Devon Myers

Dr. Patrick Juola

CPMA 582

20 April 2022

Predicting the Severity of a Volcanic Eruption

1. **Introduction**

Volcanoes are powerful land formations whose eruptions have the potential to be deadly. The volcanic explosivity index (VEI) is a logarithmic scale between 0 and 8 that describes the severity of an eruption based on magnitude and the intensity. Power of an eruption increases 100-fold for successive VEI values up to 2, and increases tenfold for successive VEI values greater than 2 [[1]](https://www.nps.gov/subjects/volcanoes/volcanic-explosivity-index.htm). Being able to predict VEI of an eruption could save lives and mitigate damage from an eruption. This would also be cool because volcanoes are cool!

The primary goal of this study was to attempt to construct a model that can predict VEI of a volcanic eruption. Secondary goals are to determine if imputation can improve model performance and if oversampling can help alleviate the adverse effects of an imbalanced dataset.

1. **Background**

The outermost layer of the Earth is made up of large pieces of rock called tectonic plates. These plates float on molten rock, or magma in the Earth’s interior. Inevitably, tectonic plates bump and rub against one another, creating many geologic formations [[2]](https://www.nationalgeographic.org/encyclopedia/plate-tectonics/). One such formation is the volcano. A volcano is a land formation from which magma and gas erupts from the Earth’s interior onto the surface of the Earth. Volcanoes have many shapes and sizes, but all have a central vent, or group of vents, from which magma can flow to Earth’s surface.

Volcanoes can form nearly anywhere on Earth - on land, where two plates meet, or in the ocean, where plates are spreading away from one another [[3]](https://www.amnh.org/explore/ology/earth/power-of-plate-tectonics/volcanoes). The area where plates spread apart in the ocean is called a mid-ocean ridge. There are many types of volcanoes on Earth, but most fall into one of the following categories: stratovolcanoes, cinder cones, shield volcanoes, lava domes, submarine volcanoes, calderas, and complex volcanoes [[4,](https://volcano.si.edu/volcanolist_holocene.cfm) [5,](https://www.bgs.ac.uk/discovering-geology/earth-hazards/volcanoes/how-volcanoes-form/#:~:text=An%20eruption%20of%20highly%20viscous,volcano%20known%20as%20a%20stratovolcano) [6]](https://pubs.usgs.gov/gip/volc/types.html). Figure 1 contains counts of volcanoes of each type on Earth.

Submarine volcanoes are simply volcanoes in the ocean. They form along mid-ocean ridges and account for most of the volcanic activity on Earth [[3]](https://www.amnh.org/explore/ology/earth/power-of-plate-tectonics/volcanoes). Interestingly, ships sailing over an erupting submarine volcano can be sunk by the gas bubbles that escape from the volcano.

Stratovolcanoes, the most prevalent type of land volcano on Earth, are distinguishable due to their steep sides. These volcanoes generally have viscous lava that does not flow far from the volcanic vent [[5]](https://www.bgs.ac.uk/discovering-geology/earth-hazards/volcanoes/how-volcanoes-form/#:~:text=An%20eruption%20of%20highly%20viscous,volcano%20known%20as%20a%20stratovolcano). After many repeated eruptions and lava cooling, the peaks of stratovolcanoes can become very tall. Mount Fuji in Japan and Mount St. Helens in Washington State are examples of stratovolcanoes.

Cinder cones, also called pyroclastic cones, form around a central vent and normally have a crater at their summit. These volcanoes grow into a cone-shape when small rock fragments, called cinders, are ejected during an eruption and harden on the slopes of the volcano [[7]](https://www.nps.gov/articles/000/cinder-cones.htm). Mounts Etna in Italy and Talbert in Oregon state are examples of cinder volcanoes.

Shield volcanoes are shorter volcanoes that cover wide areas. These volcanoes have low-viscosity magma that flows very easily [[6]](https://pubs.usgs.gov/gip/volc/types.html). The Hawaiian Islands, namely Mauna Loa and Mauna Kea, are examples of shield volcanoes

Lava domes can form atop volcanoes with very viscous magma that have previously erupted. The viscosity of the magma prevents it from flowing far from the volcanic vent and instead, the magma builds up around the vent creating a dome over the vent once cooled [[8]](https://www.britannica.com/science/volcanic-dome).

Calderas are large bowl-shaped that form in one of two ways: first, when the summit of a stratovolcano is destroyed in a powerful eruption, or when an eruption empties out the majority of magma in the volcano and the top of the volcano collapses [[5]](https://www.bgs.ac.uk/discovering-geology/earth-hazards/volcanoes/how-volcanoes-form/#:~:text=An%20eruption%20of%20highly%20viscous,volcano%20known%20as%20a%20stratovolcano). The caldera atop Mount Pinatubo in the Philippines was created in the latter fashion.

