1-1-2020

Statistical Analysis And Machine Learning For Coal Classification For Rare Earth Elements + Y (REY)

Zachary Bartley Young

STATISTICAL ANALYSIS AND MACHINE LEARNING FOR COAL CLASSIFICATION

FOR RARE EARTH ELEMENTS + Y (REY)

by

Zachary Bartley Young

Bachelor of Science, University of Arkansas, 2013

A Thesis

Submitted to the Graduate Faculty

of the

University of North Dakota

In partial fulfillment of the requirements

for the degree of

Master of Science

Grand Forks, North Dakota

May

2020

This thesis   , submitted by Zachary Young   in partial fulfillment
of the requirements for the Degree of Master of Science in Energy Systems Engineering
from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the
work has been done and is hereby approved.

Michael Mann
_____
Michael Mann

Bruce Folkedahl
_____
Bruce Folkedahl

Sean Hammond STH
_____
Sean Hammond

_____
Name of Committee Member 3

_____
Name of Committee Member 4

_____
Name of Committee Member 5

      This   thesis   is being submitted by the appointed advisory committee as having
met all of the requirements of the School of Graduate Studies at the University of North Dakota and is
hereby approved.

Chris Nelson
_____
Chris Nelson
Dean of the School of Graduate Studies

5/5/2020
_____
Date

ii

PERMISSION

Title   Statistical Analysis and Machine Learning for Coal Classification for Rare Earth Elements + Y (REY)

Department   Energy Systems Engineering

Degree   Master of Science

       In presenting this thesis in partial fulfillment of the requirement for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my thesis work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this thesis or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my thesis.

<div align="right">

Zachary Bartley Young
May 6th 2020

</div>

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

ABSTRACT

Due to their exceptional properties, rare earth elements (REEs) are critical to technological

innovation in renewable energy production, electronics, health care, and national defense. They

make up key components for many applications in the above areas. Many countries rely upon

rare earth element imports. The high demand for rare earth elements has led to the development

of alternative methods for exploration and capture. Coal has been labeled a viable potential

source of rare earth elements and yttrium (REY). Statistical evaluation of REY concentrations

and the properties of various coal samples is critical for successful characterization.

The USGS COALQUAL database Version 3.0 is an industry standard database for coal

research that contains 7658 non-weathered, full-bed coal samples from the United States. 5485

of these samples contain a full spectrum of REY concentrations. The data quality in the

COALQUAL database will be analyzed to ensure that the data is reliable, and characteristics will

be analyzed using conventional statistical methodology. This methodology includes accounting

for samples with REY concentrations below the lowest limits of detection. Mean concentrations

for each REY will be adjusted to fit a distribution of mean REY concentrations from the National

Coal Resources Data System (NCRDS) normalized by the Upper Continental Crust standard

dataset of REY mean concentrations. All samples are classified as unpromising or promising

using total rare earth oxide concentration and the ratio of critical REYs to excess REYs called

the outlook coefficient.

Machine learning is a powerful tool that can utilize data to classify new data points added

to a database based on data attributes. A machine learning model was developed to use existing

data from the COALQUAL database to train and test algorithms to classify coal samples as

unpromising or promising based on the samples ASTM ash percentage. The 5485 adjusted coal samples from the COALQUAL database were used and subjected to synthetic minority over-sampling technique (SMOTE) to eliminate label bias, and imputing methods were used to format the data for computational purposes. The adjusted coal samples were tested amongst various machine learning algorithms for the best performance. Accuracy and the number of false positives were the key performance indicators used to test each algorithm. The k-nearest neighbors (KNN) algorithm emerged as the best performer with 92% accuracy and 2% false positives. A brief economic analysis is included to justify using the model to save costs associated with obtaining trace element concentrations from laboratory analysis. Recommendations are given with details on how to utilize this research for future endeavors.

CHAPTER 1: INTRODUCTION

The global outlook on rare earth element supply is precarious. Due to growth in technology and the demand for such, the world supply of these essential elements is dwindling. Dedicated mining for rare earth elements has proven to be uneconomical in most areas of the world, but alternative extraction ideas have been considered. One of which is to extract rare earth elements from coal. Extensive research has been conducted that suggests that rare earth elements from coal can be a viable supplement to the global supply of rare earth elements.

Growing investment in the science of rare earth element extraction from coal has led to innovation in the areas of exploration, characterization, processing, and refining. One of the most important research areas in characterization. It is important to understand whether coal from a particular geological and geographical nature will yield significant returns in terms of rare earth element content. The consequences of poor characterization is wasteful spending and inadequate supply recovery.

Due to growing technological capabilities, machine learning has been a tool that many researchers are turning to in order to provide fast, low cost solutions to complex problems. Given reliable and numerous data points, machine learning algorithms can learn from the data and provide powerful insights.

The purpose of this work is to utilize machine learning algorithms to create a useful model that aims to classify coal samples as promising or unpromising in terms of their rare earth element concentration and economic outlook. The goal is ultimately to lower the cost and time associated with using analytical techniques to measure the rare earth element concentrations of coal samples collected for research purposes.

CHAPTER 2: LITERATURE REVIEW

The following chapter provides background information on rare earth elements, how they relate to coal, and how statistical analysis and machine learning have the potential to add value to research pertaining finding significant rare earth element concentrations in coal. Additional information can be located in Appendix A.

2.1. Background on Rare Earth Elements

Rare earth elements (REE) are a group of elements characterized by having atomic numbers 57 – 71. They are comprised of the lanthanide series (lanthanum (La), cerium (Ce), praseodymium (Pr), neodymium (Nd), promethium (Pm), samarium (Sm), europium (Eu), gadolinium (Gd), terbium (Tb), dysprosium (Dy), holmium (Ho), erbium (Er), thulium (Tm), ytterbium (Yb), and lutetium (Lu)). The other REEs are scandium (Sc) and yttrium (Y) which can be considered REEs due their similar properties. See Fig. 1 for the locations of REEs on the periodic table. The term REE is used variably in literature, so a scope needs to be defined. This research only includes the lanthanides and Y in the list of REEs. Lanthanides + Y is referred to as REY. Yttrium is associated with lanthanides due to its ionic radius being very similar, and its ionic charge is the same as Ho. Because of this, yttrium is often placed between Dy and Ho in standardized REY charts.

It is not exactly true that rare earth elements are "rare," but rather they are uniformly distributed throughout the earth's crust. Therefore, it is "rare" to find REYs concentrated in one place. The origins of the nomenclature begin with their discovery in the 18th century. When they were discovered along with types of oxides labeled "earths," the elements were deemed to be rather uncommon.

Figure 1. REEs denoted by red boxes [1, 2, 3, 4, 5, 6].

2.2. Properties of REYs

The physical, chemical, and metallurgic activity of REYs is dictated by their electron configuration. The common electron configuration shared by all REYs is a trivalent oxidation state. Sm, Eu, and Y are able to obtain divalent oxidation states, and Ce, Pr, and Tb are able to obtain tetravalent oxidation states. The lanthanides all have the same valence electron configuration 5d6s2 while Y has a valence electron configuration 4d5s2.

An extraordinary property of lanthanides is with increasing molecular weight, ionic radius decreases. This is called lanthanide contraction. This applies only to lanthanides that obtain the trivalent oxidation state. The contraction is due to the 5s and 5p orbitals entering into

the 4f subshell. This means the 4f orbital is unprotected from increasing nuclear charge. With poor shielding, the positively charged nucleus has a larger affinity to the electrons towing them. The result is as the atomic number increases the ionic radius decreases. Lanthanide contraction allows lanthanide separation techniques to work. Basicity is the key property used in separation techniques. As basicity decreases, hardness, density, and melting point will increase with molecular weight increasing [7].

