Beyond CIPW: A Novel Approach to Igneous Rock Classification Using Statistical Modeling and Machine Learning Algorithms

Evan Mosseri

James Madison High School

Prepared for the 2014-2015 Intel Science Talent Search

Abstract:

There is a long and complex history to the need for the classification of igneous rocks. Approximately 200 years ago, the Saint Petersburg Academy of Sciences in Russia offered a prize for the best essay on the classification of rocks. Rock classification has undergone a plethora of expansions, reviews, and alterations. It is a vast subject. Like all modern sciences, it is evolving once again with the introduction of “Big Data” and new technologies.

The Cross, Iddings, Pirsson, and Washington (CIPW; 1965) norm, is regarded by most geologists as the gold standard for determining mineralogy on the basis of major element chemistry. It is well known that there are some issues with its accuracy, especially with regards to mantle rocks, such as peridotites, and exotic rocks, such as carbonatites and lamprophyres. With the advent of cyberinfrastructure for the geological sciences, there is an opportunity to reassess how igneous rocks are classified using commonly reported data such as chemical composition.

Through the use of the EarthChem petrological database, multidimensional analysis and machine learning algorithms were performed on mantle rock data. Classifications resulting in accuracies ranging from 70-85% accuracy depending on rock type and number of samples in the dataset were obtained. In addition, some of the qualitative findings may aid geologists in understanding relationships between igneous rocks. This program will be available online as an open source project for geologists to use for classification and analysis of peridotitic rocks. Future research will include the improvement of the current classifier and the creation of classifiers for different types of rocks. This research has the potential to evolve into an entirely new standard for the classification for igneous rocks.

Background

Approximately 200 years ago -- the Saint Petersburg Academy of Sciences in Russia offered a prize for the best essay on the classification of rocks. Kirwan in 1794 coined the phrase “igneous rock”. “Today, the study of igneous rocks is a vast subject, and the task is still to create a systematic and sustainable classification of the many different types now recognized.” (M.J. Le Bas p.825)

M.J. Le Bas & A.L. Streckeisen (1991) investigated the principles of igneous rock classification using the International Union of Geological Sciences (IUGS) subcommission’s 10 principles for constructing a system of “classification and defining the appropriate nomenclature”. Prior to this work, early attempts to classify igneous rocks were based largely on petrography and mineralogy. Other authors utilized a classification system based on minerals and textures and other authors used mineral assemblages, magma origins, magma types, mineral assemblages, and other attributes. As this work continued, new names proliferated, particularly for alkaline rocks. Johannsen (1939) sought to systematize the growing nomenclature,and published ‘A descriptive petrography of the igneous rocks’ in four volumes (1932-1939), which came to dominate the thoughts of petrologists in the English-speaking world.” (Le Bas, p.825) At approximately the same time, Niggli (1931) presented a system for the classification and nomenclature of igneous rocks. He followed this (1936) with a system of classifying igneous rocks by their chemical compositions, which system was based on molecular numbers (‘Niggli numbers’), systematized as ‘magma types’. These ‘magma types’ were not rock names but objective attributes." This became the preferred method of identification in the German-speaking world. In the 1960s, as the internationalization of science continued, the need to create an all-inclusive, objective and uniform method for characterizing igneous rocks became imperative. In addition, with more than 1500 individual igneous rock names being used, there was a need to consolidate and reduce the number of individual rock names.

While much has been done to simplify and modernize the scheme for igneous rock classification, it has remained a labor intensive and somewhat subjective task based primarily on mineral identification, crystal structure and visual characteristics. Many other sciences have embraced quantitative classification and identification techniques although when it comes to petrology, the field has been slow in terms of embracing the emerging technologies such as machine learning and artificial intelligence. Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets” (Joseph A. Cruz and David S. Wishart)

Introduction

As igneous rock identification has become more uniform and quantified, it becomes more suitable for large-scale computer-based schemes for identification and analysis, making use of relatively new technologies including spectrographic analysis, remote sensing and machine learning algorithms. In addition to terrestrial applications, remote sensing, quantitative analysis when used alongside machine-learning processes are applicable for use on large-scale geological surveys both on Earth and even in extraterrestrial applications.

Examples of rock classification schemes

M.J. le Bas and A. L. Streckeisen expressed their classification and nomenclature of igneous rocks with a set of formulas that can be represented as ternary diagrams. These diagrams are based on the proportion of concentration of any n number of minerals in a sample. Each additional substance added to the classification adds an extra dimension to these diagrams.

> "Second, to utilize the large computerized database CLAIR (Le Maitre & Ferguson, 1978) which contains not only chemical analyses of igneous rocks but also the rock-name of each analysis. The database could be used to plot each volcanic rock (e.g. all trachytes) on frequency distribution diagrams and so put best-fit boundaries between adjacent fields. Despite some overlap between adjacent frequency distribution plots, the clear clustering of the points for any one rock type indicated where boundaries should be placed according to generally accepted usage" (Le Bas p.831)

Pearson et al. (2008) reviewed the complexity of identifying mantle xenoliths and sought to simplify the classification of inclusions of this kind by grouping rocks by geochemical parameters. They write:

> "To simplify matters and to circumvent the petrographic complexities of alkaline volcanic rocks in general, we will use the term "alkalic and potassic mafic magmas" to include alkalic basalts, nephelinites, melilitites, and lamprophyres. Occurrence of xenoliths in such magmas can be compared to those occurring in kimberlites and related rocks. As a general rule, the spectrum of mantle xenoliths at a given location varies with host rock type. "

C.H. Kelsey laid out his own method used to classify igneous rocks based on their oxide components rather than their mineralogical components in his paper "Calculation of the C.I.P.W Norm", published in 1965. In the paper, Kelsey illustrates a set of steps compiled from both manual observation of igneous rocks and mathematical analysis based on the chemical formulas of the idealized representations of different types of rocks. Today, Kelsey's CIPW Norm remains the gold standard in classifying rocks through chemical composition. The problem with the CIPW norm is that it has been found to be inaccurate, especially when dealing with peridotitic rocks. More so, the CIPW norm converts oxidation data to mineralogical composition. Hence, the actual classification of these rocks is based on Streckeisen classification schemes of igneous rocks, which elicits more inaccuracies. The methodology

between the CIPW norm and Streckeisen's classification of igneous rocks leaves a lot of room for error particularly when dealing with more chemically variable rock types. Although much has been done to try to remedy the faults of this classification (see "Calculation of the CIPW norm: New formulas" by Kamal L Pruseth), it remains rather arbitrary, which renders it insufficient for the needs of contemporary geologists.

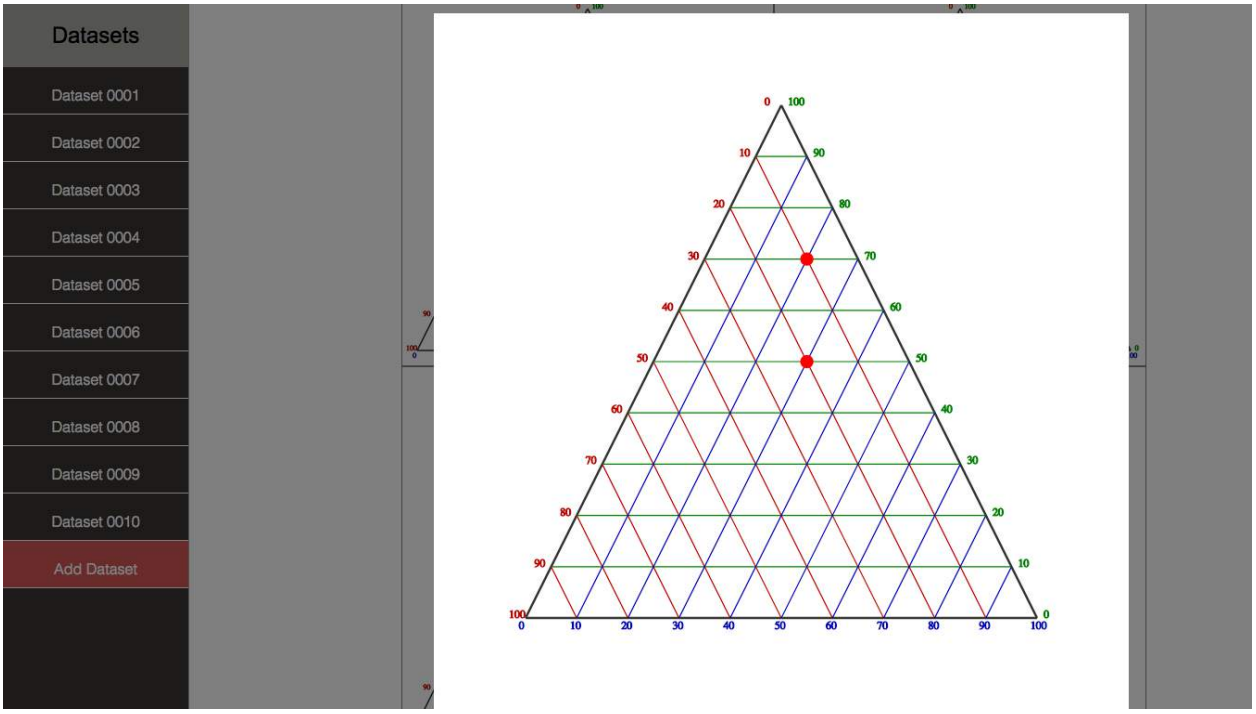
Geologists work with large datasets and collaborate with other scientists at an ever-increasing rate, and therefore would benefit from a more accurate and scalable classification.

> The U.S. National Science Foundation (NSF) places significant emphasis on computational and data-rich science and engineering, with the goal of providing a sustainable, community-based and open cyberinfrastructure for researchers and learners. This is a major challenge because the number and volume of data sets have grown to proportions well beyond the range of applicability of traditional data handling tools. Transformative approaches and innovative technologies are needed for heterogeneous data to be integrated, made interoperable, explored and re-purposed by researchers in disparate fields and for myriad uses across institutional, disciplinary, spatial and temporal boundaries. (NSF; 07/09/11)

In general, it can be stated that one of the primary difficulties faced by geologists and geochemists as substantiated in the literature is the complexity of igneous rock nomenclature and classification, necessitating the use of a standardized, quantitative based automated method of igneous rock identification. It is well established that it is difficult to classify igneous rocks due to the large number of rocks containing slight variations in composition and structure. There is a need for the development of automated tools and algorithms to rapidly and uniformly classify large datasets of igneous rock samples. For the purpose of this study, the investigation was limited to peridotites due to both of the large variation in rock/mineral characteristics and the fact that it is a subset of igneous rocks that are particularly hard to classify using traditional means and normalization methods. Future work will expand this approach to cover additional types of rocks.

Methodology

The initial work consisted of the development of visualization software for geologists. A GUI to facilitate use of the classification system in order to compare different samples based on their ternary plots was developed. A traditional ternary plot uses three dimensions represented by scales on the sides of the triangle.