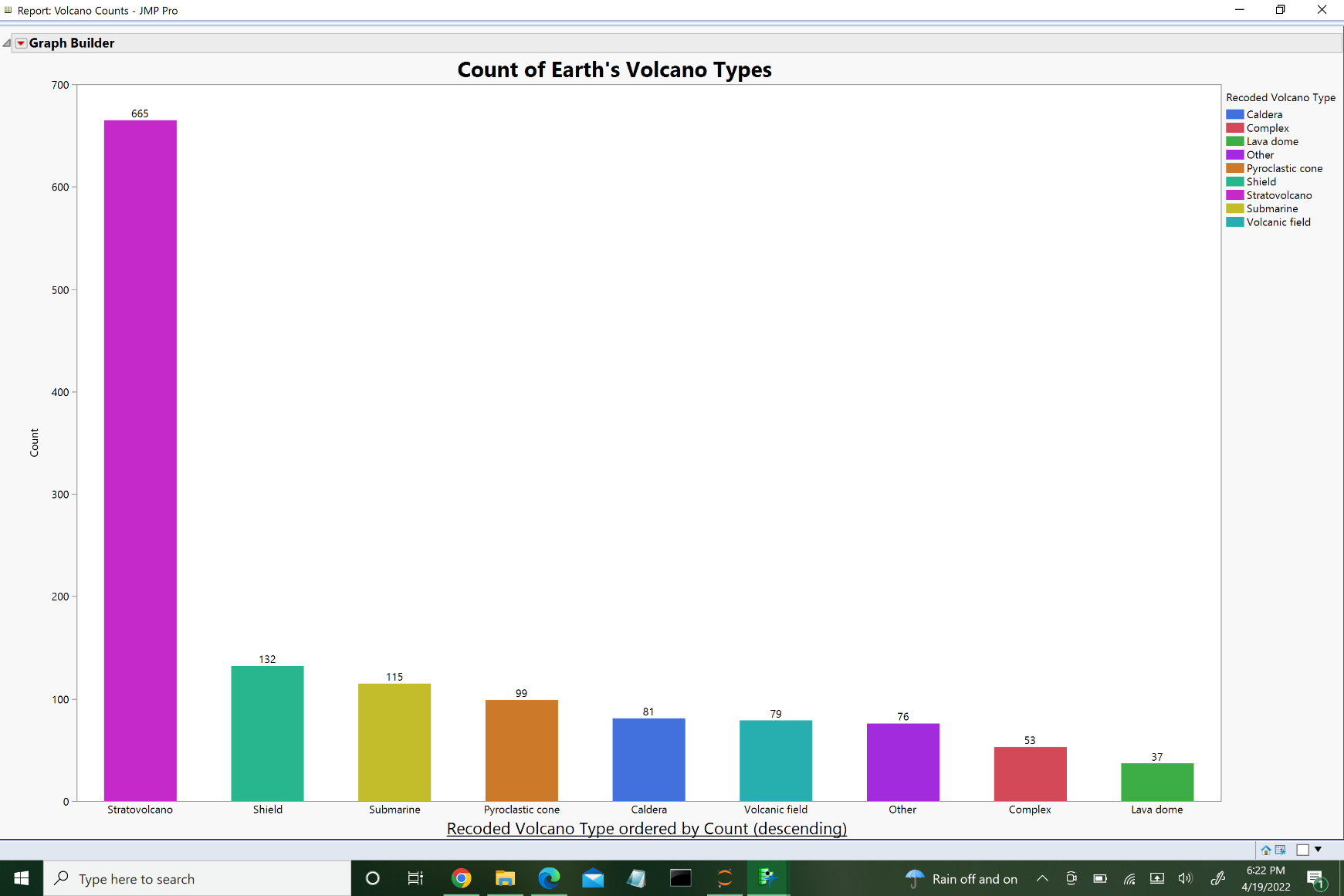
Some volcanoes consist of more that one vent connecting to a single magma chamber. These clustered volcanoes are referred to as complex volcanoes [[9]](https://volcano.oregonstate.edu/dieng-volcanic-complex). When these clusters of volcanoes are made up of mainly cinder cones, especially young cinder cones, the cluster is referred to as a volcanic field [[10]](https://www.britannica.com/science/volcanic-field.). The volcanoes at Yellowstone form a complex volcano.

Figure 1: Counts of each volcano type for Earth’s volcanoes.

Since volcanoes are not only cool, but also have deadly potential, the Smithsonian created a program to further research of volcanoes. This program, called the Global Volcanism Program (GVP), compiled, and continue to update datasets pertaining to information about the nearly 1,400 volcanoes on earth as well as volcanic eruptions.

1. **Materials and Methods**

Python 3 and Jupyter Notebooks were used to construct models capable of predicting VEI. JMP 14 was used for some preprocessing and visualization. Additionally, the Python libraries pandas and sklearn were used to store, manipulate, and preprocess the data.

**III.I Data Collection**

Two datasets were obtained from the GVP website. First, a dataset containing VEI information for 11,114 volcanic eruptions. Second, a dataset containing information about the 1,337 volcanoes on Earth such as volcano type, rock type of the magma, region, country, latitude, longitude, and elevation. Both datasets also contained names and identification numbers for each volcano. These two datasets were merged to form one dataset with 11,114 eruptions with information about the volcano that erupted. Table 1 indicates the type of each of the variables that will be used for analysis. It is also worth noting that there were no eruption observations in the dataset with a VEI level of 8.