The physical properties of lanthanides vary amongst each element unlike their chemical properties. In Fig. 2, the melting points of each element are plotted with respect to molecular weight, and the transition of crystal structures of each REY are also displayed.
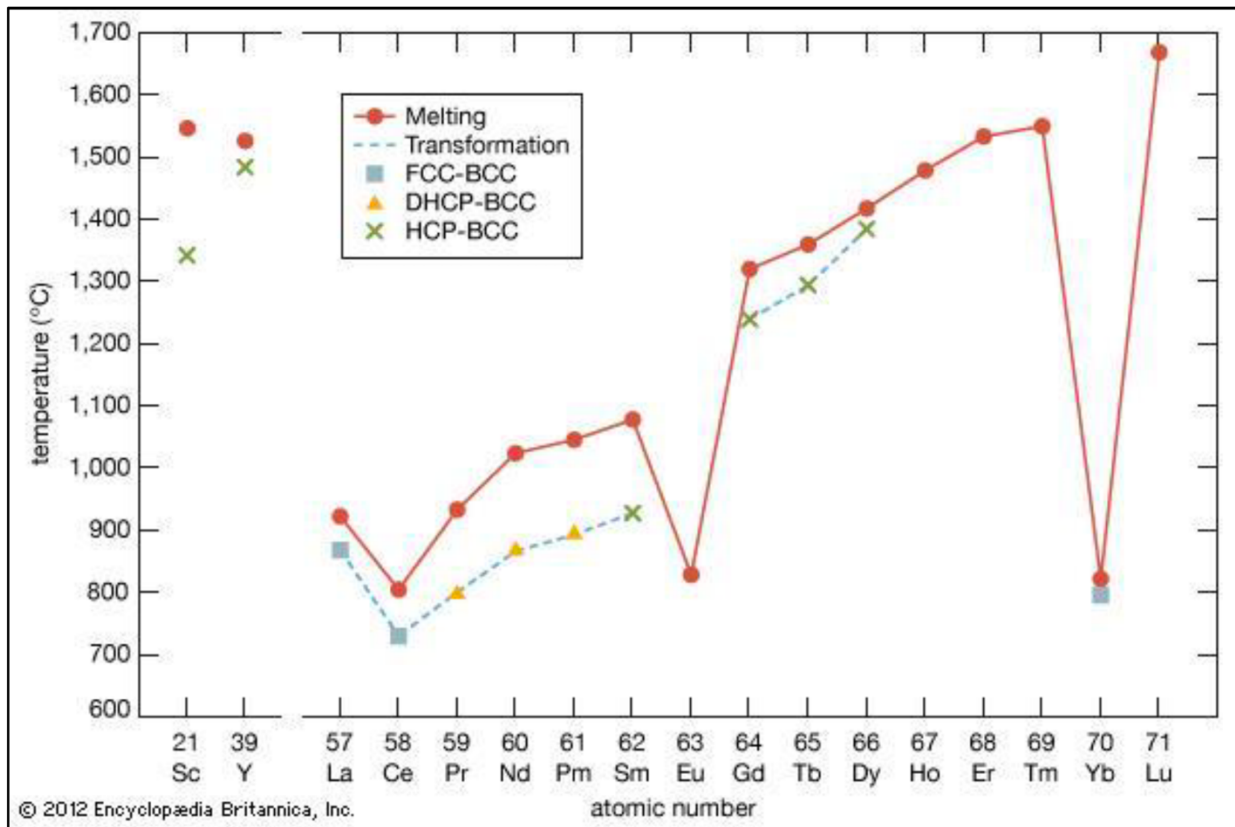
Figure 2. Transition temperature plotted against atomic number specifying melting temperatures and transformations of crystal structures [8].

Vapor pressure and boiling point also vary amongst REYs. This can be attributed to the differing oxidation states that affect crystalline structure and magnetic properties.

Rare earth magnets are among the strongest magnets in known existence. Due to their extraordinary magnetic properties, some lanthanides are extremely valuable in creating permanent magnet configurations. The cause of these properties lies in the number of unpaired 4f electrons. With increasing lanthanide atomic number, a 4f electron is added incrementally and a magnetic moment is created due to electron spin. Additional electrons align parallel until the 4f level is filled at Gd. After Gd, the newly added electrons align antiparallel until Lu is reached and no magnetic moment is created due to all electrons being paired. REYs consisting of all paired electrons have weak magnetic properties. REYs with unpaired 4f electrons are highly magnetic and make up the largest cluster of magnetic metals amongst known elements.

REYs are extremely reactive with all acids except hydrofluoric acid and produce hydrogen gas as a byproduct. They are normally ionic and act as strong reducers. In the presence of water, lanthanides will release hydrogen gas at a temperature dependent rate. They react with hydrogen gas to produce $REYH_2$. In the presence of strong hydrides, REYs produce $REYH_3$. REYs can react with organic molecules yielding organic complexes. The most researched and critical categories of REYs are halides, hydrides, and oxides.

2.3. REY Applications

REYs have been labeled as "chemical vitamins" due to having the quality of producing astonishingly different properties when reacted with different materials. With their unique chemical and physical properties, the REYs allow technology to work functionally with lower weight, energy consumption, and emissions and can improve efficiency, size, speed, structural and thermal integrity, and performance [9, 10]. NETL data shows that catalysts embody the largest

amount of application in the United States. Areas of the market that are dependent on products manufactured using REYs include military technology, health care, electronics, transportation, lighting, communication and audio devices [10]. Fig. 3 shows various applications for REEs. Prices shown are for 2008 and can experience significant fluctuations. Fig. 4 displays application percentages in the United States versus the world and details more specific REE applications.

Overview of the Rare Earth Elements		Oxide price
Element	Uses	US$ per kg
Lanthanum	Batteries, Catalyst, Lenses	$40
Yttrium `hr`	Lasers, Superconductors	$50
Neodymium	Lasers, Magnets, Computers	$60
Cerium	Catalyst, Fuel additive, Optical polish	$65
Praseodymium	Lasers, Magnets, Lighting, Alloys	$75
Gadolinium `hr`	Lasers, Magnets, Computers, X-rays	$150
Dysprosium `hr`	Lasers, Magnets, Cars	$160
Erbium `hr`	Lasers, Alloys, Photography	$165
Samarium	Lasers, Magnets, Neutron absorbtion	$350
Ytterbium `hr`	Lasers, Alloys, Gamma rays	$450
Holmium `hr`	Lasers, Magnets, Optics	$750
Terbium `hr`	Lasers, Phosphors, Lighting	$850
Europium `hr`	Lasers, Phosphors, Lighting	$1,200
Thulium `hr`	Lasers, X-rays	$2,500
Luteium	Catalyst, Medicine	$3,500
Scandium	Lasers, Lighting, Aerospace	$14,000*
Promethium	Nuclear batteries	No price

☐ Lanthanides   `hr` Heavy REE   * Refined metal

Figure 3. REEs and their uses with price per kilogram on an oxide basis [11]
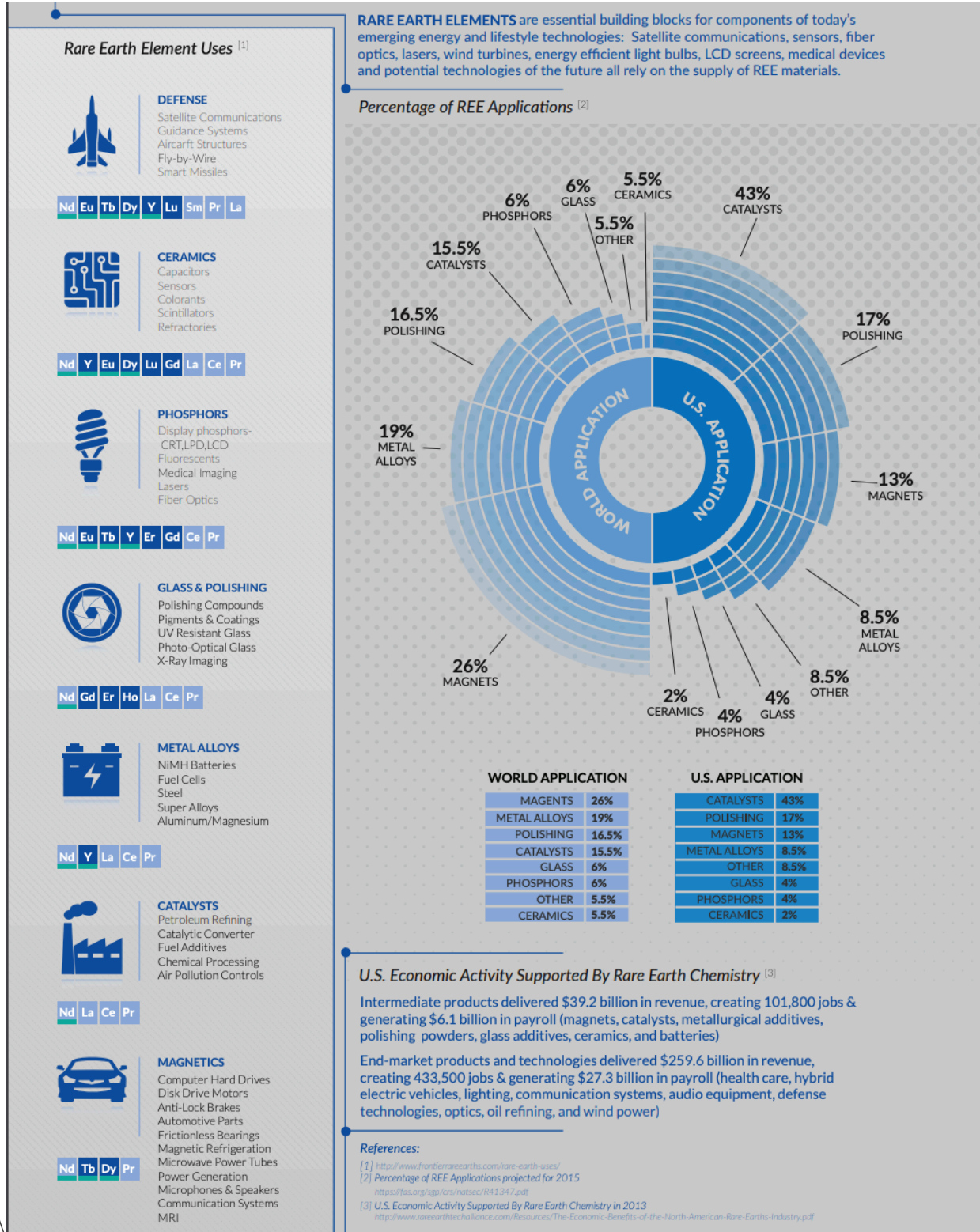
Figure 4. REE Application Percentages in the United States versus the World and more specific REE applications [10]