Simple ternary diagram with two data points generated using GUI

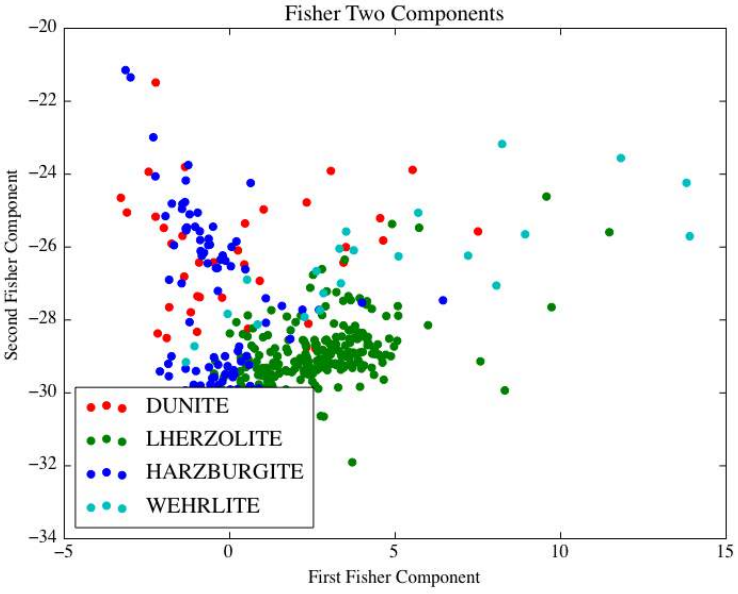
In addition, vector drawing with SVG and the D3 JavaScript library, and scaling algorithms were used alongside standard GUI development to produce functioning and interactive ternary diagrams that could be compared to one another both quantitatively and visually, and compared to various chemical compositions within an individual sample. A comprehensive web application was built on a Django backend, a python web framework, that featured a user management system, an automatic database parser, and the saving and sharing of analysis. The original parameters may contain 4 or more values. 3 distinct values may be normalized and compared to other combinations of these values on the same diagram. One requirement of the GUI is for it to be able to take chemical data and attempt to convert it to mineralogical composition data. This was attempted through a process known as CIPW normalization, the traditional method used to convert chemical composition data into mineralogical composition data that could be used to classify rocks using Streckheisen's classification scheme. This tool has been packaged with the classification software developed for this research, which follows, for use as a comparison and/or validation of a new classifier.

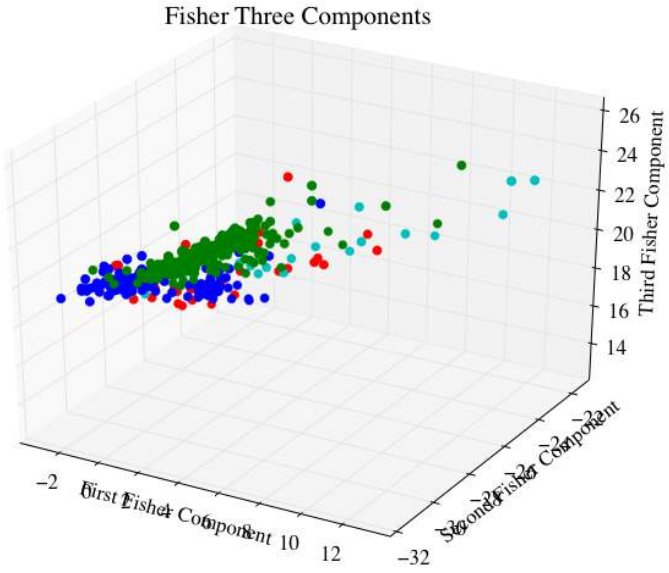
Due to some inherent weaknesses with CIPW, the need for an improved system for igneous rock identification is apparent. Therefore, the creation of a more accurate conversion of chemical composition data to mineralogical composition data and/or new, more robust classification techniques for these rocks

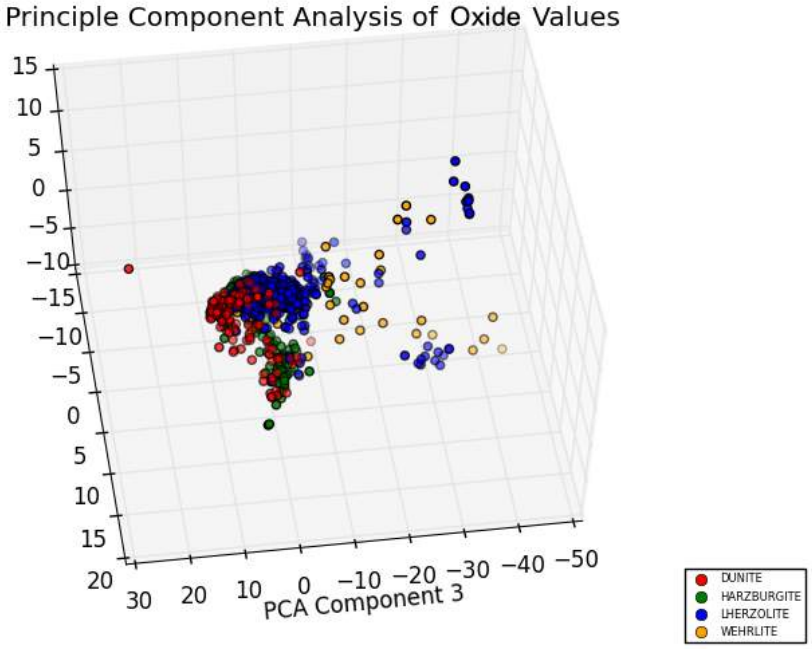
would be considered of paramount importance. This was the inspiration for the use of machine learning algorithms to create a new type of classification for igneous rocks based on their chemical content.

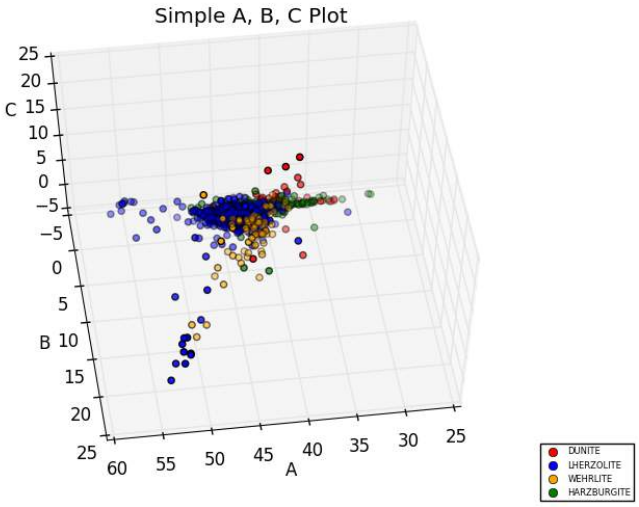
The first step required to create classification of igneous rocks consisted of performing a statistical analysis by applying machine learning analysis onto a database of peridotitic igneous rock compositions. A dataset downloaded from the NSF-sponsored Earthchem Portal was used as test case. The dataset consisted of data and metadata from geological samples including rock chemistry and geospatial coordinates. Most of the samples did not have complete chemical analyses available. As a result, out of the 1727 samples, only a small subset was available for classification based on all oxide components (466 samples).

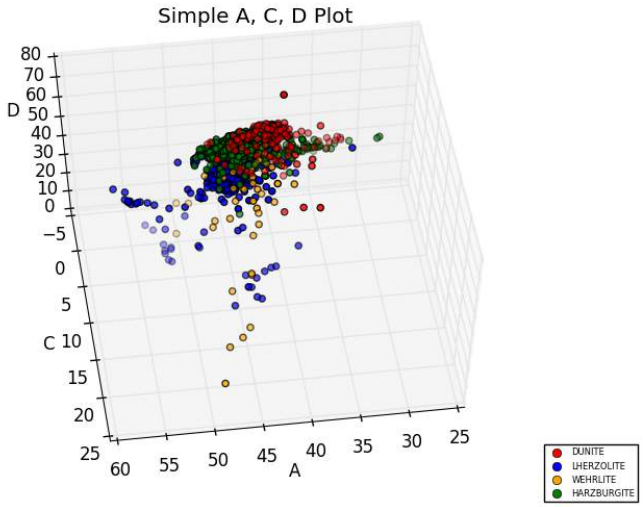
The initial step within the statistical analysis of the dataset was to take a "snapshot" of the dataset using both simple 2D component comparisons, Principle Component Analysis (PCA), and Fisher Discriminant Analysis (FDA) to analyze the dataset. PCA was selected to visualize the datasets used in this paper because it takes high dimensional data and visualizes it in fewer dimensions by identifying and averaging the most variable dimensions. The initial analysis identified four parameters (A, B, C, and D) which provided the more pronounced differences between different types of rocks. This work provided an insight into the appearance of a typical dataset, when evaluated and converted to graphic form, and served as a control for future analysis. The remaining eight parameters were plotted in the scatter matrices.

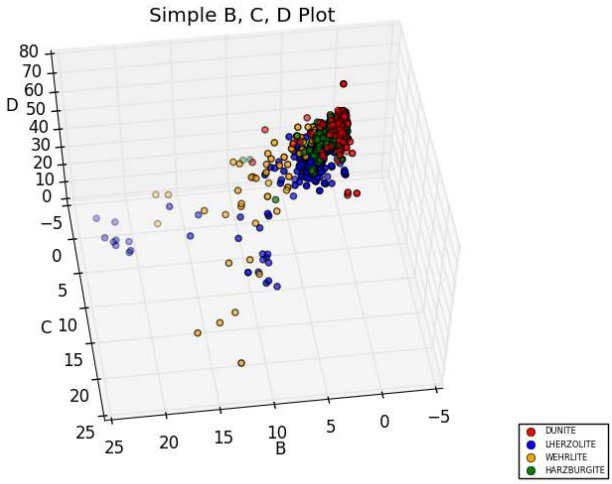


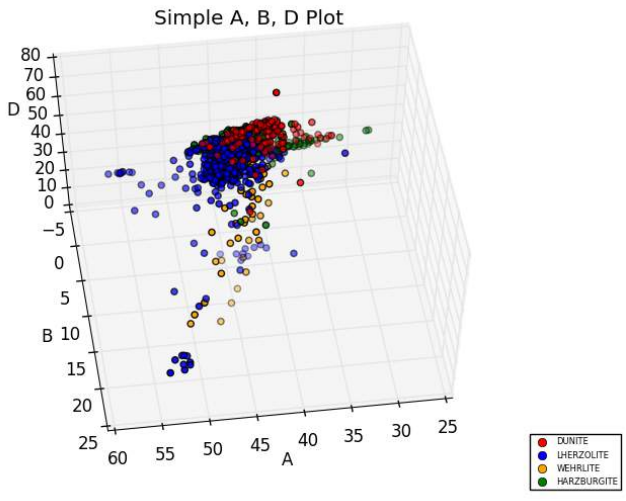



Principle Component Analysis of Oxide Values
PCA Component 3
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