JMP was used to recode volcano type into the nine categories in Figure 1. This was done to combine some uncommon volcano types, and to correct any misspellings. JMP was also used to analyze the distributions of features in the dataset.

|  |  |  |
| --- | --- | --- |
| **Ordinal Outcome Variable** | **Continuous Independent Variables** | **Nominal Independent Variables** |
| VEI | Latitude | Country |
|  | Volcano Type |
| Longitude | Region |
|  | Rock Type |
| Elevation | Tectonic Setting |

Table 1: Variables used in analysis.

**III.II Preprocessing**

In this dataset, as with many datasets, there was plenty of missing data. The Explore Missing Patterns functionality in JMP was used to identify observations with missing data. Out of 11,114 observations, 8,384 had complete records for all features. Three observations were missing tectonic setting, 31 missing dominant rock type, and 2,831 missing VEI. Note that the observations missing rock type were also missing VEI and the observations missing tectonic setting were missing both rock type and VEI. A copy of the merged dataset was made that accounted for missing data through k-nearest neighbors imputation.

Continuous independent variables were scaled using sklearns Min-Max scaler. This transforms each of the three independent variables to a value between 0.0 and 1.0. This accounts for negative values, which can affect some classification models, and helps to mitigate the effect of any potential outliers (of which there were none detected using JMP’s Quantile Range Outlier Detection feature). Additionally, the nominal independent variables were dichotomized using dummy variables. Data was then split into validation and training sets, where validation was 30% of the total dataset.

The last thing to account for before analysis is the imbalance in the outcome variable. Notice the vast difference in frequencies of different VEI levels in Figure 2. An imbalanced dataset is one where a nominal outcome variable has a large difference in frequencies for each outcome class. This can cause a classifier to misclassify outcome classes that have smaller frequencies [[11]](https://d1wqtxts1xzle7.cloudfront.net/40045253/0c960517fefa59fa6b000000.pdf20151115-68247-1ohdc8j-with-cover-page-v2.pdf?Expires=1650467336&Signature=NCVkA-iSBei7lbbtuxv7SOS8~HU0hEpSAxujGLrH0HmZr2810hyynZeNMo9EwI3iLGQ~JM7ZDURI6ZyvRZDLNR1XGJyFRC2sP6H5QiqiijbwXrA2453wB6Egrti3-oNHqR6zwslvI2Ic8WWg~TA4arZJzlHzV6MpX60FblVBkKFCvvQP~IMNgx~Ynbn9opxlUsW-5GA2cTuin1jJBS6VcHMG0xFFxuI0sNIJkz9Ids4k477m9eVzP3JJxK69X3JlsvxPgtrC-9YmSrhYa~pEIcnJgP~BV~s3jAaTlrPevhVWt0ygu2AaTdaDmf5tTLCEfeL8C58RMiZ5hgijd2DcKQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA). VEI level of 2 has most of the observations, so we will refer to this as the majority class. There are many ways to handle imbalanced data. In this experiment, a simple oversampling algorithm was implemented. To do this oversampling, we simply resample each of the minority classes with replacement until frequency of each minority VEI level match the frequency in the majority class. It is important to note that this almost surely reduces the variance of the model, since many of the datapoints will be replicated in the dataset.

