almost every comparison. Here are some conclusions of the analysis:

We can see that the types of VEI with more outliers in general are 2 (moderate) and 1 (small). No surprise here because they are the two types with more data by far.

-Latitude: we can see that there are volcanoes with non-explosive eruptions or very large VEI almost equally in both hemispheres. There are more volcanoes with small, moderate and large VEI in the north hemisphere. No significative outliers.

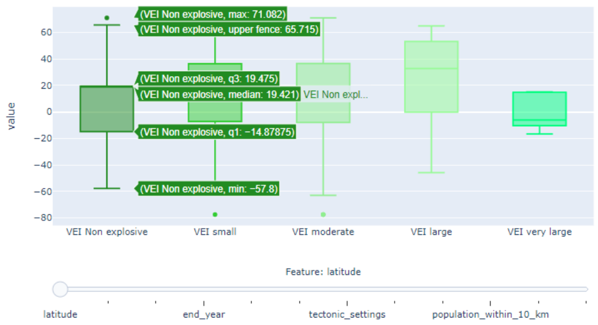


Figure 3. Boxplot VEI-latitude

-Longitude: we can see that there are more volcanoes with non-explosive VEI at the left (west) of the Greenwich meridian: American continent, big part of occidental Africa countries, small part of European countries. There are more volcanoes with small, moderately large and very VEI at the right (east) of the Greenwich meridian. The differences are not huge. No outliers.

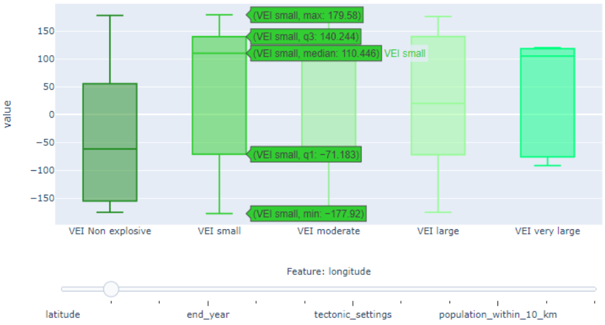


Figure 4. Boxplot VEI-longitude

-Start year: the major part of the eruptions of our dataset occurred from 1700 to nowadays, though volcanoes with moderate and large VEI have many outliers because most of the eruptions are from these two types.

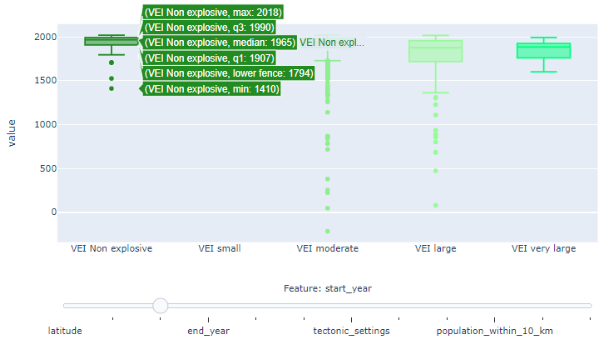


Figure 5. Boxplot VEI-start year

-End year: same issue of the start year happens with the end year, of course.

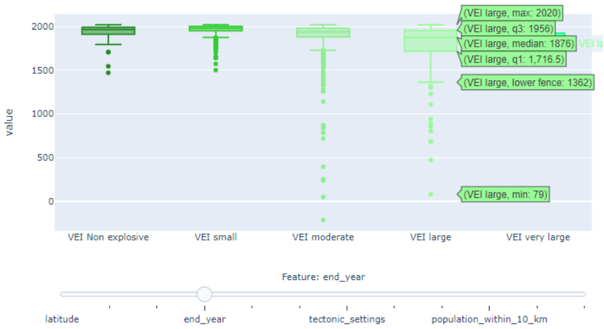


Figure 6. Boxplot VEI-end year

-Elevation: most of the volcanoes have an elevation from 1000 to 3000m. The submarine ones, of course, tend to be non-explosive ones and the ones with very large VEI are the highest ones in general.