2.4. REY Geology

As mentioned previously, REYs are considered "rare," due to their presence in the Earth's crust rarely being concentrated; therefore, economically viable opportunities are difficult to identify. REY occurrence is strongly related to geological circumstance. Because REYs are hard to separate chemically, they often appear together in various mineral forms. There are 200 identified minerals where REYs are found [12], and six particular minerals that are the most successful in terms commercial production [13]. They are Bastnasite [(Ce,La)(CO$_3$)F], Monazite [(Ce,La)PO$_4$], Xenotime [YPO$_4$], Loparite [(Ce,Na,Ca)(Ti,Nb)O$_3$], Apatite [(Ca,REE,Sr,Na,K)$_3$Ca$_2$(PO$_4$)$_3$(F,OH)], and Ion-adsorption clays. Bastnasite, Monazite, and Xenotime make up most of the world's known reserves ~ 95% [14]. There are two major classifications of REY deposits: magmatic deposits and sedimentary deposits. These deposits can be broken down into mineral types. Magmatic deposits consist of carbonitite, perlalkaline, and pegmatitic minerals. Sedimentary deposits consist of residual, placer, phosphorite, phosphate, ion-adsorption clays, coal and associate minerals [15]. These deposits are described more in-depth in Appendix A [16].

2.5. REY Criticality

The criticality of REYs have recently been characterized in terms of supply risk and the effect of supply reduction or significance to a combination of technologies and applications. A critical materials assessment was recently completed by the United States Department of Energy (DOE) where each REY was evaluated based on five year and fifteen year outlooks [17]. A visualization how REYs were assessed is displayed in Fig. 5 for both outlooks. The criticality was assessed based on their associations to clean energy and their supply risk. REY clean energy importance can be broken down and weighted by clean energy demand (75%) and

limitations in sustainability (25%). Supply risk can be broken down and weighted by basic

availability (40%), producer diversity (20%), political, regulatory, and social factors (20%),

competing technology demand (10%), and codependence on other markets (10%). From this

analysis, Nd, Eu, Tb, Dy, Y, and Er are considered critical REYs [10]. Dy is considered the most

critical REY in both time scenarios. Seredin (2010) also categorized REY criticality based on

IMCOA market evaluations. REYs can be broken down into categories based on Seredin's

research: critical (Nd, Eu, Tb, Dy, Y, and Er), uncritical (La, Pr, Sm, and Gd), and excess (Ce,

Ho, Tm, Yb, and Lu) [18].
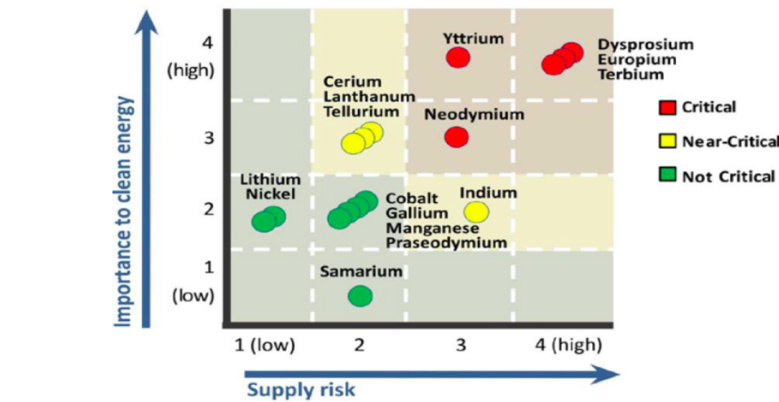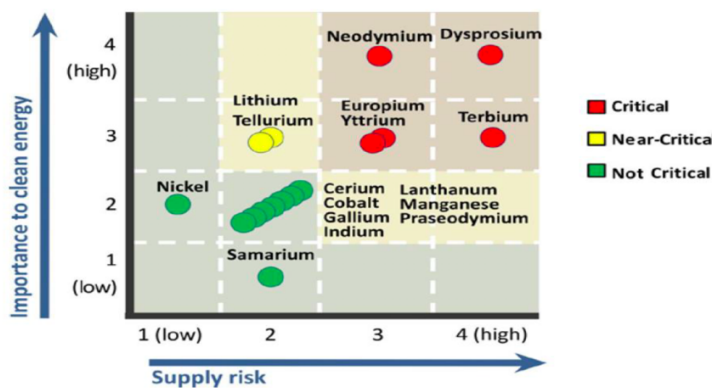
Figure 5. U.S. Department of Energy 2011 critical materials assessment for clean energy

for short and medium terms [17]

9

For simplicity, Seredin's REY categories will be used in the rest of this document to describe REY criticality. Based on these evaluations, it is fair to say that REYs are some of the most important of mineral commodities in the United States and the rest of the world.

2.6. Global Outlook of REYs

It is estimated that China possesses 30% - 50% of all global REY reserves [19, 20]. This is consistent with China being the dominant country in terms of global REY supply, totaling 83% in 2016. In 2010, China's global supply contribution was about 95% [19]. This was not always the case as the United States controlled the global market in first half of the 20th century. Fig. 6 shows a history of REY production comparing the U.S., China, and the rest of the world throughout the 20th century. In the 1980's, China began to dominate the market for a number of reasons including inexpensive human capital and less severe environmental regulations. In 2002, the last remaining rare earth element mine in the United States, Mountain Pass, closed. It has since been reopened and is currently owned by stakeholders from the U.S. and China. In 2005, long-term unstable production rates incentivized China to enact export quotas on REYs that benefit domestic supply chains and created a global monopoly for China. A Chinese REY embargo against Japan catalyzed speculative supply fears for western countries and caused high significant global price surge for REYs. As a result, China assigned export quotas for local and foreign REY exporting businesses. China was forced to change their export policies after a World Trade Organization (WTO) dispute amongst the European Union (EU), Japan, and the United States. Because of relational disputes in the global REY market, stability is fragile [20].
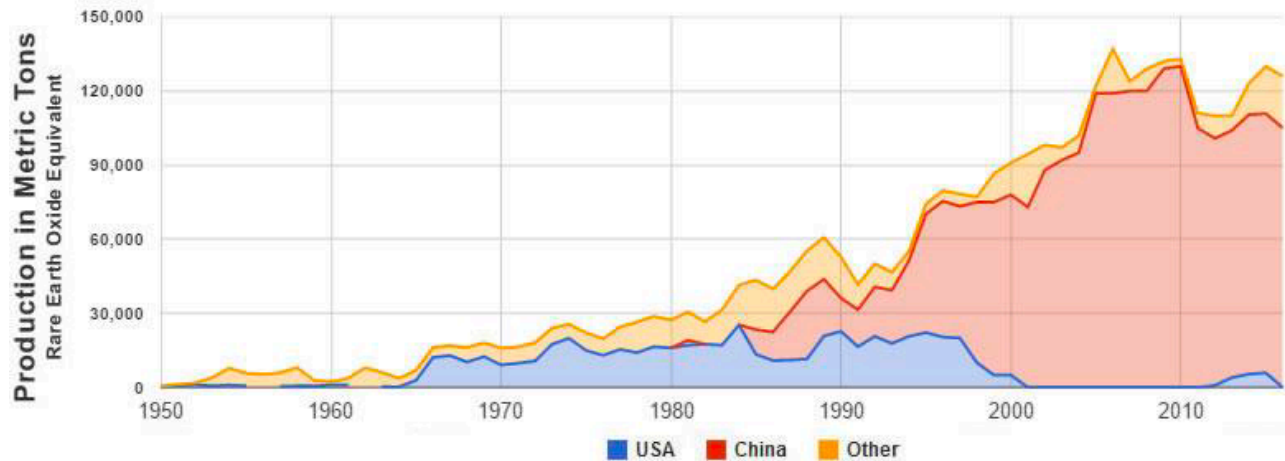
Figure 6. Rare earth element production through 2016 [21]