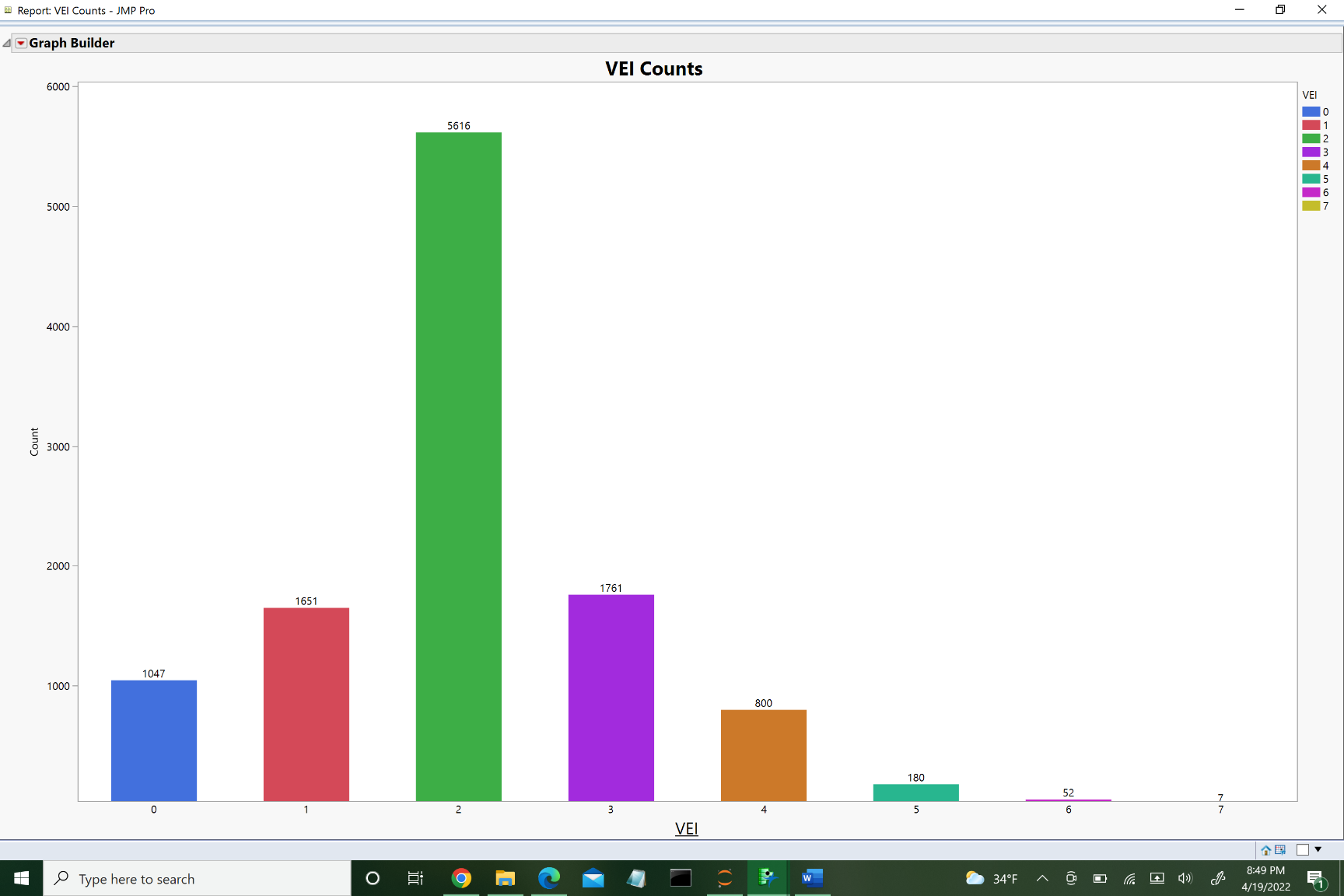


Figure 2: Frequency of each level of VEI.

**III.III Models and Performance**

To determine if imputation and oversampling increase model performance, four different datasets were constructed from the original merged dataset as follows: one with oversampling and imputation, one with oversampling and no imputation, another with imputation and no oversampling, and a final dataset with neither imputation nor oversampling. Performance metrics for models were subsequently compared for each of these datasets.

Six multiclass classification models were constructed for each of the four datasets. Each model employed one of the following classifiers: logistic regression, multinomial naïve bayes, k-nearest neighbors, decision trees, random forests, and gradient boosted trees. Metrics used to determine model performance were fit time, accuracy, and F1 score. Moreover, model performance metrics were estimated using 10-fold cross-validation. Summaries of each model can be found in table 2.

**III.IV Feature Selection**

Feature selection was performed by manually changing the independent variables present during analysis, and prediction accuracy was measured. This selection was done using the imputed, oversampled dataset, since this dataset has the highest model prediction accuracy (see Table 2(d)). Models with all eight of the independent variables had the highest accuracy, which indicates that all the variables should be used for analysis.

Moreover, to determine which variable is the best predictor of VEI, one independent variable at a time was removed from the study and accuracy was recorded. Note that the gradient boosted trees seemed robust to the changing of independent variables present – the accuracy of this model only decreased slightly when variables were removed. Interestingly, elevation and country seem to be the variable with the greatest predictive capabilities as the removal of each of these variables caused the accuracy of all six models to drop by at least 5%, whereas the other independent variables decrease model accuracy by less than 3%.

1. **Results**

Figure 3 shows VEI by average elevation in meters. Figures 4 and 5 show average

VEI by country and volcano type, respectively.

Tables 2 (a) – (d) contain summaries of model performance. From table 2(d), we see that gradient boosted trees, though by far the slowest of any of the models, has the highest F1 and accuracy scores