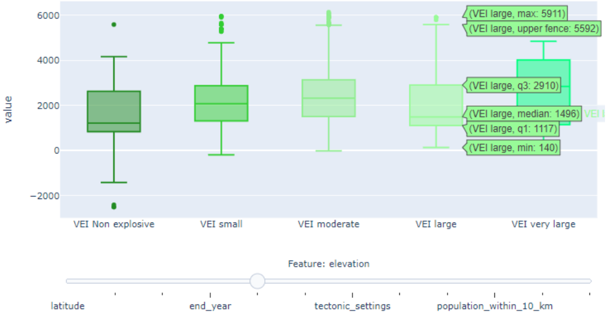


Figure 7. Boxplot VEI-elevation

-Primary volcano type: we can see here that most of the volcanoes are cinder cones, composite or shield ones, especially the ones with non-explosive eruptions or very large VEI. The VEI types with more data are the ones with more outliers too.

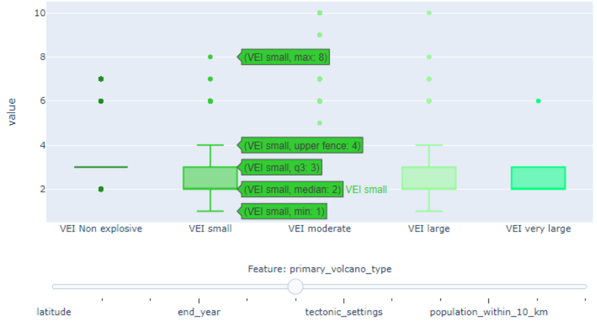


Figure 8. Boxplot VEI-primary volcano type

-Tectonic settings: the eruptions with large or very large VEI are located in the subduction zone, especially in the continental crust. The eruptions with non-explosive VEI are located, generally, or in the subduction or in the intermediate zones. Small and moderate VEI eruptions tend to be located in the subduction area too, but with outliers in the other zones.

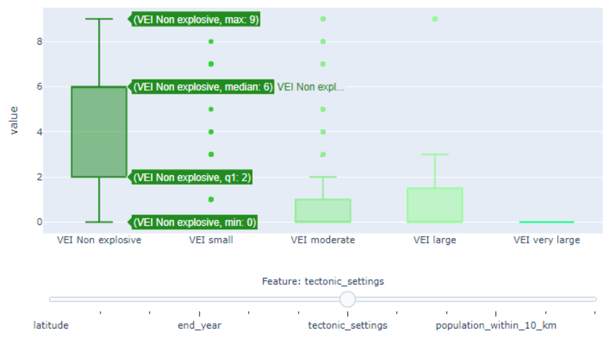


Figure 9. Boxplot VEI-tectonic settings

-Total days of eruption: fortunately, the large and very large VEI eruptions are the ones with less days of duration. As expected, moderate and small VEI eruptions have a lot of outliers and we can see there are eruptions lasting more than 100.000 days!!

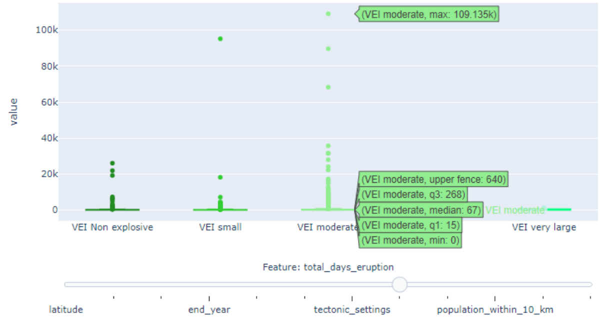


Figure 10. Boxplot VEI-total days of eruption

-Population within 100km: population around the very large VEI eruptions is little compared to the other VEIs, thankfully.