The potential for instability in the global REY market has catalyzed the need for several countries to begin looking into domestic exploration, mining, and production opportunities. A few of the top REY producers outside of China are Australia, the U.S., Russia, Myanmar, and India. With the reopening of Mountain Pass mine and the increases in mining and production active globally, the global production of REEs in 2018 was 170000 MT. China is still the largest producer at 120000 MT (~71%). Australia is the second largest producer with 20000 MT (~12%), the U.S. is the third largest producer with 15000 MT (~9%). The rest of the global market makes up the remaining 8% of the produced REEs [22].

Another factor that contributes to global outlook instability in REY supply stems from REY over-production in China. 88% of Chinese REY exports come from three domestic mining provinces: Baotou, Sichuan and Jiangxi. 83% of China's REY reserves are located at the Bayan Obo mine in the Baotou province. This is the world's largest reserve of REYs. REYs are produced at the Bayan Obo mine as a byproduct of iron ore mining. The chemical composition of the ore mined there is Fe-REE-Nb. From a global outlook perspective, Bayan Obo is where a vast majority of the world's heavy REY supply is generated [23]. Experts are predicting that this

reserve will be heavy REY depleted by 2025 which suggests that there will be global instability

in the supply of HREYs in the near future [24]. Fig. 7 shows the geographic spread of major

REE mining operations throughout the globe.
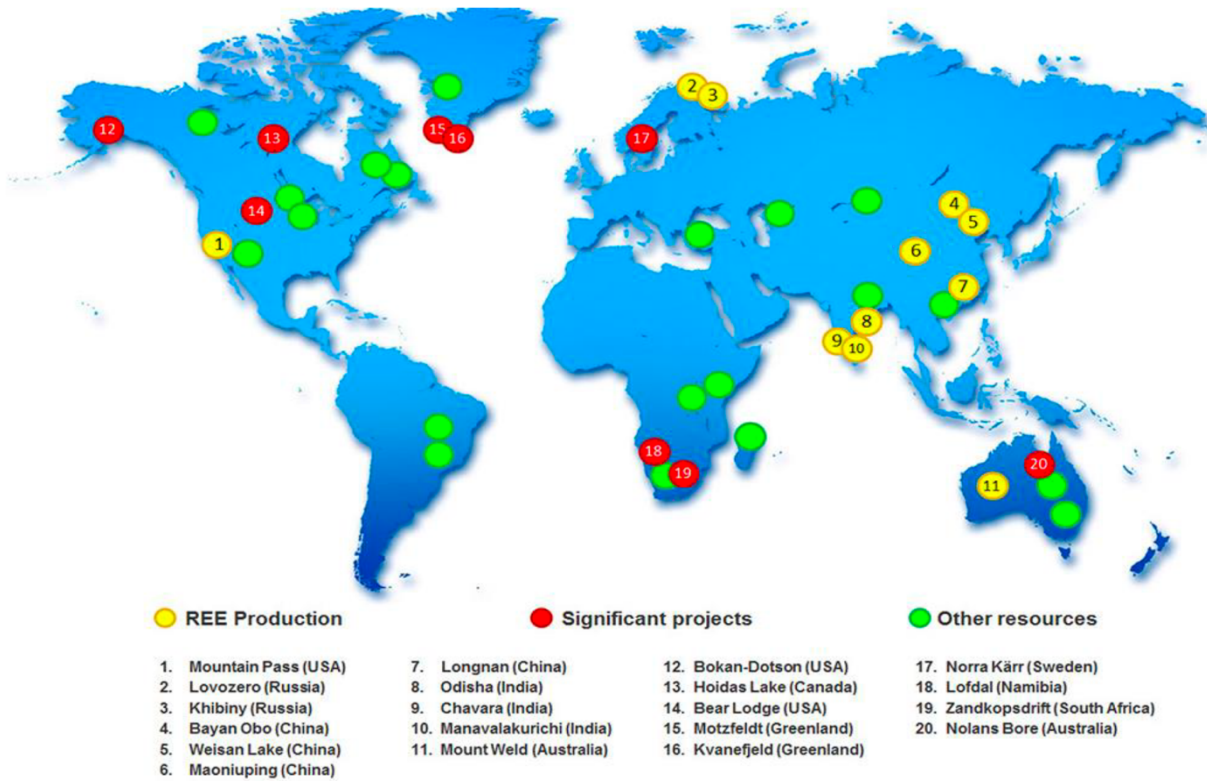
Figure 7. Global REE mines, exploration projects and other resources [20]

2.7. REYs in Coal and Coal Byproducts

It is previously mentioned in this text that REY concentration can be enriched in some

coals and sediments to a greater degree than that of the Earth's crust. Total REY content > 0.1%

has been found in some coal seams and basins which is considered high in concentration. Coal

resources have not been utilized because conventional methods of mining ore deposits containing

REYs was deemed sufficient in terms of supplying the global market. As discussed in the

previous section, global demand for REYs is increasing while supply is in danger of drastically

waning.

Seredin and Finkelman (2008) suggest that REYs can be concentrated in coal at many formation phases from early peat stages to anthracite stages. Coal processes can be divided into syngenetic, diagenetic, and epigenetic processes. Syngenetic processes take place during the peat accumulation phase. Diagenetic occurs after the peat burial phase and during the coalification steps. Epigenetic processes take place after coal has been compacted and solidified, and this process is stratified on an age basis (lignite, subbituminous, bituminous, etc.). There are many different ways in which REYs are transported into coals. Surface and ground water are the main transport methods in which REYs enter coals. Trace REYs can leach out of ore deposits or volcanic ash material as a means of surface water transport. Ground water in coal basins typically have large concentrations of trace elements and through percolation these trace elements will move through coal ranks to a rank with low permeation. Typically, the higher the coal rank the less permeable it is. Volcanic or cosmic impact events can distribute REYs geographically and wind can displace particulates from these events. Polygenic transport phenomena can cause spikes in REY accumulations. This is generally a side effect of ore-forming processes that take place throughout coal forming phases [25].

Three geological conditions are responsible for creating coals containing REYs during the formation of coal basins. Tuffaceous type coals are formed primarily as a result of volcanic activity. Tuffaceous REY enrichments are accompanied by high concentrations of hafnium (Hf) and zirconium (Zr) and have less Eu than normalized crustal abundances. Minerals that typically form in tuffaceous type coals are zircon, phosphate contain minerals such as apatite and monazite, and aluminophosphates such as crandallite, Infiltration type coals are typically formed from epigenetic percolation transportation of REYs into coal basins where they are adsorbed into organic matrices. This is more common in upper sediment layers and lower rank coal.

Exfiltration type coals are created as a result of the flow of sedimentary hydrothermal solutions or deep-water solutions to coal basins at the peat accumulation step. Generally, the organic matrix of coals at this stage can readily absorb trace metals [25].