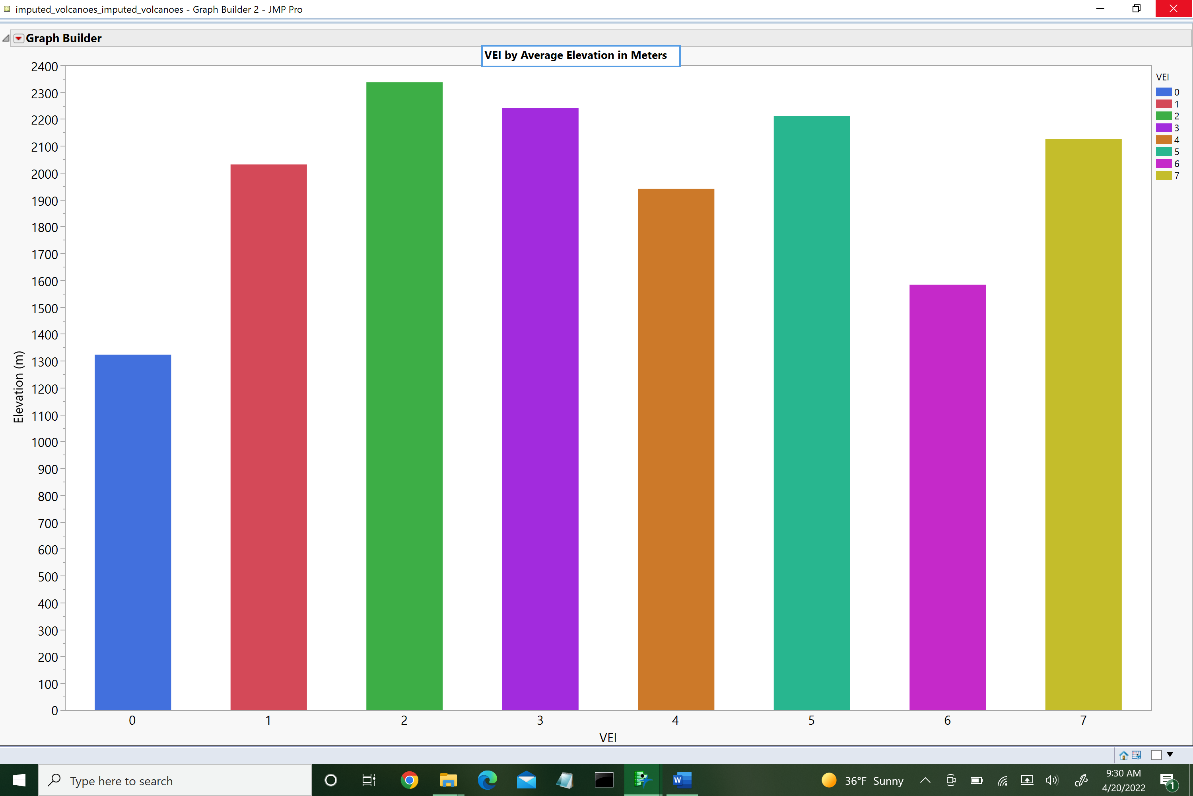


Figure 3: Average elevation in meters for each level of VEI

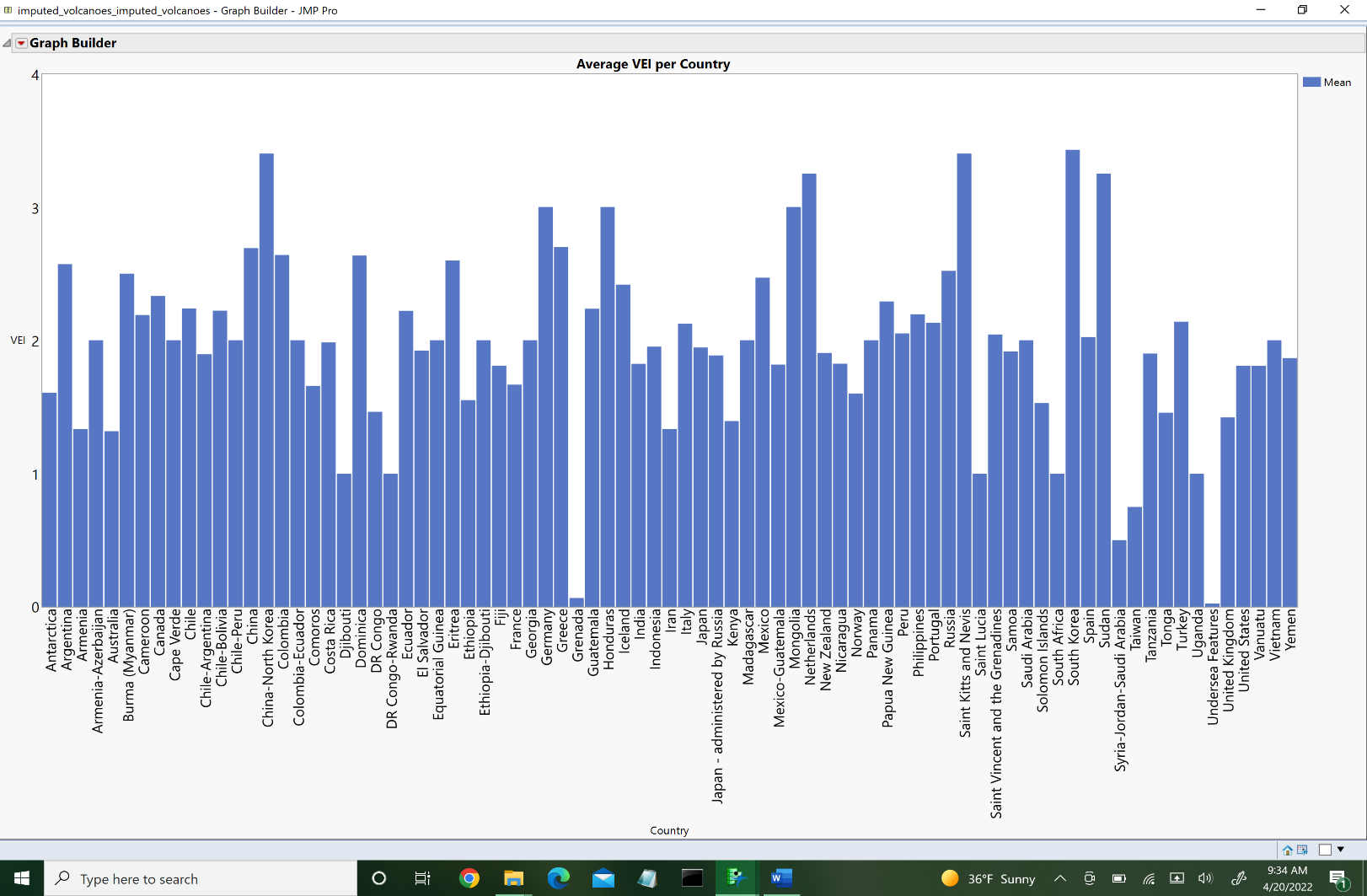


Figure 4: Average VEI per country

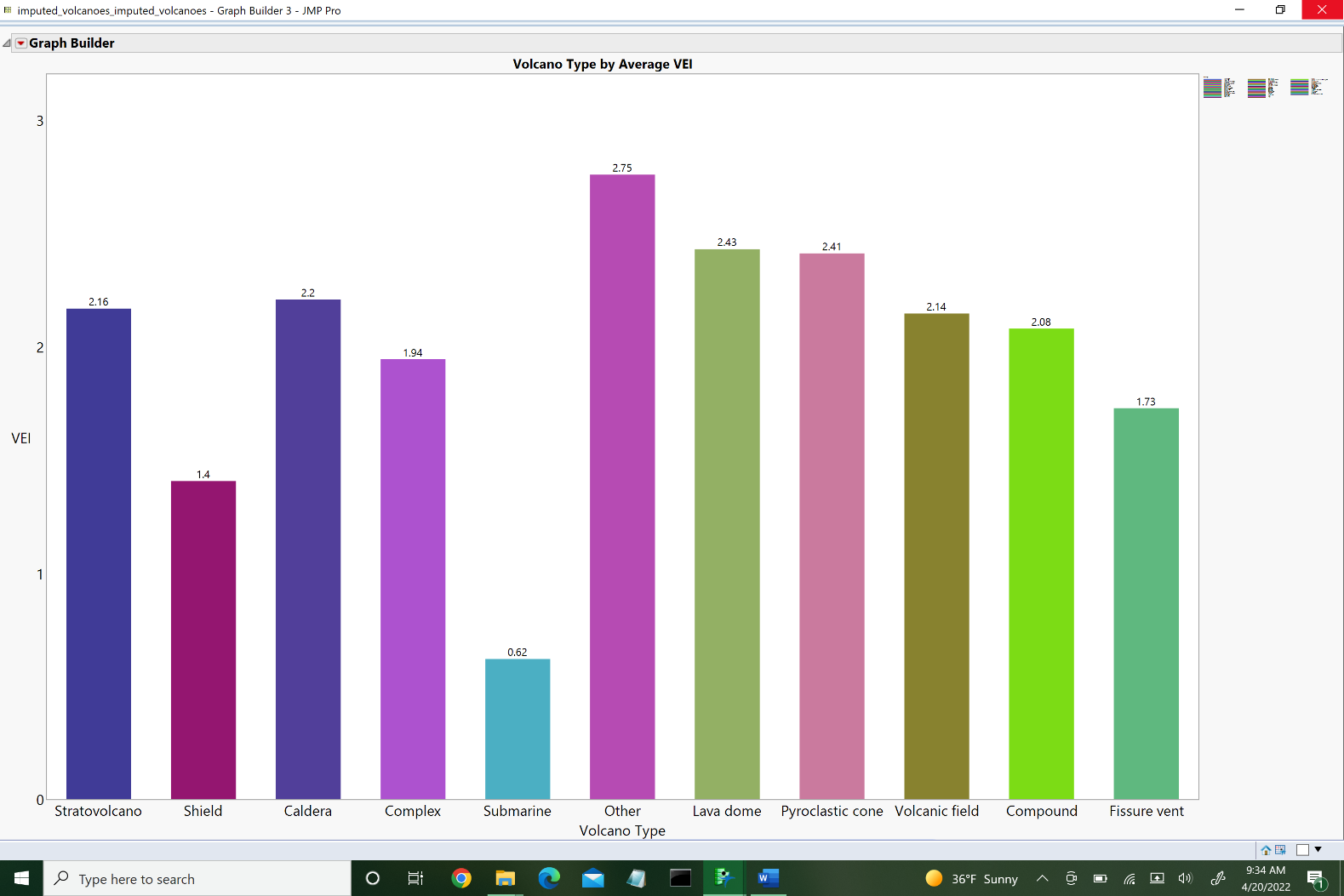


Figure 5: Average VEI per Volcano Type

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Logistic Regression | Naïve Bayes | k-Nearest Neighbors | Decision Trees | Random Forests | Gradient Boosted Trees |
| Fit time (s) | 1.465 | 0.025 | 0.015 | 0.065 | 8.891 | 67.00 |
| F1 | 0.443 | 0.436 | 0.518 | 0.537 | 0.511 | 0.544 |
| Accuracy | 0.543 | 0.517 | 0.563 | 0.571 | 0.577 | 0.579 |