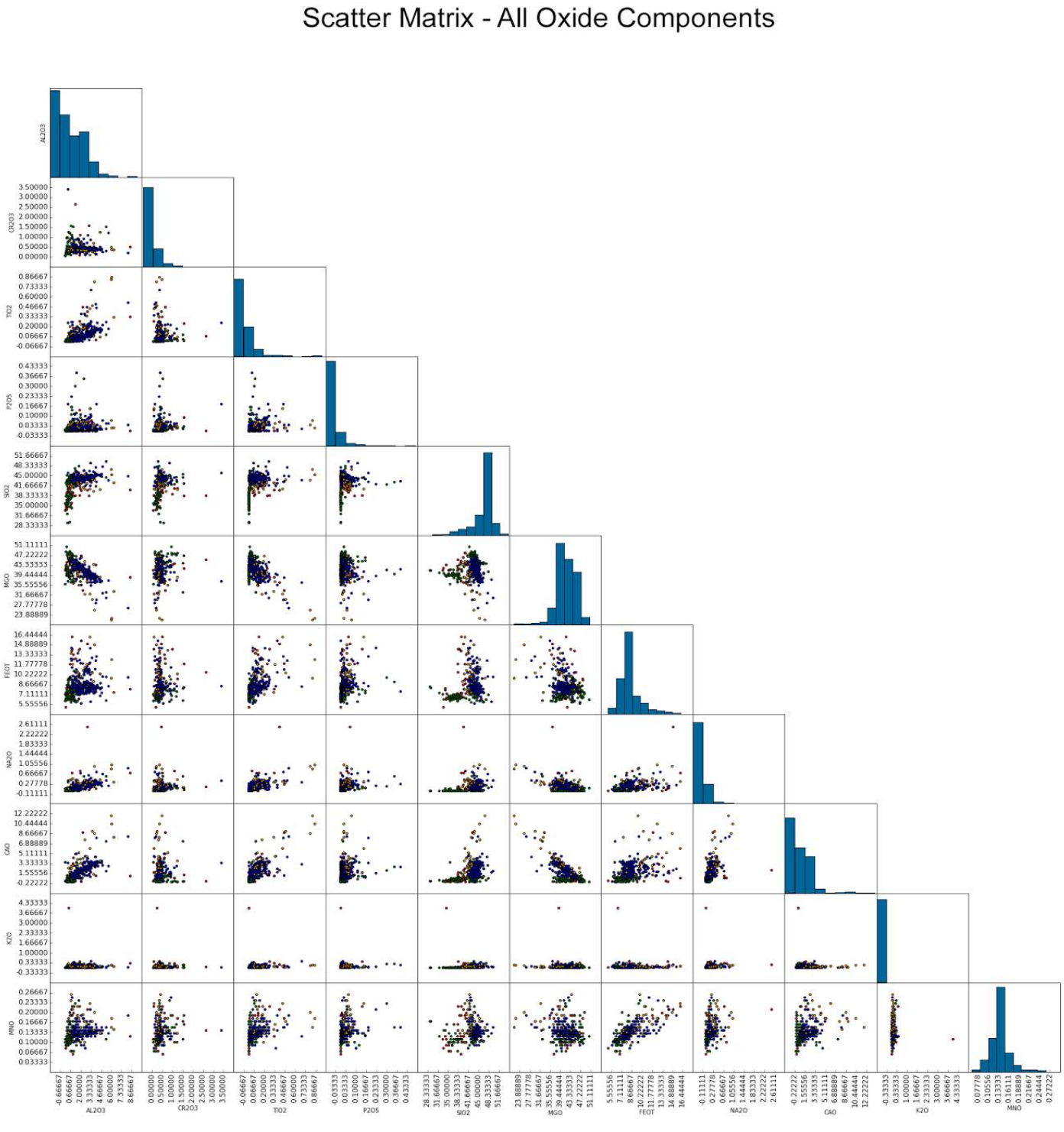

Simple A, B, C Plot
A
B
C
DUNITE
LHERZOLITE
WEHRLITE
HARZBURGITE


Simple A, C, D Plot
A
C
D
DUNITE
LHERZOLITE
WEHRLITE
HARZBURGITE


Simple B, C, D Plot
B
C
D
DUNITE
LHERZOLITE
WEHRLITE
HARZBURGITE

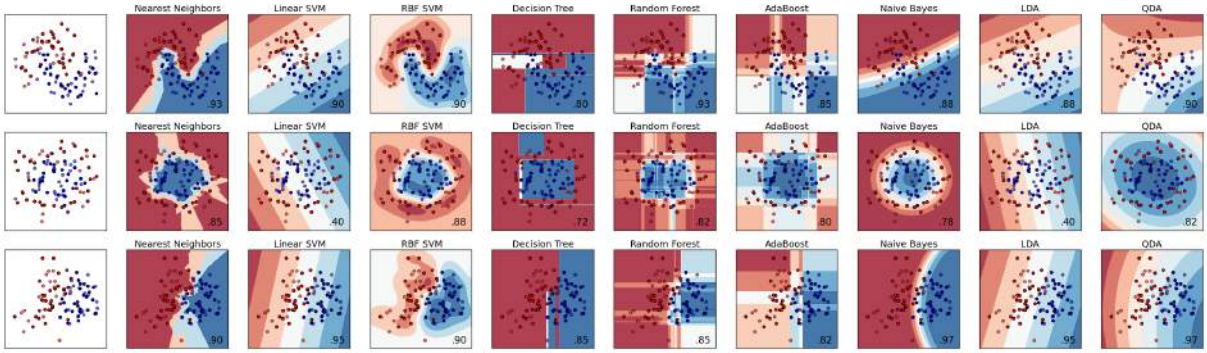

Simple A, B, D Plot
A
B
D
DUNITE
LHERZOLITE
WEHRLITE
HARZBURGITE

The following scatter matrix consists of individual 2D comparisons of different combinations of chemical components. This provides an insight into the degree in which the plots of different rock type representations overlap in a simple 2 dimensional space and therefore shows that the analysis can be easily represented in a few dimensions, which can be valuable in terms of feature selection and data preprocessing.

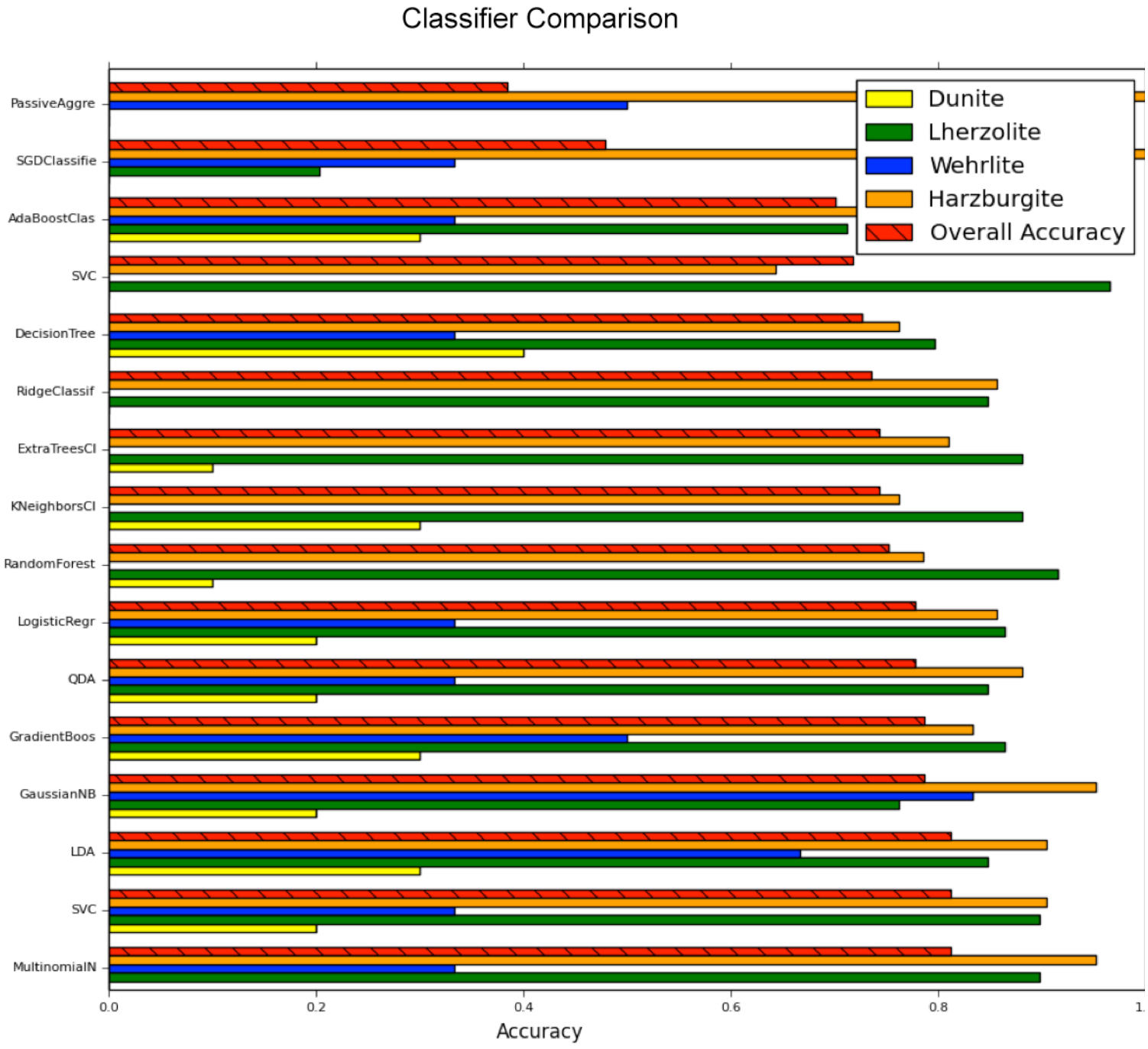


After a careful analysis of both the volatility and spread of the dataset (in reality, fairly overlapped, suggesting that it would be hard to classify without proper tuning of an appropriate algorithm), a classifier comparison was used in order to determine which classifier utilized the optimal algorithm for classifying the chemical composition data. Every classifier used data that was separated into both a “training set” which was analyzed in order to test the performance of classification of the “testing set”. While it would be optimal to shuffle the data and split it into two equal training and testing sets under normal conditions, the limited available dataset rendered it necessary to use K-Fold Cross Validation to representatively split a 75% subset of the data into a training set, and a 25% subset of the data into a testing set. K-Fold Cross validation does this without over fitting, meaning without the test being run on the training data which would be misleading in terms of its accuracy. A confusion matrix was generated for every classification, pinpointing the samples that the classifiers identified correctly and the samples that they identified incorrectly. This allowed for the fine tuning of the classifiers through the use of optimal parameters for the specific dataset. The analysis was accomplished using a combination of both the numpy, matplotlib, and scipy python libraries and the scikit-learn machine learning library. The actual algorithms used were: Naïve Bayes (Gaussian NB), Multinomial NB, SVC, SVC (Altered Parameters), K Nearest Neighbors, Decision Tree, Ada Boost, Random Forest, Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), Logistic Regression, Gradient Boosting, Extra Trees, Ridge, Passive Aggressive, SGD Classifier. These classifiers were compared using the 466 samples with full oxidation data.