Seredin and Dai (2012) grouped REYs into light (LREY - La, Ce, Pr, Nd, and Sm), medium (MREY - Eu, Gd, Tb, Dy, and Y), and heavy (HREY - Ho, Er, Tm, Yb, and Lu) categories based on molecular weight [26]. The authors' research shows that coals with LREYs are formed based on tuffaceous or infiltrational mechanisms. Acidic water activity in coal basins and high adsorption properties of the organic matrix in coals are consistent with MREYs [27, 28]. Circulation of marine water, alkaline terrestrial water, and volcanogenic water with high concentrations of HREYs in coal basins are responsible for coals with HREYs [29]. Due to the variety of different enrichment conditions for the three different coal categories, REY distributions in coal are highly variable.

2.8. Occurrence of REYs in Coal

There are three main groupings shown by numerous studies that describes the presence of REYs in coal. One of these groupings is syngenetic minerals from tuffaceous and terrigeneous type formations. These minerals are mostly monazite, apatite, zircon, and xenotime. Diagenetic and epigenetic minerals make up the other mineral grouping, and it is made up of aluminophosphates, sulfates, water-bearing phosphates, oxides, carbonates, and fluorocarbonates. Organic compounds are the third grouping that can house REYs in coal. Literature indicates that REYs will be present in both organic and mineral forms in most coals [26]. The concentrations of REYs in both forms are variable and will depend on numerous features including the environment, coal rank, etc. Typically, LREYs are most commonly found in the minerals in coal. Minerals can be readily collected from the clay sections of coal and in the

partings, margins, or seams. Coals with high concentrations of REYs often have a large amount of its REY content present in the organic matrix. This is common in lignite and subbituminous coals with low ash content. It is commonly observed that there is direct correlation between ash yield and REY concentration in light specific gravity fractions [30, 31, 32, 33, 34, 35, 36, 37]. This empirical evidence is validated by experimental research regarding REY adsorption in coals [38, 39]. It follows that MREYs and HREYs are more commonly associated with the organic matrix of coal due to LREYs being more associated with the mineral content of coal. Seredin and Shpirt(1999) [28] showed that 50% of the REY concentration were found in the organics of coal samples and of that 50% the majority of REYs found were MREYs with direct experimentation. Robert Finkelman's (1993) [40] research found that HREYs were also associated more with the organic matrix of coal rather than minerals.

2.9. Machine Learning Overview

Machine learning is the study of utilizing statistical methods and algorithms to give computers the ability to perform without being explicitly programmed. It is a subcategory in the field of artificial intelligence because it allows computers to learn given data input, identify patterns in data, and make decisions based on the data. All of this is accomplished with a minimum amount of human involvement. Because of growing technological capabilities, machine learning has experienced a boom in popularity in recent years. New and innovative machine learning algorithms and software that utilizes these algorithms are being developed at a rapid rate to meet the demand for applicable capabilities.

There are two main types of machine learning algorithms. Algorithms that learn given labeled independent variables called features and labeled dependent variables called targets or labels are called supervised learning algorithms. Examples of this are regression, classification,

decision tree algorithms, and neural networks. There are also algorithms that utilize unlabeled

datasets and recognize patterns and connections within it. These are called unsupervised learning

algorithms. Examples of this are clustering, anomaly detection, neural network, and latent

variable model algorithms.

One of the main concepts in machine learning is splitting a dataset into a portion of data

that will be used to train algorithms and a portion of the data that will be used to test algorithm

performance. Depending on how large the dataset is, the user will want to split a dataset into

training data and testing data into particular portions. Training data will make up the majority of

the original dataset and testing data will always be in the minority. A machine learning algorithm

is fitted to the training dataset and then the test dataset is plugged into the trained algorithm. The

results from the tested algorithm are then subjected to performance testing and validation testing

to determine how well the algorithm works.

Feature engineering is also an important aspect of machine learning. Attributes need to be

at least somewhat correlated to targets or classes. Generally, the more correlation, the better.

Uncorrelated data creates noise in machine learning algorithms, and noise creates bias in the

model. In the next section, more details will be covered regarding how noise affects specific

machine learning algorithms. Attributes that have no relation to the given target need to be

expunged from the machine learning algorithm to prevent this. The user can iteratively select

different combinations and ratios of features in order to obtain the best performing algorithm. In

some cases, features need to be standardized in order to be useful in various algorithms. If

multiple attributes are used in a machine learning algorithm, it is likely that they will need to be

standardized because it is likely that the attributes vary by measurement and order of magnitude.

Sometimes there are inequalities in the size of attribute vectors given a selected group of

attributes used for machine learning. This is most often due to missing values in the dataset. Imputing is a method utilized to fill in the blanks in attribute vectors, making computation possible. Often there are no missing values in a dataset, but in classification problems there might be extreme bias in a training set for a particular class. There are methods to handle this issue by adding values to features that will create more equal class distributions in the training data. One such method will be covered in the methods section of this paper. For most instances, data is not in the correct format to be computed. The data must be cleaned in order for it to be suitable for computation. This is can mean that data types might need to be changed, or 0's need to be changed to null values. There is also a major difference in how numerical values and non-numerical values are computed. Numerical values are computed given they have the correct data type, but categorical values need to be encoded for computation. For example, if there is a class set that is labeled "yes" or "no", those entries need to be converted to 1's for yes and 0's for no.

2.10. Machine Learning Algorithms

This work is focused on a category of supervised learning called classification algorithms. Also called classifiers, these algorithms essentially recognize patterns. Classifiers place instances into categories called classes based on the features from a dataset that are fed into the algorithm. Classifiers are trained with features from the training dataset where their association to the algorithm's classes is already known. The classic application example where this type of algorithm is used is a spam filter. Known spam email data attributes are feed into an algorithm, and it learns from those examples. There are many different types of classifiers, but this research will be focused five particular algorithms: k-nearest neighbors, logistic regression, decision tree classifiers (CART), random forest, and the Adaboost algorithm.

To visualize a dataset, let a dataset have p features and n observations. Below is a matrix of the features in the dataset:

$$X = \begin{pmatrix} 1 & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{np} \end{pmatrix} \qquad \textit{Eq.1}$$

where x_{ij} denotes the jth feature for the ith observation. From this definition, x_i can be shown as:

$$x_i = \begin{bmatrix} 1 \\ \vdots \\ x_{ip} \end{bmatrix} \qquad \textit{Eq.2}$$

This is the representation for a dataset that will be used for the rest of this work.

2.10.1. K-Nearest Neighbors

K-nearest neighbors is a machine learning algorithm that uses the entire dataset to test the algorithm. Algorithms that learn as such are called "lazy learners [41]." A dataset should not be split into training sets and test sets when dealing with k-nearest neighbors. There is no actual learning involved when using this algorithm. When new data is tested, it is added directly to the original dataset and then the updated dataset is fitted to the algorithm. The model assumes that things that are similar are spatially close. The way the model does this is quite simple. The algorithm makes predictions for new observations by looking through the fitted dataset for the k most similar observations called neighbors and defines the class for those k observations. For k-nearest neighbors, k is the only hyperparameter. The classifier is based on the nearest neighbor density estimator in Eq. 3 [42]. Let x be from population G, it follows that $P(x|G) \approx$ (proportion of observations in the neighborhood around x) / (neighborhood volume). The population that relates to the largest value of Eq. 3 is used to classify x.

$$\frac{P(G_i)P(x|G_i)}{\sum_j P(G_j)P(x|G_j)}, i = 1, \dots, k. \qquad \textit{Eq.3}$$

Hyperparameter tuning techniques should be used to determine the value of k. It is important to note that k-nearest neighbors is a nonparametric statistical method. It does not assume any statistical distribution of data. k-nearest neighbors works very well with a small number of input features, but struggles when the number of features becomes very large. The number of input features is the number of dimensions in j-dimensional space. With increasing dimensionality, there is increasing volume in j-dimensional space and k-nearest neighbor's performance is based on proximity. When there are a large number of dimensions, similar observations can be far away from one another which defies the notion of similarity and proximity being directly related. This property is sometimes called the "curse of dimensionality. [43]"