Table 2(a): Model performance metrics for non-oversampled, non-imputed training data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Logistic Regression | Naïve Bayes | k-Nearest Neighbors | Decision Trees | Random Forests | Gradient Boosted Trees |
| Fit time (s) | 2.966 | 0.045 | 0.028 | 0.183 | 19.83 | 138.0 |
| F1 | 0.467 | 0.456 | 0.589 | 0.608 | 0.577 | 0.618 |
| Accuracy | 0.552 | 0.541 | 0.616 | 0.634 | 0.628 | 0.647 |

Table 2(b): Model performance metrics for non-oversampled, imputed training data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Logistic Regression | Naïve Bayes | k-Nearest Neighbors | Decision Trees | Random Forests | Gradient Boosted Trees |
| Fit time (s) | 11.52 | 0.080 | 0.050 | 0.343 | 40.99 | 280.6 |
| F1 | 0.486 | 0.416 | 0.619 | 0.642 | 0.642 | 0.649 |
| Accuracy | 0.506 | 0.445 | 0.628 | 0.654 | 0.654 | 0.660 |

Table 2(c): Model performance metrics for oversampled, non-imputed training data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Logistic Regression | Naïve Bayes | k-Nearest Neighbors | Decision Trees | Random Forests | Gradient Boosted Trees |
| Fit time (s) | 14.64 | 0.065 | 0.054 | 0.288 | 50.60 | 573.5 |
| F1 | 0.515 | 0.443 | 0.673 | 0.700 | 0.697 | 0.702 |
| Accuracy | 0.533 | 0.460 | 0.680 | 0.706 | 0.704 | 0.707 |

Table 2(d): Model performance metrics for oversampled, imputed training data

Since gradient boosted trees run on the oversampled, imputed data has the highest F1 and accuracy scores, validation was run using this model. Table 3 contains the confusion matrix for this validation. Values along the main diagonal (boldface values) are correctly classified observations. All other observations are misclassified. Note that the misclassification rate on the validation set is .

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Actual | Predicted | | | | | | | | |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 0  1  2  3  4  5  6  7 | **312** | 14 | 19 | 17 | 19 | 5 | 8 | 0 |
| 45 | **240** | 119 | 61 | 26 | 28 | 21 | 1 |
| 109 | 249 | **818** | 215 | 75 | 105 | 49 | 5 |
| 18 | 43 | 95 | **200** | 43 | 39 | 8 | 3 |
| 7 | 16 | 27 | 43 | **111** | 35 | 19 | 1 |
| 1 | 1 | 4 | 8 | 8 | **25** | 1 | 0 |
| 0 | 1 | 4 | 1 | 3 | 4 | **4** | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | **0** |

Table 3: Confusion matrix for gradient boosted tree VEI prediction validation.

1. **Discussion**

Figures 3 and 4 display VEI by average elevation and country, respectively. One can see from Figure 3 that a VEI level of 0 or 6, on average, occur when volcanoes of comparatively stouter volcanoes. Moreover, VEI level of 2 has the highest average elevation. From Figure 5, we see that stratovolcanoes, which are generally taller volcanoes, have an average VEI of just short of 2.2, so the average height of volcanoes with VEI-2 eruptions seems to be dragged upward by the heights of stratovolcanoes. Additionally, the category with the largest VEI is the “other” category. This category is the ‘catch-all’ group where non-prevalent volcano types were grouped together. Since the VEI is greatest for this group, it may be worth studying these volcanoes in greater detail instead of grouping them all in one.

One can also see from Figure 4 that the countries with the greatest average VEI eruptions are the China, North and South Korea, the Netherlands, Saint Kitts and Nevis (a Caribbean country), and Sudan. On the other hand, Grenada (another Caribbean country) and ‘undersea features’ have very low average VEI. This ‘undersea features’ refers to submarine volcanoes. It is interesting that there is such a spread in average VEI of eruptions just in the Caribbean area.

From Tables 2 (a) and (c) we see that oversampling makes the fit time increase for each of the models. This is because oversampling increases the size of the dataset. Similarly, oversampling seems to increase the test set accuracy and F1 score for each model, except for naïve bayes, which drops from 0.517 to 0.445 and from 0.436 to 0.416, respectively. Comparing the results in Tables 2 (a) and (b), we see that imputation increases fit time, F1 score, and accuracy for each of model.