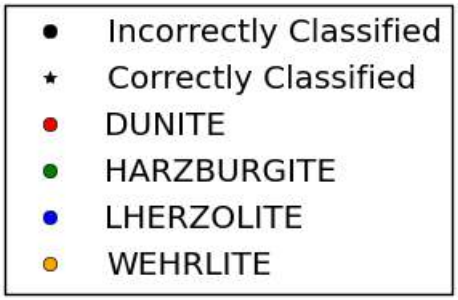
A comparison of a several classifiers in scikit-learn on synthetic datasets. The point of this example is to illustrate the nature of decision boundaries of different classifiers. The following diagram illustrates how different classifiers will separate the same data points. All of these classifiers were used in the analysis of the peridotitic tock data. ()



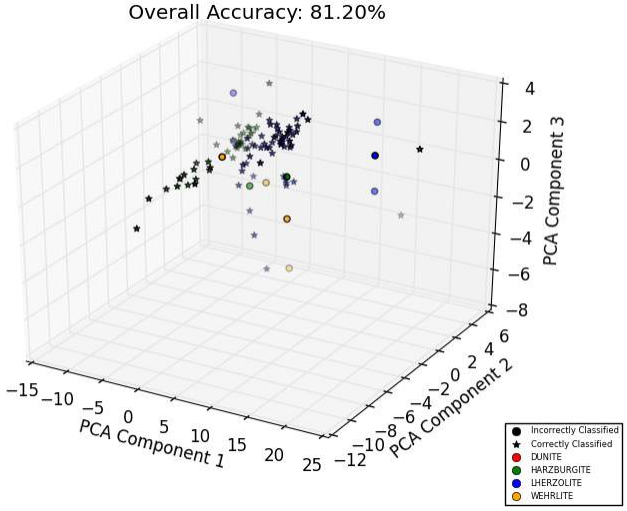
This graph appears on the following page. It visualizes the accuracy of the classifiers for different types of rocks and the simple accuracy of the classifier. It is sorted by simple accuracy from lowest to highest from top to bottom. All 16 of the classifiers are represented on the graph.



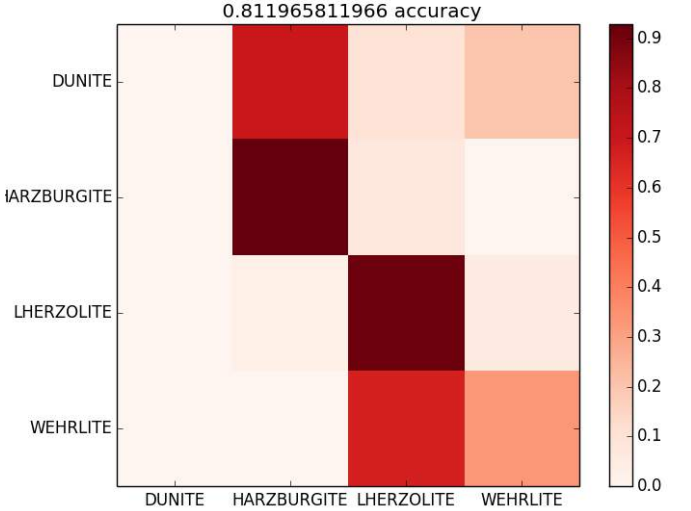
3D Representation and Confusion Matricies of the classifiers of the dataset with complete oxide values are shown on the following page. Most classifiers (11) were ommited for the due to the 20 page limit.



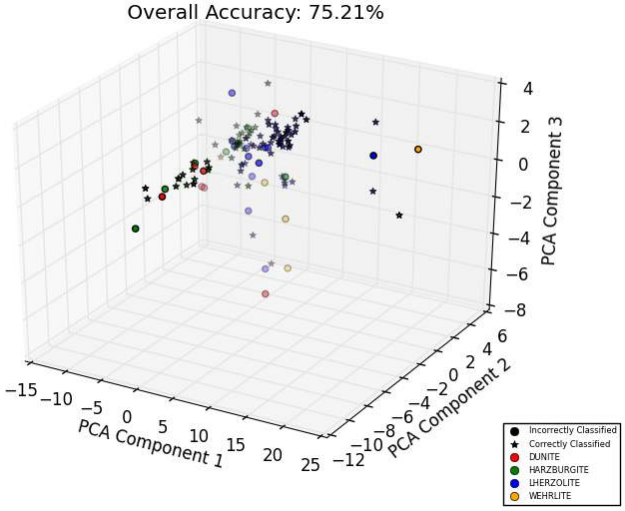
AdaBoostClassifier: :
Overall Accuracy: 81.20%



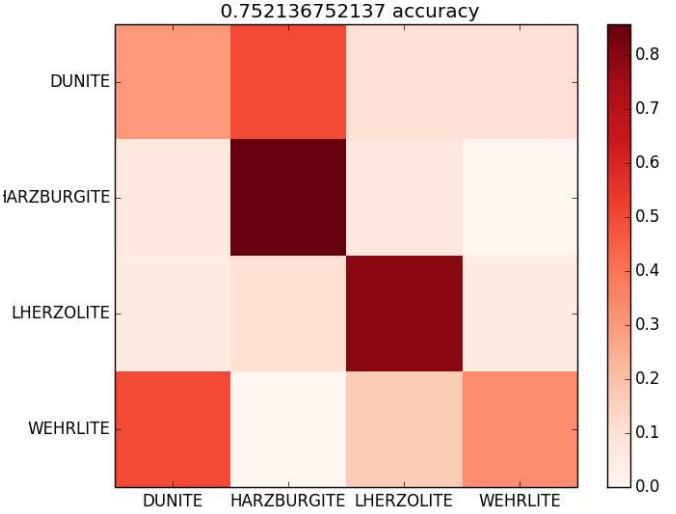
AdaBoostClassifier Confusion Matrix
0.811965811966 accuracy



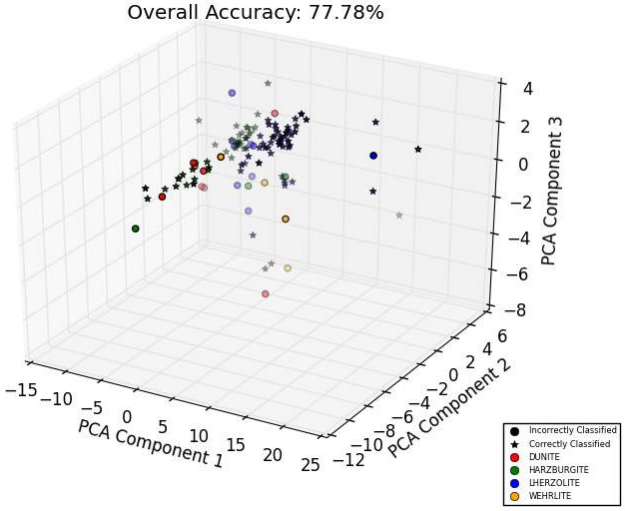
DecisionTreeClassifier: :
Overall Accuracy: 75.21%



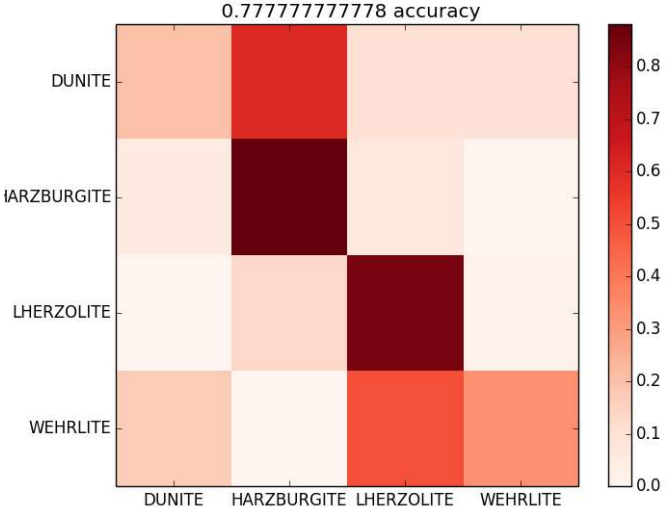
DecisionTreeClassifier Confusion Matrix
0.752136752137 accuracy

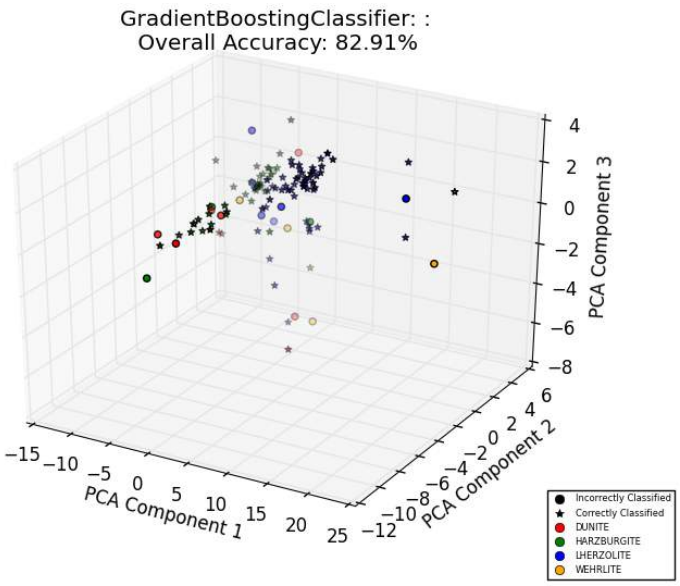


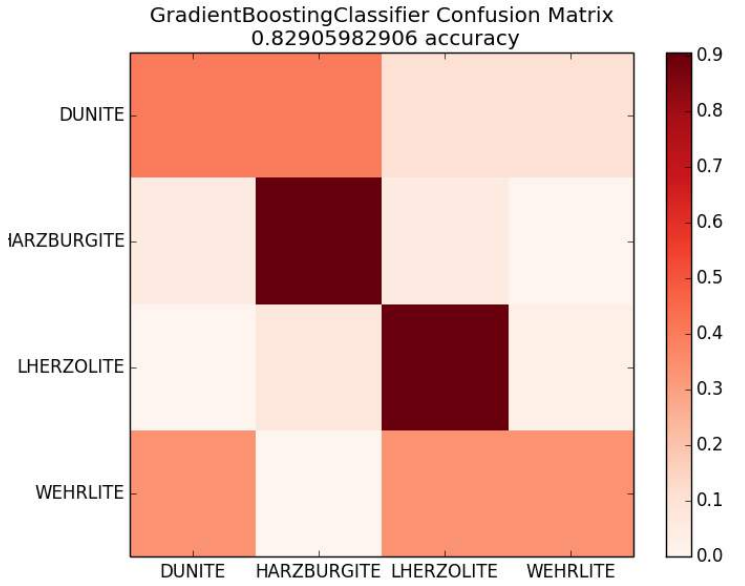
ExtraTreesClassifier: :
Overall Accuracy: 77.78%