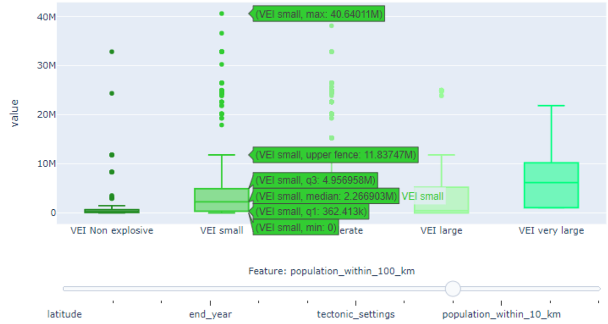


Figure 11. Boxplot VEI-population within 100km

-Population within 10km: population around the very large VEI eruptions is little compared to the other VEIs, thankfully.

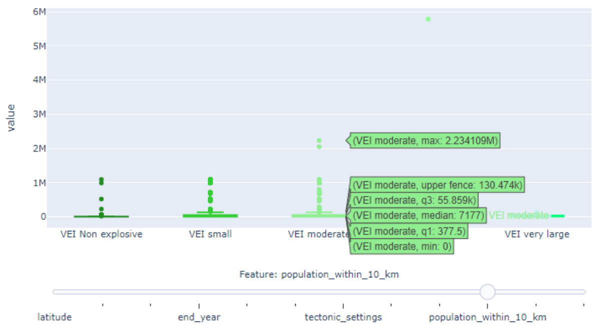


Figure 12. Boxplot VEI-population within 10km

-Population within 30km: population around the very large VEI eruptions is little compared to the other VEIs, thankfully.

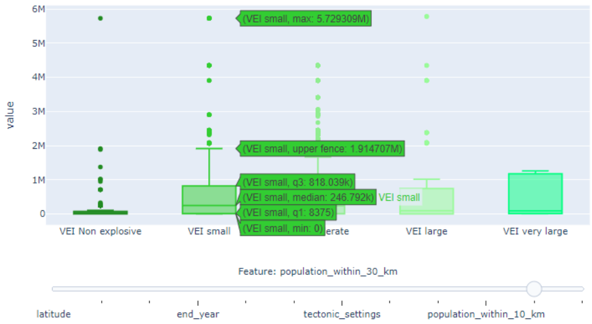


Figure 13. Boxplot VEI-population within 30km

-Population within 5km: thankfully, population around the very large VEI eruptions is little compared to the other VEIs, specially within 5km.

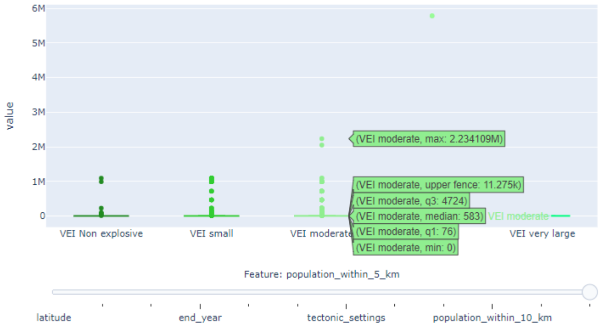


Figure 14. Boxplot VEI-population within 5km

Model building

Preprocessing:

We performed supervised learning. First, using the library Scikit-learn [8], we scaled the numerical data with Robust Scaler because our data has a lot of outliers. This Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).

As we had a very imbalanced dataset, with one class having more than 2000 values and another just 5, we performed the Borderline SMOTE (Synthetic Minority Over-sampling Technique) of the imb-learn library to our dataset. A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Borderline samples will be detected and used to generate new synthetic samples.

Selection of the model:

Then, with the help of the Pycaret library [9], we found out that the best performing models with our dataset are Extra Trees Classifier or Random Forest.

Extremely Randomized Trees Classifier (Extra Trees Classifier) [10] is a type of ensemble learning technique that implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees.

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to his/her choice.

Selection of the metrics:

We will use the most important metrics to compare th e different models [11]:

- **Confusion matrix**

The Confusion Matrix created has four different quadran ts:
True Negative (Top-Left Quadrant) - TN
False Negative (Top-Right Quadrant) - FN
False Positive (Bottom-Left Quadrant) - FP
True Positive (Bottom-Right Quadrant) - TP

True means that the values were accurately predicted, Fa lse means that there was an error or wrong prediction.