Comparing Tables 2 (a) and (d), we see that oversampling and imputation increase all three metrics for each of the models, except for naïve bayes again. In fact, decision trees, random forest, and gradient boosted tree models all have nearly equal F1 and accuracy scores on the training set. The fit time, however, increases greatly from decision trees to random forest, and random forest to gradient boosted models. The k-nearest neighbors model also has comparable accuracy and F1 scores but is still outperformed by the three aforementioned models.

Logistic regression and naïve bayes underperform compared to each of the four remaining models for each set of training data. Interestingly enough, naïve bayes performance on the testing data is only increased by imputation and is decreased when oversampling is introduced.

Though gradient boosted trees have the best performance on the training data, it is clear from the confusion matrix in Table 3 that this model does not perform as well as on the validation data as it did on the training data. This could indicate that the model is overfit, or that the imputation and oversampling muddied the generalizability of the model, perhaps both.

1. **Conclusion and Future Work**

The purpose of this study was to determine if a model could be constructed to predict VEI of volcanic eruptions. Though we can build many models capable of doing this prediction, it seems like more work is required to get a model with better generalizability. Overfitting and reduced variance from the simple oversampling technique seem to have a large impact on the results of this experiment.

A suggestion for future work to increase the quality of this study would be to try different ways of handling the imbalance in the dataset. A popular oversampling technique, called SMOTE, samples from the feature space to generate synthetic observations [[12]](https://doi.org/10.1613/jair.953). Though these are not true observations, this oversampling procedure could possibly not lower the variance in the dataset as much as the oversampling algorithm used in this study.

An additional suggestion would be to further study the effect of imputation on VEI, and to test multiple forms of imputation. For instance, perhaps it is the case that mean imputation, or multivariate SVD imputation, can increase the predictive capabilities of the model.

Lastly, it would be interesting to try additional modeling techniques, such as artificial neural networks, to attempt to train a better model. It may be that with the right size of network and hidden layers that one can get a model that very accurately predicts VEI based on input features such as country and elevation.

**References**

[1] “Volcanic Explosivity Index.” *National Parks Service*, U.S. Department of the Interior, <https://www.nps.gov/subjects/volcanoes/volcanic-explosivity-index.htm>.

[2] “Plate Tectonics.” *National Geographic Society*, 2 Apr. 2020, <https://www.nationalgeographic.org/encyclopedia/plate-tectonics/>.

[3] “Power of Plate Tectonics: Volcanoes: AMNH.” *American Museum of Natural History*, <https://www.amnh.org/explore/ology/earth/power-of-plate-tectonics/volcanoes>.

[4] “Global Volcanism Program: Holocene Volcano List.” *Global Volcanism Program | Holocene Volcano List*, <https://volcano.si.edu/volcanolist_holocene.cfm>.

[5] “Types of Volcano.” *British Geological Survey*, 23 June 2021, <https://www.bgs.ac.uk/discovering-geology/earth-hazards/volcanoes/how-volcanoes-form/#:~:text=An%20eruption%20of%20highly%20viscous,volcano%20known%20as%20a%20stratovolcano>.

[6] Watson, John. “Principal Types of Volcanoes.” *Volcanoes: Principal Types of Volcanoes*, <https://pubs.usgs.gov/gip/volc/types.html>.

[7] “Cinder Cones (U.S. National Park Service).” *National Parks Service*, U.S. Department of the Interior, <https://www.nps.gov/articles/000/cinder-cones.htm>.

[8] “Volcanic Dome.” *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., <https://www.britannica.com/science/volcanic-dome>.

[9] “Dieng Volcanic Complex.” *Volcano World*, 18 Dec. 2018, <https://volcano.oregonstate.edu/dieng-volcanic-complex>.

[10] “Volcanic Field.” *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., <https://www.britannica.com/science/volcanic-field>.

[11] Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. ["Handling imbalanced datasets: A review."](https://d1wqtxts1xzle7.cloudfront.net/40045253/0c960517fefa59fa6b000000.pdf20151115-68247-1ohdc8j-with-cover-page-v2.pdf?Expires=1650467336&Signature=NCVkA-iSBei7lbbtuxv7SOS8~HU0hEpSAxujGLrH0HmZr2810hyynZeNMo9EwI3iLGQ~JM7ZDURI6ZyvRZDLNR1XGJyFRC2sP6H5QiqiijbwXrA2453wB6Egrti3-oNHqR6zwslvI2Ic8WWg~TA4arZJzlHzV6MpX60FblVBkKFCvvQP~IMNgx~Ynbn9opxlUsW-5GA2cTuin1jJBS6VcHMG0xFFxuI0sNIJkz9Ids4k477m9eVzP3JJxK69X3JlsvxPgtrC-9YmSrhYa~pEIcnJgP~BV~s3jAaTlrPevhVWt0ygu2AaTdaDmf5tTLCEfeL8C58RMiZ5hgijd2DcKQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)*GESTS international transactions on computer science and engineering* 30.1 (2006): 25-36.

[12] Chawla, N. V., et al. “Smote: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321–357., <https://doi.org/10.1613/jair.953>.