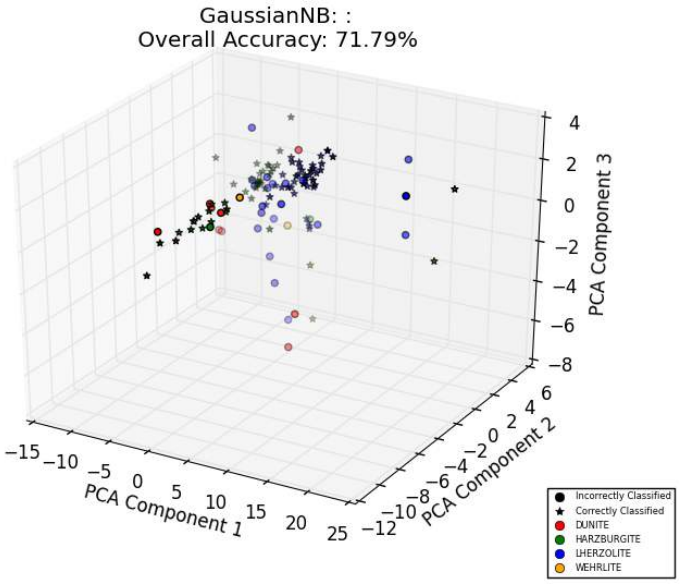


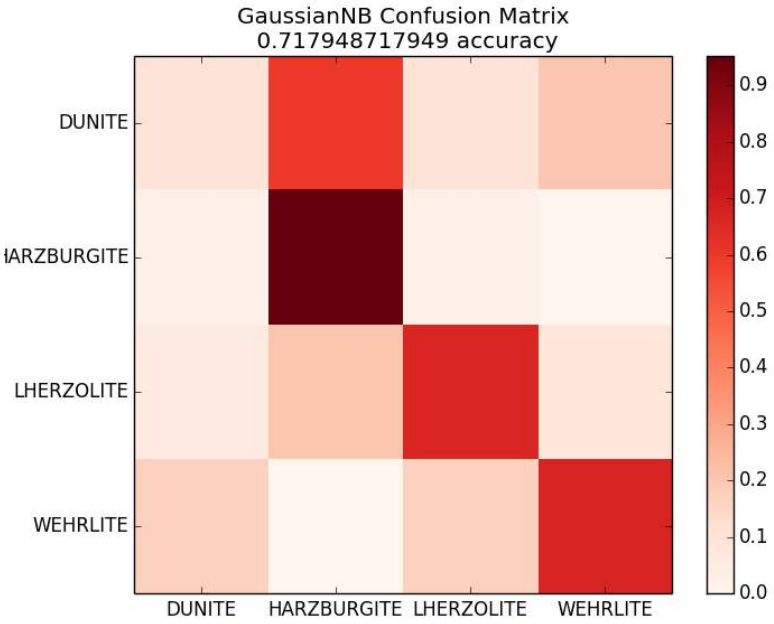
ExtraTreesClassifier Confusion Matrix
0.777777777778 accuracy




GradientBoostingClassifier: :
Overall Accuracy: 82.91%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

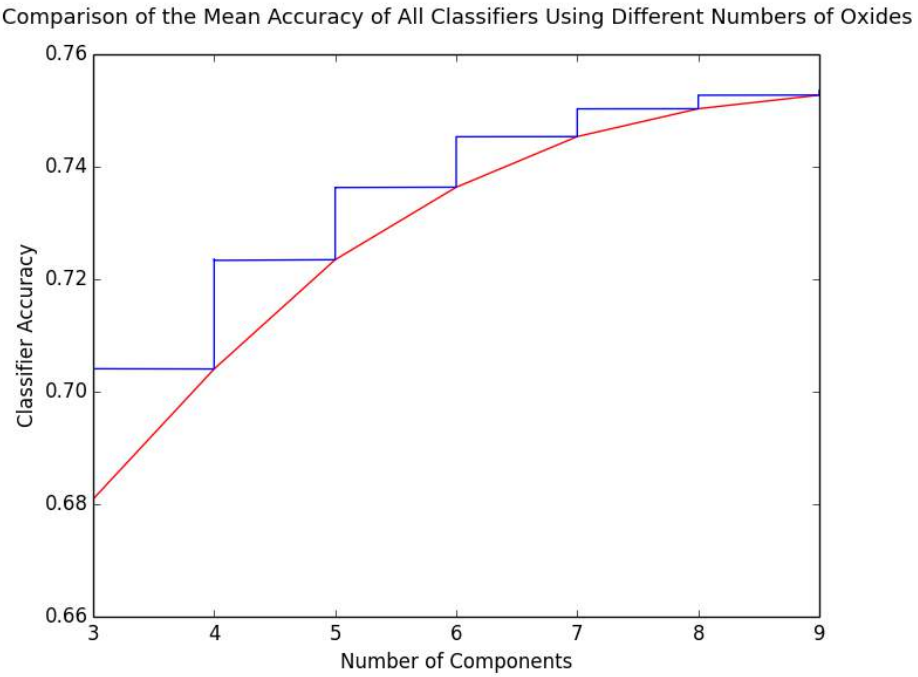

GradientBoostingClassifier Confusion Matrix
0.82905982906 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE


GaussianNB: :
Overall Accuracy: 71.79%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE


GaussianNB Confusion Matrix
0.717948717949 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

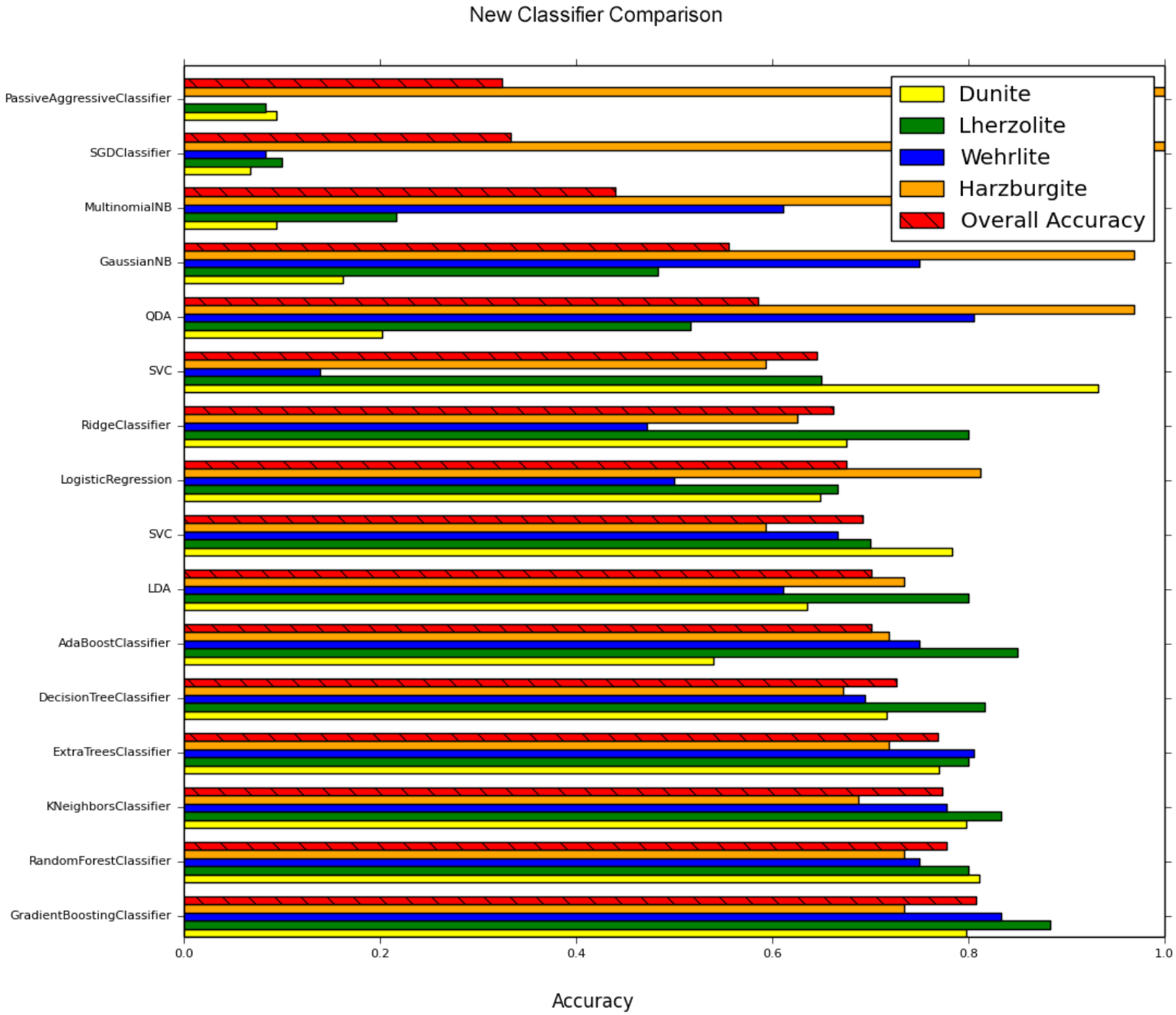
The bar graph shows the accuracies of different algorithms in their classification of various rock types from the full oxide dataset. Although the simple accuracy of these classifiers are relatively high, the results are skewed due to the large inequality in the distribution of rock types in the testing set and due to the classifiers mistaking Dunites and Werhlites with Harzburgites and Lherzolites as demonstrated in the confusion matrices. It was believed that this was due to the number of samples with all oxide values being low even though tests indicated that having more oxide values increased classifier accuracy albeit logarithmically (see below).

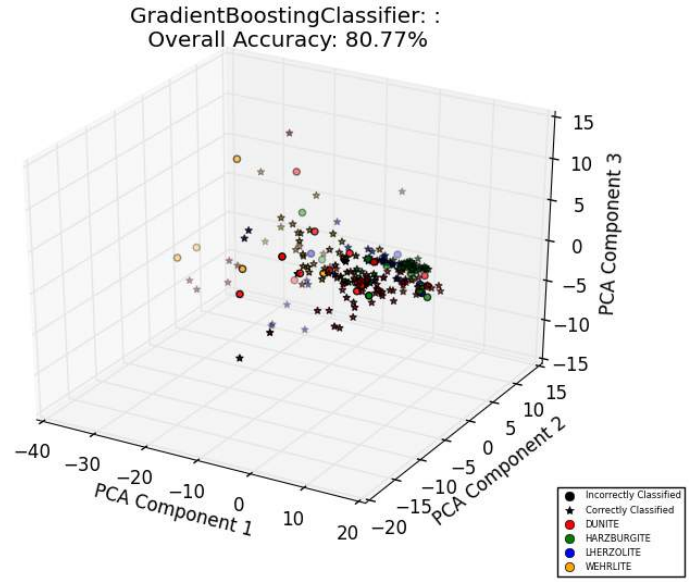
The following diagram was generated by selecting n number of oxide values in every possible combination from the database of 466 values (from left to right) and finding the mean accuracy of the classifiers run on the subset of the database. It shows that more oxide values equates to higher classifier accuracy regardless of which oxides are used and how large of a dataset is available.