2.10.2. Logistic Regression

A logistic regression classifier, also referred to as logit regression, is a statistical method that utilizes a basic logistic function to predict whether an observation belongs to a certain class or not. If the probability is estimated to be larger than 50% then the observation belongs to one class and if the estimated probability is below 50% it belongs to another class. The logistic function or sigmoid function was developed by statisticians to explain population growth models [44]. It has an S-shaped curve that maps any value between 0 and 1, but those limits are never met.

$$g(x) = \frac{1}{(1 + e^{-x})} \qquad \text{Eq.4}$$

The sigmoid function in Eq. 4 is extremely important in not only logistic regression but in many other forms of machine learning, particularly neural networks.
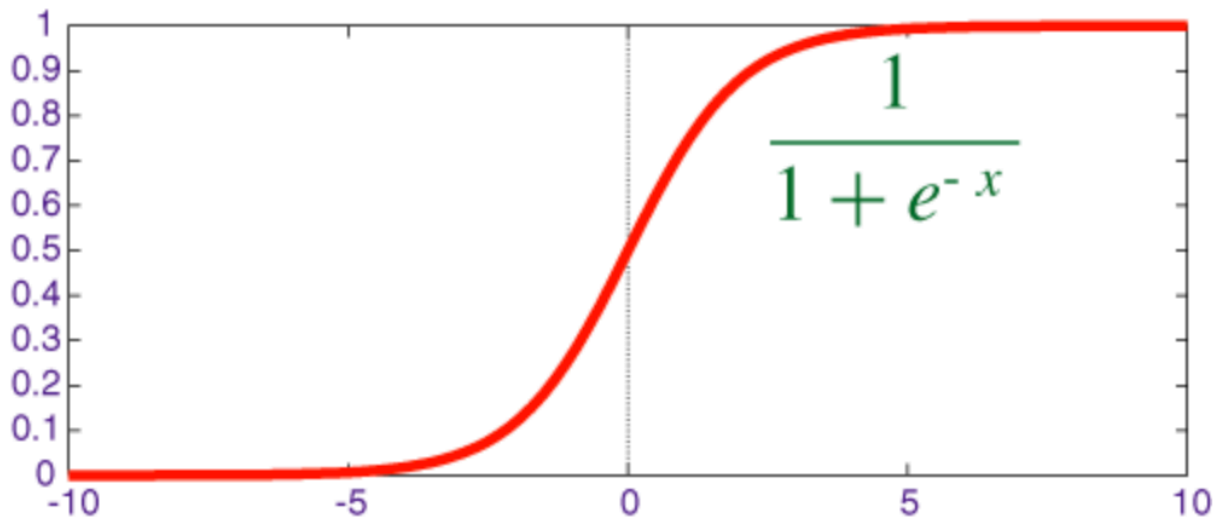
Figure 8. The graph of the sigmoid function [45]

From Fig. 8 it is apparent that g(x) approaches 1 as x approaches ∞ and g(x) approaches 0 as x

approaches -∞. This means that the range of g(x) is between 0 and 1, but g(x) never reaches

either boundary.

Logistic regression is mapped by an equation similar to linear regression. Input values, x,

are plugged into a formula that is weighted by coefficients represented as β.

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \qquad\qquad Eq.5$$

Let h(xᵢ) be the estimation of the iᵗʰ observation called the hypothesis.

$$h(x_i) = \beta^T x_i \qquad\qquad Eq.6$$

Logistic regression uses the sigmoid function with respect to the hypothesis to make

classifications.

$$g(h(x_i)) = \frac{1}{(1 + e^{-h(x_i)})} \qquad\qquad Eq.7$$

20

The outputs of a logistic regression equation are numbers between 0 and 1. In order convert these numbers to classes, conditional probabilities are created. Let $\hat{y}$ be the predicted class resulting from observations plugged into Eq. 7.

$$\hat{y} = \begin{cases} 0, & g\big(h(x_i)\big) < 0.5 = h(x_i) \\ 1, & g\big(h(x_i)\big) \geq 0.5 = 1 - h(x_i) \end{cases} \qquad Eq.8$$

The logistic regression classifier is trained using maximum likelihood of parameters to estimate β values. Likelihood is the probability that the training data supports possible β values. Let y_i be the target for each observation in the training data.

$$L(\beta) = \prod_{i=1}^{n} P(y_i|x_i;\beta) = \prod_{i=1}^{n} [h(x_i)]^{y_i}[1 - h(x_i)]^{1-y_i} \qquad Eq.9$$

L(β) can be more easily calculated by taking the logarithm of the function.

$$\log[L(\beta)] = \sum_{i=1}^{n} y^i \log[h(x_i)] + (1 - y_i)\log[1 - h(x_i)] \qquad Eq.10$$

Logistic Regression's cost function is the inverse of the logarithm of the likelihood function. A cost function is a function that, given test data, will measure the performance of a machine learning algorithm. Cost functions represent the error between predicted values from the training set and the test values as a number. The goal of logistic regression's cost function is to minimize the error associated with estimating the value of β [46].

$$J(\beta) = \sum_{i=1}^{n} -y^i \log[h(x_i)] - (1 - y_i)\log[1 - h(x_i)] \qquad Eq.11$$

Logistic Regression is the most widely used binomial classifier. It does not require much computational power and is easily understood. A very important positive to using logistic regression is that features do not have to be standardized, and no hyperparameters require tuning. As with linear regression, logistic regression performs better when uncorrelated features are

removed from the training data. The major disadvantage associated with logistic regression is that the input data must be linear because the hypothesis is linear. Because logistic regression is a rather simple classifier, it can be outperformed by much more complex algorithms. Due to its simplicity, it is a good algorithm to develop a baseline for using more complex algorithm [47].

2.10.3. Decision Tree Classifiers

Decision Trees are powerful algorithms capable of processing complex datasets. The anatomy of tree is used to visualize how the decision tree classification algorithm works. The first step of the algorithm is to visualize the entire training set as the "roots." Individual observations in the dataset are the "branches," and classes are the "leaves."

The official name for this algorithm is Classification and Regression Tree (CART). This basic mechanism of the algorithm is that initially the training set is split into two subgroups using a single attribute, ϕ, and a threshold, t_ϕ. The algorithm then looks for the values of ϕ and t_ϕ that produces the purest subgroups. Eq. 12 describes the cost function associated with minimizing the CART algorithm.

$$J(\phi, t_\phi) = \frac{m_l}{m} I_l + \frac{m_r}{m} I_r \qquad \qquad Eq.12$$

I_l and I_r are the impurities associated with the left and the right subgroups, and m_l and m_r are representative of the observations in the left and the right subgroups. The total number of instances in the training set is m. Once the training set is split in two, the resulting subgroups are split in two, and this process is repeated recursively until a pre-determined maximum depth is reached or convergence is reached in terms of minimizing impurity. The algorithm can be described as "greedy" because it checks for the best split at each level, and it does not consider other levels in the tree. This usually leads to reasonably good results but does not guaranteed the best results [48].

The metric of most concern in a CART algorithm is the measure of impurity for each split, and the two most popular methods of measuring impurity are the Gini index and entropy. The Gini index aims to measure the probability of a randomly chosen observation being mislabeled in the training set. Eq. 13 describes this criterion. The Gini index has a range from 0 to 1. If all instances belong to a specific class, then the Gini index is 0. The Gini index would be 1 if instances are randomly distributed throughout different classes. If the Gini index is 0.5 then instances are equally distributed in some of the classes.

$$GI = 1 - \sum_{i=1}^{n}(p_i)^2 \hspace{3cm} Eq.13$$

The entropy measure of impurity measures the degree of uncertainty at each depth in the tree classifier. The aim is to reduce the amount of disorder beginning at the root node to the leaf nodes. This is described in Eq. 14.

$$E = \sum_{i=1}^{n} -p_i log_2 p_i \hspace{3cm} Eq.14$$

If all observations in a node are associated with the same class, then entropy is 0. There is maximum entropy when there is uniformity in the class distribution [49].

Generally, decision trees are advantageous because they require less preprocessing of features and they are inherently designed to handle complex datasets. This means that they do not require data to be of a specific distribution or that data need to be standardized. It is also important to note that missing data is not a problem for decision trees. The model is easy to visualize because of the tree description, and the it is not mathematically intensive. This allows it to be easily explained to non-experts.

With complexity comes larger computational demand. Depending on the dataset, decision tree classifiers can require a massive amount of computational power to compute solutions in a

timely manner. This is because the algorithm trains all features on each instance at each node. It also follows from this notion of complexity that a small change to the training set can vastly change the computed results [50].