- **Accuracy**

correct predictions / total predictions
(TP+TN)/(TP+FP+FN+TN)

- **Precision**

correct positive predictions / total positive predictions
TP/(TP+FP)

- **Recall or sensitivity**

correct positive predictions / total positives
TP/(TP+FN)

- **Specificity**

correct negative predictions / total negatives
TN / (TN + FP)

- **F1 score**

2 x (recall x precision) / (recall + precision)

III. Results

Descriptive analysis

Analyzing the data with matplotlib and Folium [12], we found out that volcanoes are located all over the world although EEUU is the country with more of them (with almost 100), followed by Indonesia, Japan and Russia. [13]

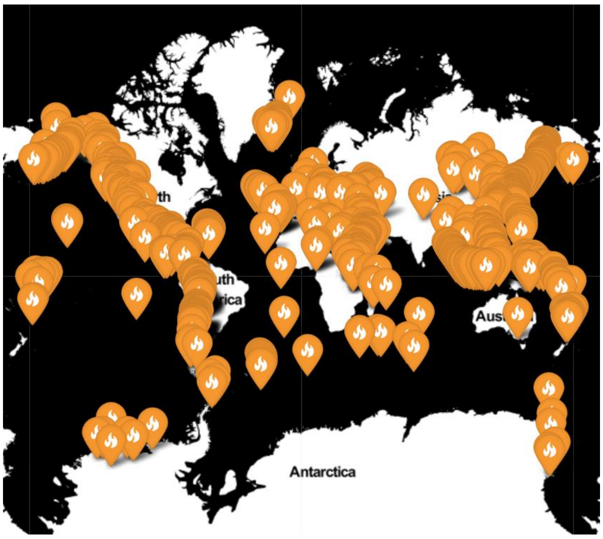


Figure 15. Volcano map

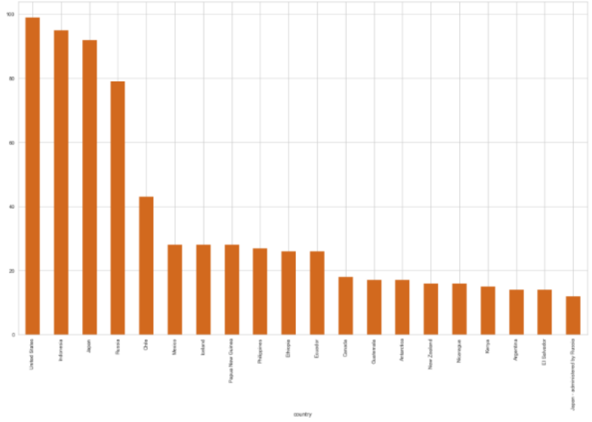
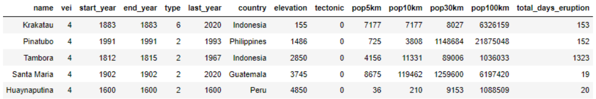


Figure 16. Volcanoes per country

If we take a look to the volcanoes with largest VEI, there are 5 and located in Indonesia, Philippines, Guatemala and Perú. The very large eruptions occurred mostly during the XIX and XXth centuries. They are part of the most noteworthy volcanoes of the world [14].



name	vei	start_year	end_year	type	last_year	country	elevation	tectonic	pop5km	pop10km	pop30km	pop100km	total_days_eruption
Krakatau	4	1883	1883	6	2020	Indonesia	155	0	7177	7177	8027	6326159	153
Pinatubo	4	1991	1991	2	1993	Philippines	1486	0	725	3808	1148684	21875048	152
Tambora	4	1812	1815	2	1967	Indonesia	2850	0	4156	11331	89006	1036033	1323
Santa Maria	4	1902	1902	2	2020	Guatemala	3745	0	8675	119462	1259600	6197420	19
Huaynaputina	4	1600	1600	2	1600	Peru	4850	0	36	210	9153	1088509	20