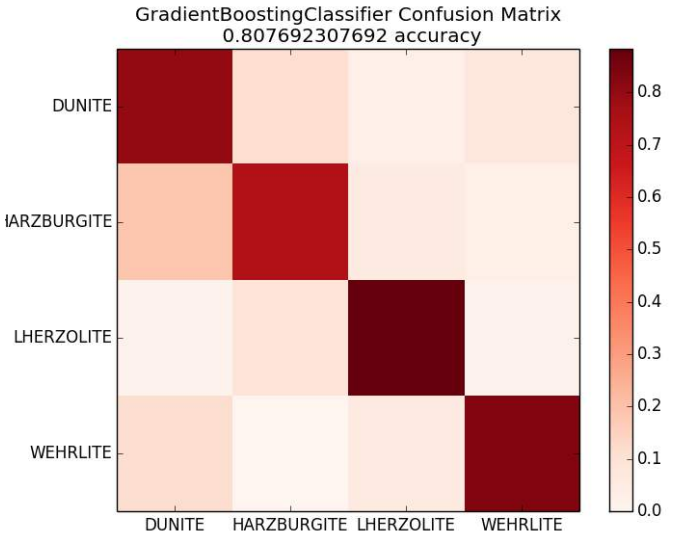


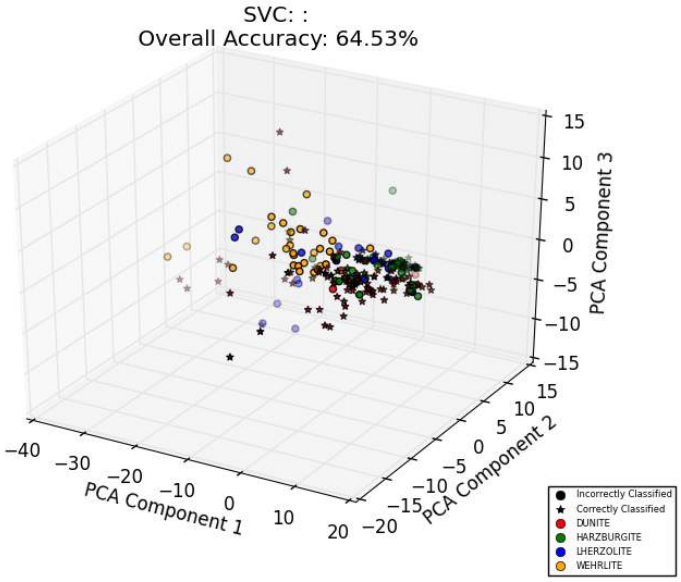
After this finding, it was decided to use statistical analysis to find a balance between number of oxide values and number of samples. In order to do this, all of the combinations of 11 oxide values were tested based on number of useable samples containing those values. This was done with the assumption that more oxide values yielded better classifier accuracies as demonstrated above. The minimum number of acceptable oxide values was determined to be 6 (anecdotally, due to it being center of the figure located above). The 6 oxide values that yielded the highest number of samples was determined to be SiO_2, TiO_2, Al_2O_3, MgO, CaO, and MnO by looping through all of the combinations of 6 oxide values (12 nCr 6) while searching for the combination yielding the highest number of samples. These six values resulted in

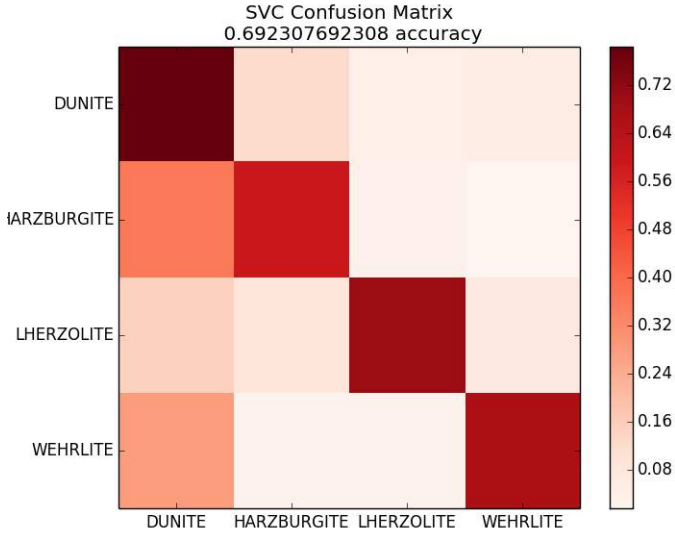
935 useable samples out of an initial database of 1220 labeled samples, the largest of all 6 oxide combinations. The same processes were then used on this larger database containing a lesser number of oxide values as was used on the smaller database containing all oxides and 466 samples.

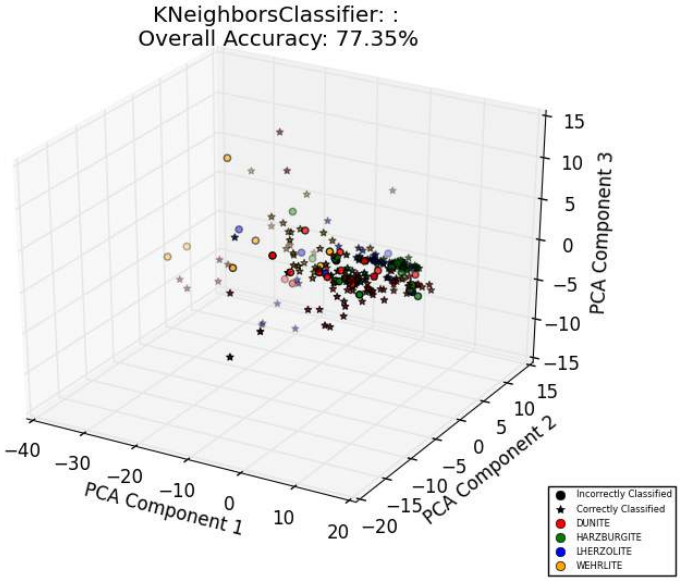


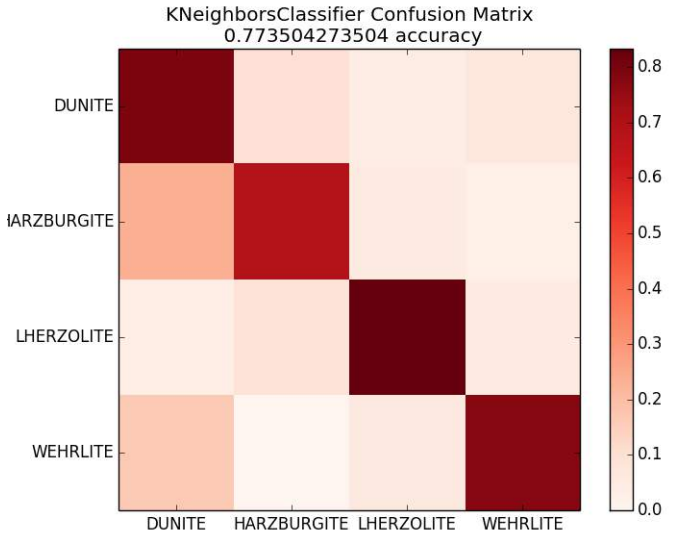

GradientBoostingClassifier: :
Overall Accuracy: 80.77%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

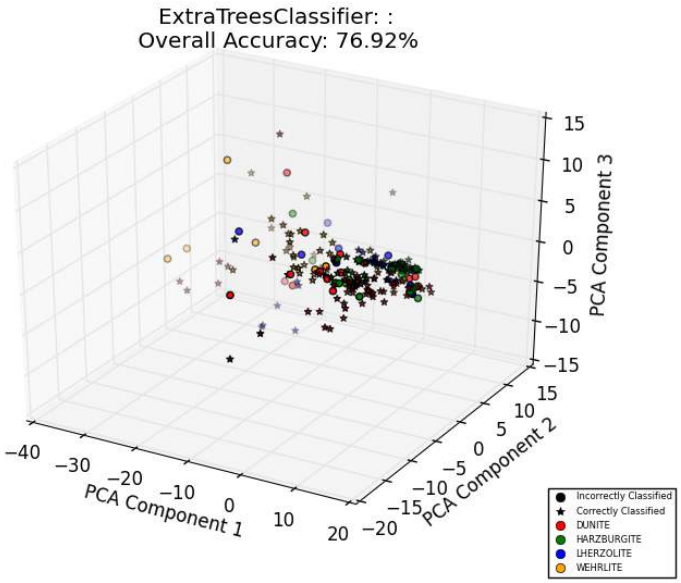

GradientBoostingClassifier Confusion Matrix
0.807692307692 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

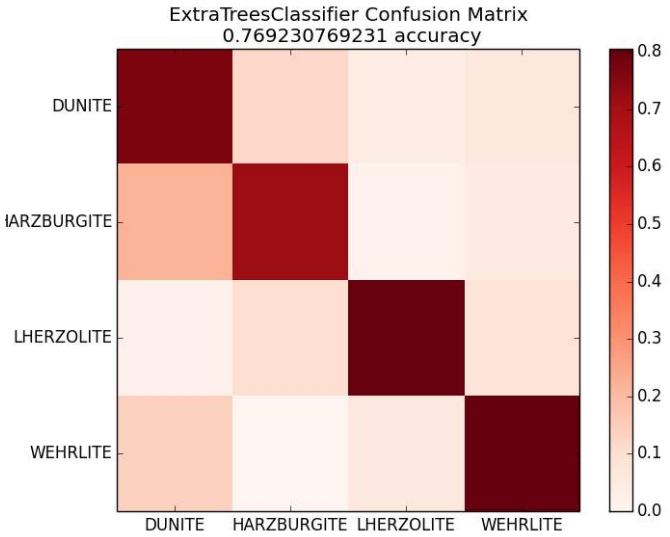

SVC: :
Overall Accuracy: 64.53%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

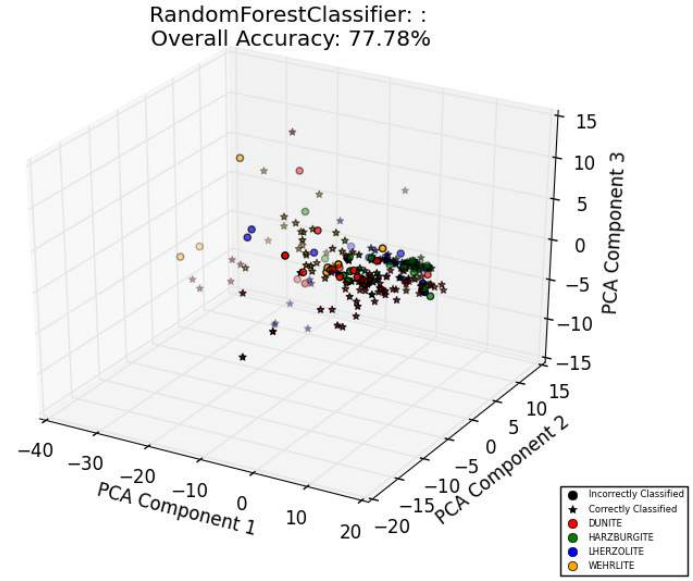

SVC Confusion Matrix
0.692307692308 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