2.10.4. Random Forest

Random Forest is a bootstrap aggregation method where a collection of decision tree classifiers is assembled, and then majority voting determines the best classifiers. It is amongst the most popular machine learning algorithms because it is easy to train and tune, and it displays superior performance [51]. Bootstrapping is a resampling technique used to estimate statistics of a population where sampling is conducted with replacement. Bootstrap aggregation, also referred to as bagging, is an ensemble machine learning technique that uses many weak models and combines them to make predictions and selects the best prediction. Ensemble methods are machine learning techniques that train multiple models by using the same algorithm. Bagging has a distinct advantage over conventional machine learning techniques because it combines many weak learners to outperform a singular strong learner. These types of algorithms are known for reducing overfitting tendencies by reducing the overall variance in a model which increases accuracy [52].

Random forest is made up of a large amount of decision trees. Every tree in the ensemble produces a prediction and the most predicted class dominates. Random forest increases randomness in the model as more trees are produced. The algorithm looks for the optimum feature amongst a random sampling of features instead of looking for the strongest feature in the node split step. The randomness generated by this approach creates a stronger model [53].

The random forest algorithm is elegantly simple. Fig. 9 and 10 displays the full random forest algorithm visually. The first step is to select samples at random from a training set. Then a

decision tree is built for each sample and prediction is made for each decision tree. Each

predicted result is voted on and the predicted result with the most votes is the final prediction.

Majority voting is carried out by making comparisons between the predicted results and a test

result. If the test result is different than the predicted result then the predicted result does not

receive a vote, and if they match then it receives a vote.
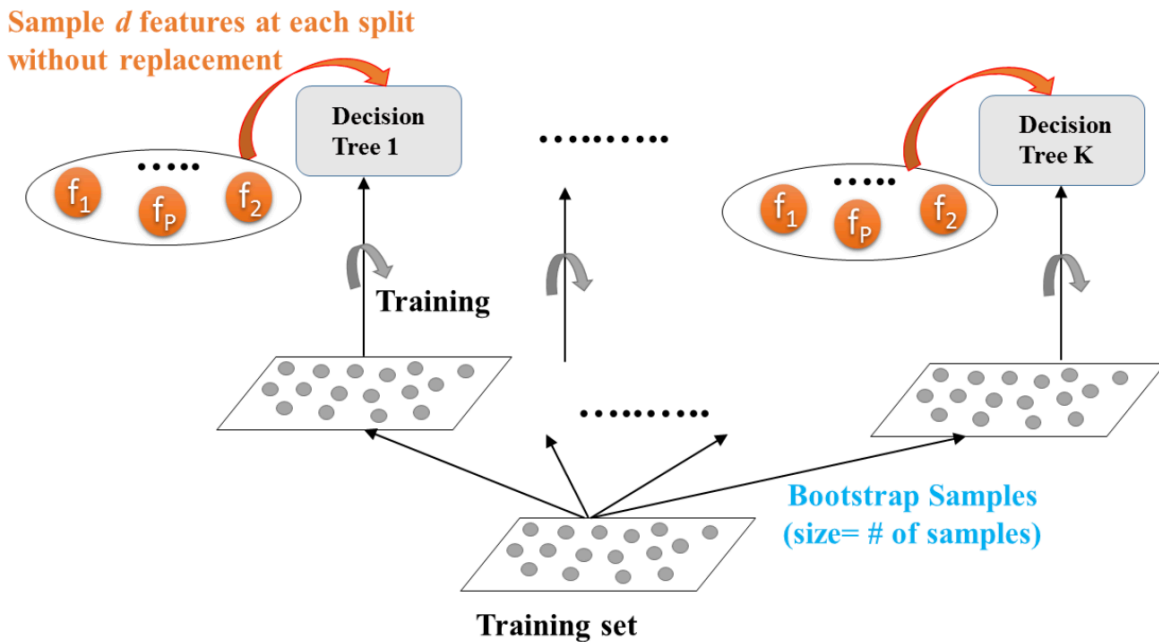
Figure 9. Random forest algorithm visualization [53]
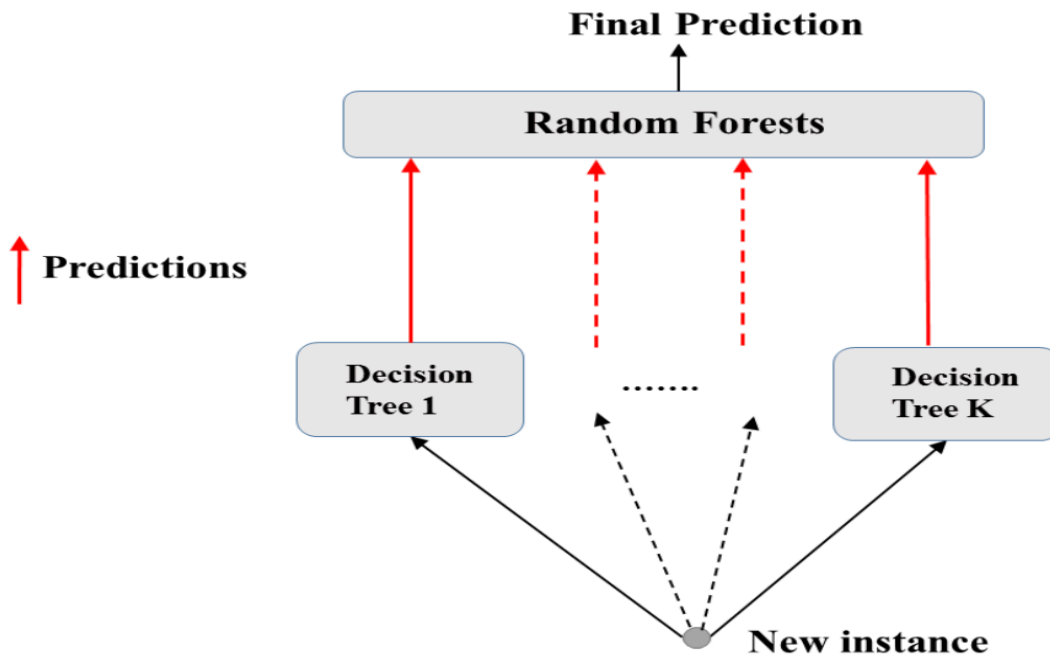
Figure 10. The flow of predictions in random forest [53]

There are many advantages to random forest. It has all the advantages that a decision tree would have except it handles complexity better because it is an ensemble algorithm (many weak models are combined to create a powerful prediction). Random forest creates an unbiased estimate of the general error as trees grow and reduces variance drastically. This can be effective at combating the bias-variance tradeoff. There is an observed disadvantage with this technique. With very noisy data, overfitting tends to occur. This is because random forest is really good at estimating with weak learners. If there is low correlation between features and classes, then the algorithm will learn based on uncorrelated features which creates model bias [54].

2.10.5. Adaboost

Adaboost is the final machine learning algorithm utilized in this work. It stands for adaptive boosting and, like random forest, is an ensemble method. Boosting handles low performance better than bagging algorithms [55]. Remember that bagging is essentially

26

bootstrapping a training set and fitting a decision tree to each bootstrapped sample, and then combining the trees and voting to select a potentially powerful classifier. Boosting is very similar, but decision trees are created using data from previously created decision trees. Boosting does not use bootstrapping. Every tree is fitted to an adapted version of the original training dataset [52].