Figure 17. Volcanoes with largest VEI

The primary type of the volcano is Composite. Stratovolcanoes have relatively steep sides and are more cone-shaped than shield volcanoes. They are formed from viscous, sticky lava that does not flow easily. The lava therefore builds up around the vent forming a volcano with steep sides. Stratovolcanoes are more likely to produce explosive eruptions due to gas building up in the viscous magma. The 5 volcanoes with largest VEI are all placed in the continental crust of the subduction zones, which are a spot where two of the planet's tectonic plates collide and one dives, or subducts, beneath the other. This tectonic process can produce some of the planet's most powerful earthquakes, tsunamis and volcanoes.

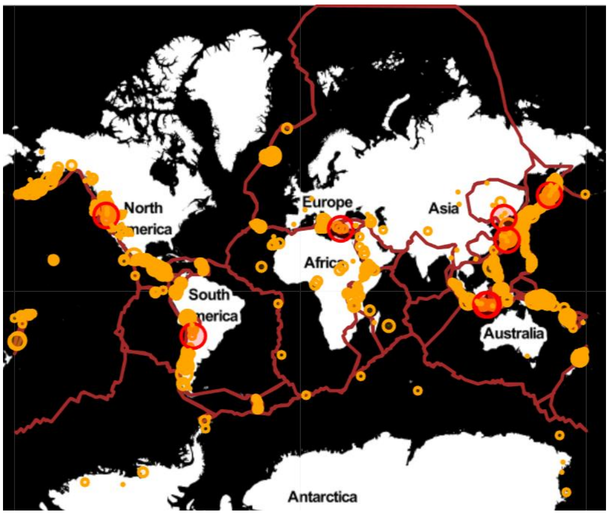


Figure 18. Eruptions and tectonic plates

Although the country with more volcanoes is EEUU, Indonesia is the one with more active volcanoes. The east Asian country has more than 500 eruptions of VEI 2, followed by Japan with more than 300 eruptions of also VEI 2.

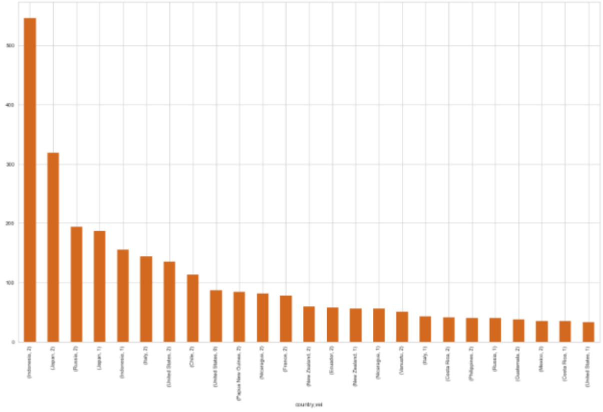


Figure 18. Countries with more eruptions

If we take a look at the most active volcano of all we find that it is Etna, in Italy. It has 145 eruptions in its history, most of them of VEI 2, followed by Pitón de la Fournaise, in Reunion Island. [15]

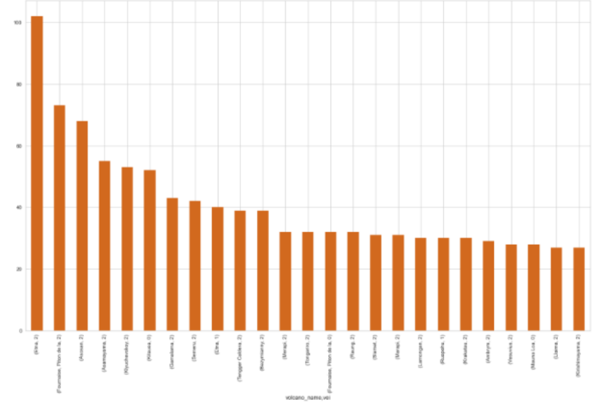


Figure 19. Volcanoes with more eruptions per VEI

Without any doubts, the most usual type of volcano is Stratovolcano, with almost 2500 eruptions coming from this type of volcano. Shield and caldera volcanoes are the second and third most usual types of volcanoes, but far away from Composite ones [16].