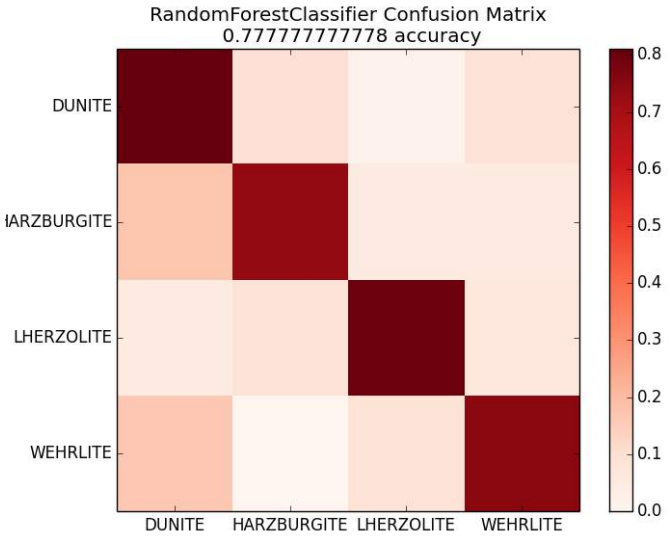

KNeighborsClassifier: :
Overall Accuracy: 77.35%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

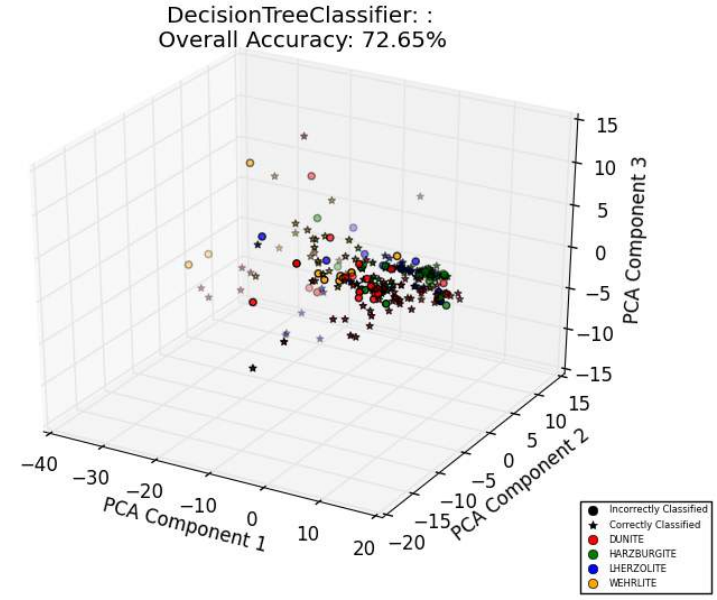

KNeighborsClassifier Confusion Matrix
0.773504273504 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

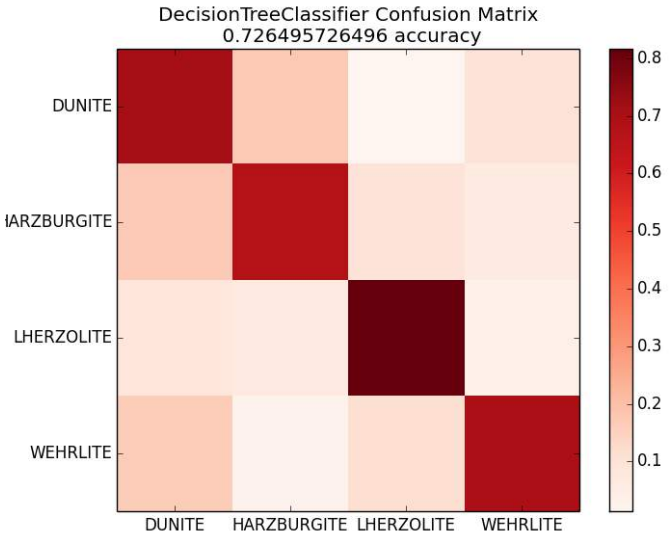

ExtraTreesClassifier: :
Overall Accuracy: 76.92%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE


ExtraTreesClassifier Confusion Matrix
0.769230769231 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE


RandomForestClassifier: :
Overall Accuracy: 77.78%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE


RandomForestClassifier Confusion Matrix
0.777777777778 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE


DecisionTreeClassifier: :
Overall Accuracy: 72.65%
PCA Component 1
PCA Component 2
PCA Component 3
Incorrectly Classified
Correctly Classified
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE


DecisionTreeClassifier Confusion Matrix
0.726495726496 accuracy
DUNITE
HARZBURGITE
LHERZOLITE
WEHRLITE

The classifier comparison was effective in terms of the selection of a suitable algorithm (Gradient Boosting was consistently more accurate in tests and in the displayed bar chart). Every classifier uses its own unique strategy (algorithm) to separate a dataset into different classes. As a result, every classifier has its own unique signature and temperament, which can therefore provide clues as to how the data provided to it, is structured from its resulting classification. In this case, Gradient Boosting and KNearestNeighbors reigned supreme in terms of classification accuracy as seen in the bar chart. This signifies that although the data has a large spread, it is rather clearly defined which allows those classifiers to form boundaries around different classes of rocks accurately. It also provided an insight into how well the algorithms scale. The classification of the 6 oxide data was much more accurate than the smaller, full oxide data in terms of real accuracy. The 6 oxide data provides relatively high accuracy, consistent classification for all rock types while not being as influenced by misclassification (due to all rock types being relatively accurate). This is clearly evident in the "staircase effect" seen in the confusion matrices. While in theory a large database with all oxide values would be optimal (based upon the number of oxide components vs. classifier accuracy diagram), that kind of data is simply not available at this point in time for peridotites and many other classes of rocks due to both the cost and labor required to collect this data. While there are other avenues being explored to improve the accuracy of these classifications, it is inferred based on these trends that in order to achieve obtain or a greater than 90-95% accuracy across rock types, a much larger dataset (around 10,000-20,000 samples) containing some oxides or a moderately larger dataset (around 10,000 samples) with full oxide values would be required.

Results:

The combination of several data handling analyses and machine learning algorithms proved capable of yielding accurate identification of peridotitic igneous rocks (ranging from 70-85% accuracy depending on rock type). To achieve these results, the data was visualized using various graphic representations, classified according to the results of the visualization, fine-tuned according to the volatility of the dataset, and packaged into an accessible application that is available for geologists worldwide to use for rapid identification of large datasets. This application is able to automatically classify both CSV and Excel files containing peridotite oxide values and is able to adjust itself to various sized datasets with different numbers of oxides. Each classifier also returns a probability matrix containing the classifier's confidence level that a sample is one of the four current rock types. For example, the following probability matrix signifies a 25% confidence that the sample is a dunite and a 75% chance that the sample is a lherzolite: probability matrix - [.25, 0, .75, 0].

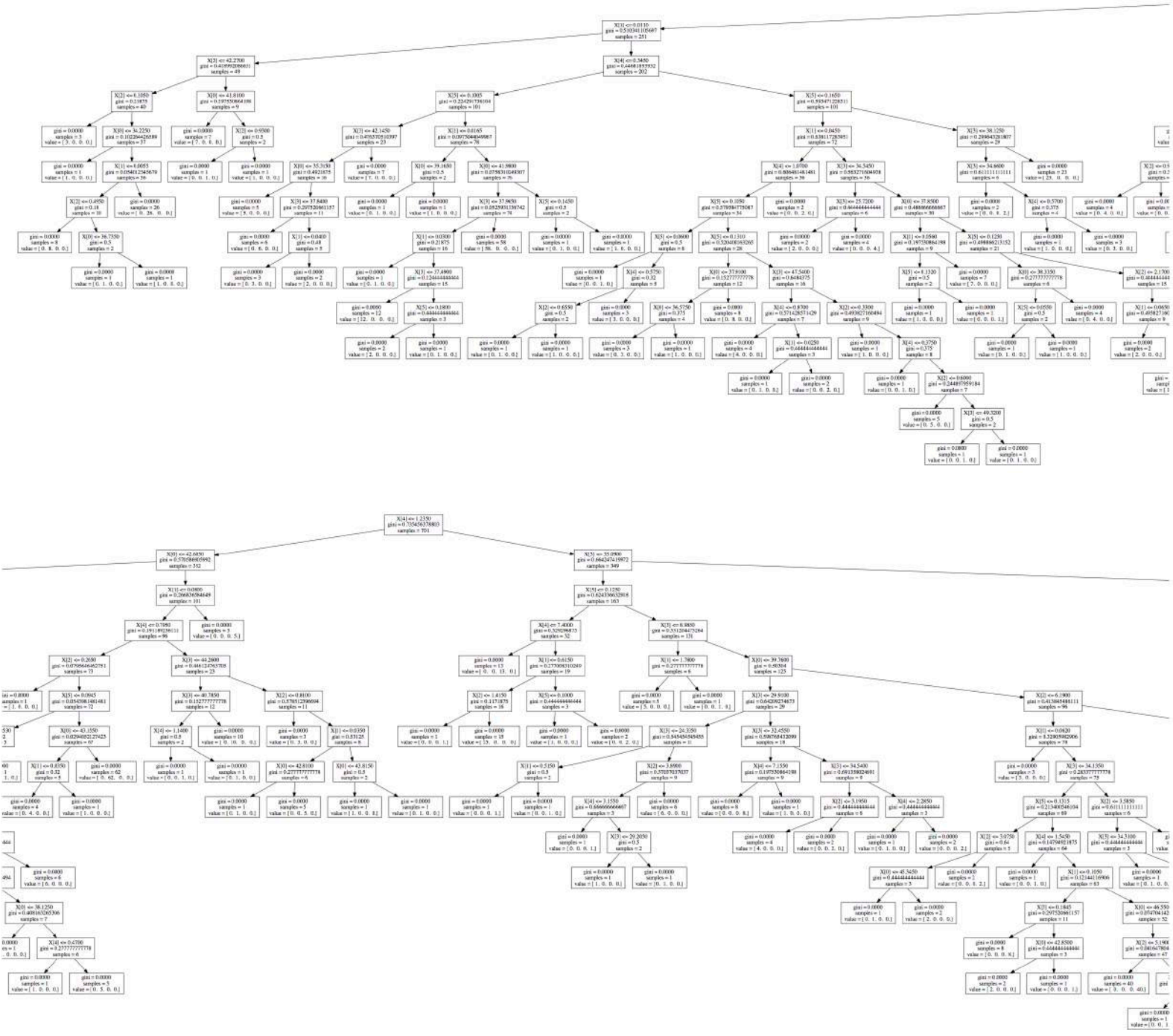
Discussion:

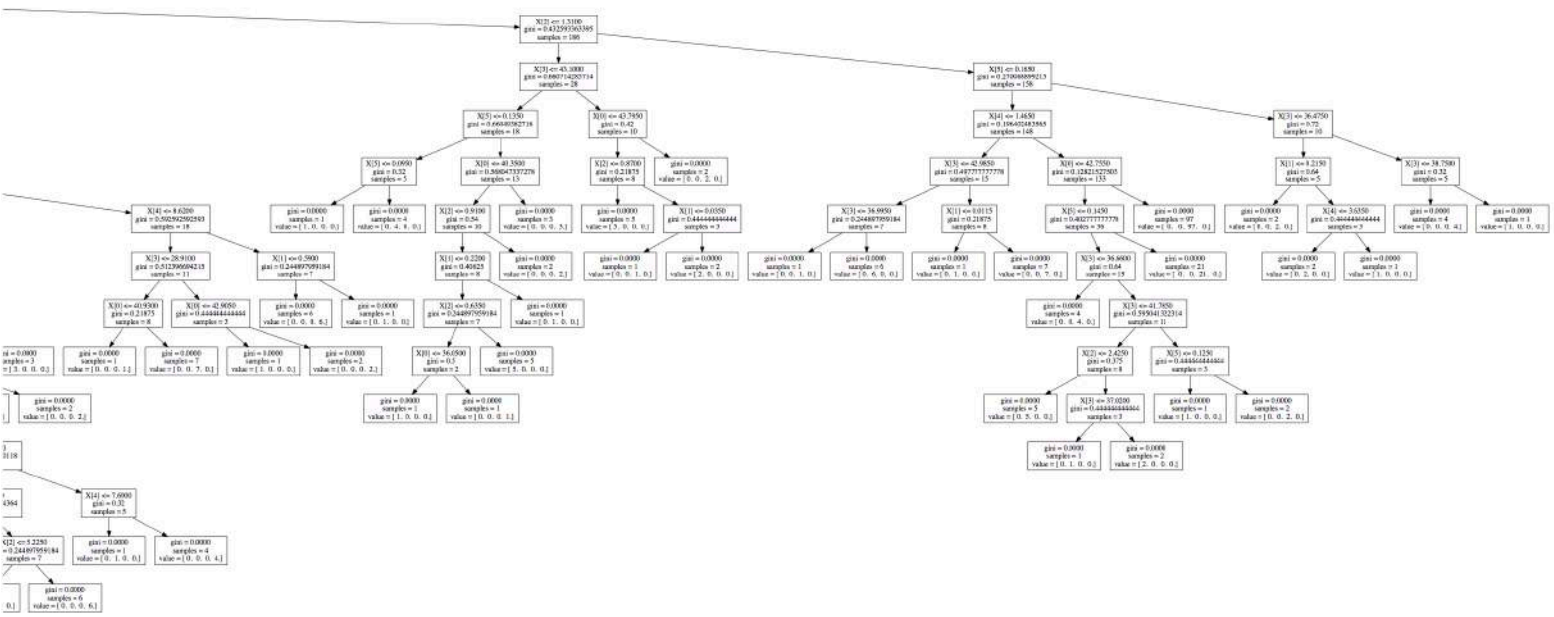
The initial work consisted of an attempt to classify igneous rocks using standard mineralogical data, and an attempt to classify igneous rocks based on oxidation data through the reverse engineering of

the Streckeisen definition of peridotitic rocks. After this approach proved to be unwieldy, a closer look was taken into the dataset using various graphical representations and statistical analysis methods in order to find relationships between different types of igneous rocks and to spot overlaps and or clustering within the data that could signify how difficult the data would be to classify. After it was determined that there were significant chemical distinctions between different types of peridotitic rocks in various dimensions, several machine learning algorithms were applied to create classifiers for these rocks using all of the dimensions available. This classification was based upon a set of 11 dimensional oxidation data. The classifiers that were generated were compared and provided fairly mediocre accuracies (accuracies lower than pure chance) on some rock types. It was determined that this discrepancy in accuracy was due to the small amount of data (a fraction of 466 samples) used to train the classifiers. As a result, since there were more samples in the dataset that contained less oxide values (6 were used), a new classifier comparison was administered using a larger dataset containing fewer oxide values. The new classifiers provided much more respectable, consistent accuracies (ranging from ~70-85% depending on rock type). Following this, additional data was obtained from the EarthChem Database, which were labeled by peridotitic rock type. This additional data allowed the classifiers to "train" with larger datasets, therefore yielding more accurate, and consistent classification across the board. This proved that the classification method easily scales and adapts to larger datasets. The concept of a "crowd sourced" training of the classifiers was considered however if users inadvertently submit a data source containing errors, it could degrade the performance of the classifiers. At this time, it is believed that this set of classifiers yields results that are more reliable than the CIPW norm classification scheme in terms of peridotitic igneous rock. In addition, with the use of a "decision tree" (see appendix), a relatively accurate classification can be accomplished based on these principles with pen and paper. Further testing and the reverse engineering of the CIPW norm would be required to prove this for igneous rocks as a whole. Ongoing work includes the development of algorithms to allow such a "crowd sourced" dataset by rejecting outlying or spurious data from contaminating the dataset by reducing the accuracy of the classifiers. Future research will also include the improvement of the current classifiers and the creation of classifiers for additional families of rocks. This research and the techniques investigated have the potential to evolve into a new, widely accepted classification scheme based on quantitative analysis and chemical composition as these techniques are applied to all major rock types.

Appendix:

This is a visual representation of the decision tree used for classification of peridotitic rocks as produced by the Decision Tree Classifier from the scikit-learn machine learning Python library. It can be used to manually classify rocks based on their oxide values (continued on next page).




X[2] <= 1.5100
gini = 0.4329593385
samples = 196
X[3] <= 43.1000
gini = 0.6807142857714
samples = 28
X[5] <= 0.1350
gini = 0.6668085827718
samples = 18
X[0] <= 43.7950
gini = 0.42
samples = 10
X[5] <= 0.0950
gini = 0.32
samples = 5
X[0] <= 40.3500
gini = 0.5000547337278
samples = 13
X[2] <= 0.9700
gini = 0.21875
samples = 8
gini = 0.0000
samples = 2
value = [0. 0. 2. 0.]
X[4] <= 8.6200
gini = 0.5025992592593
samples = 18
X[2] <= 0.9100
gini = 0.54
samples = 10
gini = 0.0000
samples = 5
value = [0. 0. 0. 5.]
gini = 0.0000
samples = 5
value = [5. 0. 0. 0.]
X[1] <= 0.0350
gini = 0.4444444444444
samples = 3
X[3] <= 28.9100
gini = 0.5123996094215
samples = 11
X[1] <= 0.5900
gini = 0.2449979591864
samples = 7
X[1] <= 0.2200
gini = 0.40625
samples = 8
gini = 0.0000
samples = 2
value = [0. 0. 0. 2.]
gini = 0.0000
samples = 1
value = [0. 0. 1. 0.]
gini = 0.0000
samples = 2
value = [2. 0. 0. 0.]
X[3] <= 42.9050
gini = 0.4444444444444
samples = 3
gini = 0.0000
samples = 6
value = [0. 0. 0. 6.]
gini = 0.0000
samples = 1
value = [0. 1. 0. 0.]
X[2] <= 0.6550
gini = 0.2449979591864
samples = 7
gini = 0.0000
samples = 1
value = [0. 1. 0. 0.]
gini = 0.0000
samples = 2
value = [0. 0. 0. 2.]
X[0] <= 56.0500
gini = 0.5
samples = 2
gini = 0.0000
samples = 5
value = [5. 0. 0. 0.]
gini = 0.0000
samples = 1
value = [1. 0. 0. 0.]
gini = 0.0000
samples = 1
value = [0. 0. 0. 1.]
X[5] <= 0.1650
gini = 0.2700888959213
samples = 158
X[4] <= 1.4650
gini = 0.1964024483565
samples = 148
X[3] <= 42.9850
gini = 0.4977777777778
samples = 15
X[6] <= 42.7550
gini = 0.1282152276865
samples = 133
X[3] <= 36.9950
gini = 0.2448979591864
samples = 7
X[1] <= 0.0115
gini = 0.21875
samples = 8
X[5] <= 0.1450
gini = 0.0027777777778
samples = 39
gini = 0.0000
samples = 97
value = [0. 0. 97. 0.]
gini = 0.0000
samples = 1
value = [1. 0. 0. 0.]
gini = 0.0000
samples = 6
value = [0. 0. 6. 0.]
gini = 0.0000
samples = 1
value = [0. 1. 0. 0.]
gini = 0.0000
samples = 7
value = [0. 0. 7. 0.]
X[3] <= 36.8600
gini = 0.04
samples = 19
gini = 0.0000
samples = 21
value = [0. 0. 21. 0.]
gini = 0.0000
samples = 4
value = [0. 0. 4. 0.]
X[3] <= 41.7850
gini = 0.5950413223314
samples = 11
X[2] <= 2.4250
gini = 0.375
samples = 8
X[5] <= 0.1250
gini = 0.4444444444444
samples = 3
gini = 0.0000
samples = 5
value = [0. 5. 0. 0.]
X[3] <= 37.0200
gini = 0.4444444444444
samples = 3
gini = 0.0000
samples = 1
value = [1. 0. 0. 0.]
gini = 0.0000
samples = 2
value = [0. 0. 2. 0.]
gini = 0.0000
samples = 1
value = [0. 1. 0. 0.]
gini = 0.0000
samples = 2
value = [2. 0. 0. 0.]
X[1] <= 0.2150
gini = 0.64
samples = 5
X[3] <= 38.7500
gini = 0.32
samples = 5
gini = 0.0000
samples = 2
value = [0. 0. 2. 0.]
X[4] <= 3.6350
gini = 0.4444444444444
samples = 3
gini = 0.0000
samples = 4
value = [0. 0. 0. 4.]
gini = 0.0000
samples = 1
value = [1. 0. 0. 0.]
gini = 0.0000
samples = 2
value = [0. 2. 0. 0.]
gini = 0.0000
samples = 1
value = [1. 0. 0. 0.]
X[4] <= 7.6000
gini = 0.32
samples = 5
gini = 0.0000
samples = 1
value = [0. 1. 0. 0.]
gini = 0.0000
samples = 4
value = [0. 0. 0. 4.]
gini = 0.0000
samples = 6
value = [0. 0. 0. 6.]

References:

1. Applications of Machine Learning in Cancer Prediction and Prognosis Joseph A. Cruz, David S. Wishart Cancer Inform. 2006; 2: 59–77. Published online 2007 February 11.

2. Holland, H. D., K. K. Turekian, and R. W. Carlson. Treatise on Geochemistry: Volume 2: The Mantle and Core. Vol. 2. Amsterdam: Elsevier, 2004. Print. P.171

3. Kelsey, C. H. "Calculation of the C.I.P.W. Norm." *Mineralogical Magazine* 34.268 (1965): 276-82. Web.

4. Le Bas, M. J., and A. L. Streckeisen. "The IUGS Systematics Of Igneous Rocks." Journal of the Geological Society 148 (1991): p. 833. Print.

5. National Science Foundation "Research Areas." US NSF. N.p., 9 July 2011. Web. 09 Nov. 2014.

6. Richert, Willi, and Luis Pedro. Coelho. Building Machine Learning Systems with Python. Birmingham: Packt, 2013. Print.

7. Scikit-learn 0.15.2 Documentation. Web. 09 Nov. 2014. — "Classifier Comparison." --

8. "Scikit-learn.": Machine Learning in Python — 0.15.2 Documentation. Web. 09 Nov. 2014.