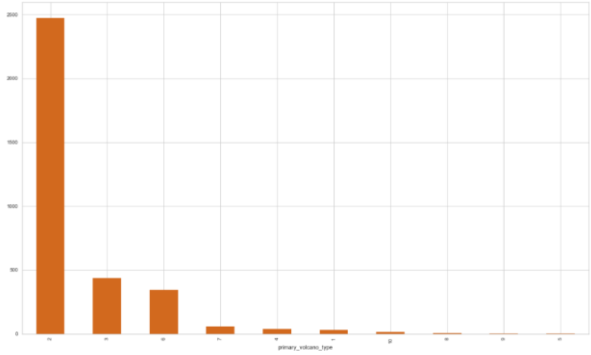


Figure 20. Eruptions per volcano type

Feature importance

Gini Importance or Mean Decrease in Impurity (MDI) calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits. [17]

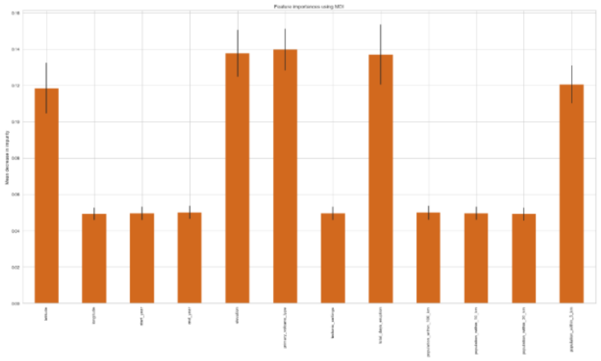
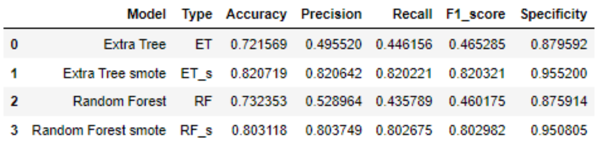


Figure 21. MDI feature importance

We can see in the graph above that the most important features in our prediction's model are elevation, primary volcano type and total days of eruption.

Results of the metrics

With the Borderline SMOTE technique, we improved our F1_score (which is a mix between the fraction of TP examples among the examples that the model classified as positive and the fraction of examples classified as positive, among the total number of positive examples) from 0'465 to 0'82 in our Extra Trees classifier.



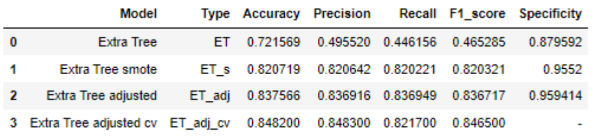
	Model	Type	Accuracy	Precision	Recall	F1_score	Specificity
0	Extra Tree	ET	0.721569	0.495520	0.446156	0.465285	0.879592
1	Extra Tree smote	ET_s	0.820719	0.820642	0.820221	0.820321	0.955200
2	Random Forest	RF	0.732353	0.528964	0.435789	0.460175	0.875914
3	Random Forest smote	RF_s	0.803118	0.803749	0.802675	0.802982	0.950805

Figure 22. Results of the metrics in train/test approach

Hyperparameter tunning and cross validation

Applying hyperparameter tunning [18] we adjusted the accepted parameters of our model to improve a little bit more our metrics and, thus, the prediction performance of our model.

After the adjustment of the parameters, we tried the cross-validation approach [19] and we were able to improve a little bit more the metrics. We went from 0'82 to 0'8367 and, finally, to 0'8465.



	Model	Type	Accuracy	Precision	Recall	F1_score	Specificity
0	Extra Tree	ET	0.721569	0.495520	0.446156	0.465285	0.879592
1	Extra Tree smote	ET_s	0.820719	0.820642	0.820221	0.820321	0.9552
2	Extra Tree adjusted	ET_adj	0.837566	0.836916	0.836949	0.836717	0.959414
3	Extra Tree adjusted cv	ET_adj_cv	0.848200	0.848300	0.821700	0.846500	-

Figure 23. Results of the metrics after adjustment and with cros s-validation approach

IV. Discussion

Regarding the supervised learning classification, we could try to scale the features without outliers with another scaler such as MinMax Scaler or Standard Scaler to see if there is any improvement.

There is another interesting possibility which is to implement unsupervised learning clustering instead of the classification. We could get other insights on the data that could help us in our analysis.

Using web scrapping, maybe we could get more interesting features to add to our dataset that probably helped us in the building of the model. Of course, to have a much more accurate predictive model, we should take in account many other geographical and climatological variables such as the slope of the volcano, the winds that can carry ashes around very large areas, if there is possibility of snow because it can mix with the ashes and create volcanic mudflows… This is an investigation line to be continued.

Vulcanology is a very difficult science and still with huge gap of improvement. This paper has no means to replace the magnificent job that vulcanologists do around the world. It needs to be taken as a will to find some patterns in the eruptions that could help the scientists in the prevention of potential catastrophes.

V. Some curious data

1. Put simply, a volcano is an opening in the Earth's surface.
2. Volcanoes are often found at meeting points of tectonic plates.
3. Volcanoes can also occur over mantle plumes. They're super-hot areas of rock inside the Earth!
4. Approximately 350 million people live within "danger range" of an active volcano. That means that around one in 20 people live in an area at risk of volcanic activity.
5. Volcanoes are classified as active, dormant or extinct.
6. Magma and lava are two different things. Magma is the name given to hot liquid rock inside a volcano. Once it leaves the volcano, it's known as lava.
7. Lava from a volcano can reach 1,250°C!
8. The world's largest active volcano is Mauna Loa in Hawaii.
9. Volcanoes exist throughout the solar system!
10. The loudest sound in recorded history was made by a volcano called Krakatau.
11. Volcanoes can produce rich, fertile land. [20]

VI. References

[1] Wikipedia, Rampino, M R; Self, S; Stothers, R B (May 1988). "Volcanic Winters". Annual Review of Earth and Planetary Sciences. 16 (1): 73–99

[2] Global Volcanism Program, 2013. Volcanoes of the World, v. 4.11.2 (02 Sep 2022). Venzke, E

(ed.). Smithsonian Institution. Downloaded 01 Oct 2022.

[3] URL: https://pandas.pydata.org/docs/user_guide/index.html

[4] *Volcanic Explosivity Index (VEI) –* URL: https://geology.com/stories/13/volcanic-explosivity-index/

[5] *Smithsonian Institution: volcanoes with longest eruptions* URL: https://volcano.si.edu/faq/index.cfm?question=longesteruptions

[6] URL: https://matplotlib.org/stable/tutorials/index

[7] URL: https://plotly.com/python/renderers/

[8] URL: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[9] URL: https://pycaret.gitbook.io/docs/

[10] URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[11] URL: https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226

[12] URL: https://python-visualization.github.io/folium/modules.html#module-folium.map

[13] *Smithsonian Institution: countries with most volcanoes* *URL:* https://volcano.si.edu/faq/index.cfm?question=countries

[14] *Smithsonian Institution: most noteworthy volcanoes* URL: https://volcano.si.edu/faq/index.cfm?question=eov_noteworthy

[15] *Smithsonian Institution: active volcanoes* URL: https://volcano.si.edu/faq/index.cfm?question=activevolcanoes

[16] *National Park service, U.S Department of the Interior: Type of volcanoes* URL: https://www.nps.gov/subjects/volcanoes/types-of-volcanoes.htm

[17] URL: https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3

[18] URL: https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/

[19] URL: https://scikit-learn.org/stable/modules/cross_validation.html

[20] *National Geographic: Explosive volcano facts* URL: https://www.natgeokids.com/uk/discover/geography/physical-geography/volcano-facts/