Each boosting step in the algorithm creates data modifications by assigning weights, α, to every training instance. In the beginning all weights are equal to the inverse of the sample size and thus trained in usual fashion. Each instance is trained iteratively after this and weights are modified individually. The classifier is retrained for each successive weighted observation. Weights are increased if their associated instance was classified incorrectly and decreased if they are classified correctly. It follows that instances that are hard to classify receive larger weights [56]. See Fig. 11 below for a graphic explanation of adaboost.
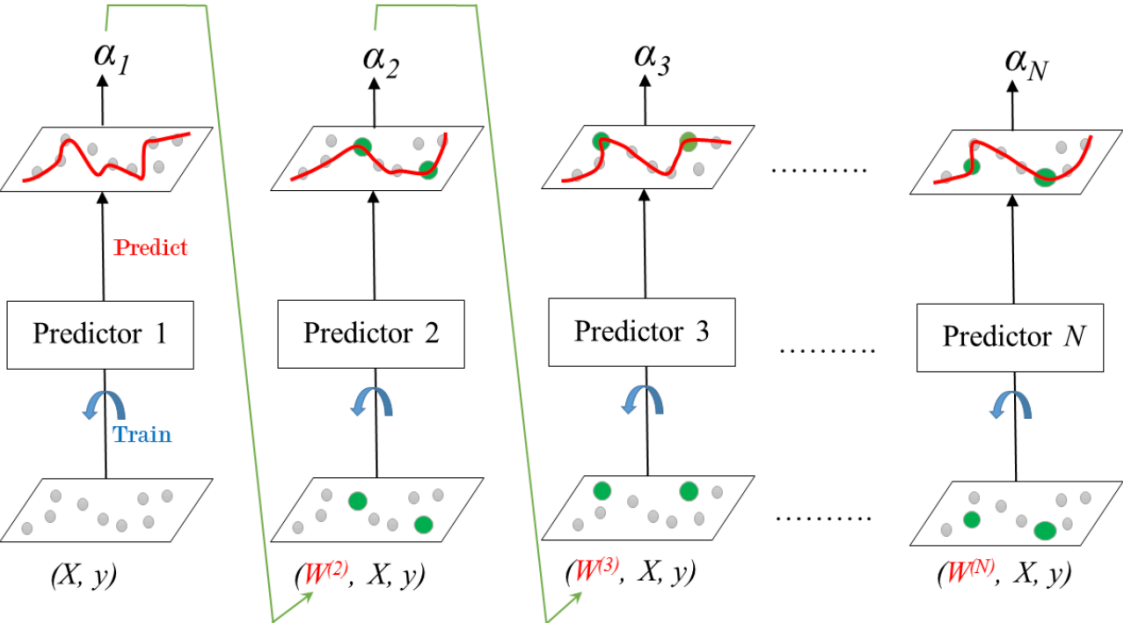
Figure 11. Adaboost algorithm visualization [57]

Adaboost essentially has the same strengths and weaknesses as bagging techniques like random forest. It can be very powerful when dealing with weak learners, but if there is too much noise in the training set, the algorithm has a tendency to overfit. This tendency to overfit is embedded in the fact that adaboost strengthens weak learners by assigning increasing weights to them until the classify correctly. If there is no correlation between features and classes then every instance will be over-trained and therefore subject to bias [58].

2.11. USGS COALQUAL Database

The USGS COALQUAL database version 3.0 was published in 2015 and provides the user access to 7658 coal samples from the United States. These coal samples are considered dry whole coal samples which means they are sampled directly from their environment without being processed. The database has 4 different tables when exported from the USGS website: Sample Description, Proximates, Oxides, and Trace Elements. When combined into one table, there are 279 features that describe each of the 7658 coal samples. Some of the information in the features is missing for certain samples. There are data qualifiers associated with the measurement data from the Proximates, Oxides, and Trace Elements tables. Data qualifiers are attributes that describe data measurements; therefore, for each data measurement there will be a data qualifier associated with it. For purposes of this research, the only data qualifier that is considered is the "L" (less than) data qualifier. This indicates that the measurement taken is less than the detection limits of the analytical testing equipment or technique used [59]. This is important to consider when conducting statistical analysis and utilizing machine learning algorithms. There are two types of data in the COALQUAL, measurements in the form of numerical data called floats and text descriptions in the form a categorical data called texts. It is

important to distinguish between the two because computations are dependent on datatypes. For example, text data cannot be added to a float.

2.12. Software Overview

Software and programming play a crucial role in this work. The four tables from the COALQUAL database are exported from the USGS website to a .csv file. This file stands for "comma separated values and means that all data is separated by commas and rows are ended with a semicolon. The .csv files are then converted to SQL tables. SQL stands for "Structured Query Language." SQL is the most popular language and tool to manage relational databases. A relational database is a grouping of tables where data can be queried or manipulated in various ways without having to change any data in the tables. MySQL is an open-source platform used to manage SQL projects, and is the platform used in this work. A python-SQL connector is used to merge MySQL with Python 3.8. Python is a general, object-oriented, high-level programming language. It was utilized for this project because it is easy to use and learn, and there are open-source, well-developed libraries to manage data and machine learning that are programmed specifically in python. Python was accessed through Anaconda platform which is an open-source package software that provides many useful libraries and programs that utilize python. All python syntax was written and processed in Jupyter Lab, an interactive development environment, accessed in the Anaconda platform.

CHAPTER 3: METHODS

The following chapter provides the methodical approach used in the statistical analysis used to account for variance and bias in the COALQUAL database and how machine learning algorithms are processed in python. Additional information can be located in Appendix A.

3.1. Exploratory Data Analysis

When first examining a dataset, it is important to have a holistic approach. Specific insights can be targeted from general insights. Exploratory Data Analysis (EDA) allows the user to get a big picture visualization of a dataset in order to target specific features for further insight. Since the end goal of the research is to classify coal samples as promising or unpromising, it is important to establish the criteria for a coal sample to be promising or unpromising. In order to do so, the concentrations of the 15 REYs were analyzed for key statistical parameters. Table 1 shows these parameters for each REY.

Table 1. Summary statistics of all REY concentrations measured in parts per million (ppm) for the COALQUAL database

REY	count	mean (ppm)	standard deviation (ppm)	minimum (ppm)	first quartile (ppm)	median (ppm)	third quartile (ppm)	maximum (ppm)
Y	7585	8.93	6.84	0.16	4.79	7.34	11.00	185.00
La	6652	11.70	9.42	0.45	5.81	9.20	14.70	236.00
Ce	6081	23.96	25.45	1.30	10.90	17.40	27.60	506.00
Pr	5601	10.21	8.33	0.30	4.60	8.20	13.00	110.00
Nd	5946	12.32	11.09	0.61	5.74	9.40	15.40	236.00
Sm	5588	2.54	3.78	0.01	1.10	1.67	2.50	68.00
Eu	5626	0.42	0.28	0.03	0.25	0.36	0.51	5.83
Gd	5602	2.91	2.39	0.12	1.40	2.21	3.67	39.70
Tb	5619	1.16	3.76	0.02	0.21	0.31	0.48	47.00
Dy	5607	3.11	2.15	0.16	1.70	2.60	3.90	23.00
Ho	5598	1.03	1.06	0.05	0.48	0.76	1.20	19.00
Er	5603	1.24	0.95	0.03	0.60	1.00	1.60	16.00
Tm	5603	0.63	0.48	0.03	0.34	0.51	0.76	7.70
Yb	7269	1.01	0.68	0.05	0.56	0.87	1.27	9.27
Lu	5587	0.37	0.89	0.01	0.10	0.15	0.22	10.10

It is interesting to note that Ce is has the highest mean in the dataset amongst other REY concentrations, and Lu has the lowest mean. Ce is considered the most abundant REY and Lu is the heaviest of the REYs. Standard deviation is a measure of the magnitude of the variance in a dataset. Ce has the highest standard deviation and Tm has the lowest. Tm is known as one of the rarest of the REYs, and its low standard deviation is indicative that there is not a lot of variance in the amount of Tm found in each sample. Since the mean of Tm is the second lowest, this indicates that small concentrations of Tm are being found in each sample at around the same concentration for each sample. It is important to note that each REY in this dataset has close to the same number of entries. This does not align with literature, where there is a large degree of variation between number of Ce and number of Tm samples recorded. The original dataset is not representative of this since Tm does not have the smallest number of samples or the smallest mean. There is evidence to suggest that the database is biased toward the rarer REYs. Outlier analysis like this is very important in the EDA process.

A potential source of variance and bias in the dataset comes from data with L qualifiers (samples below detection limits with the values reports as the lower detection limit). Below in Table 2 are summary statistics of all the data in COALQUAL without L qualifiers in terms of REY concentration.

Table 2. Summary statistics of all REY concentrations without L qualifiers in the COALQUAL database.

REY	count	mean (ppm)	standard deviation (ppm)	minimum (ppm)	first quartile (ppm)	median (ppm)	third quartile (ppm)	maximum (ppm)
Y	7560	8.94	6.83	0.26	4.80	7.36	11.00	185.00
La	6160	11.19	9.06	1.00	5.50	9.07	14.20	236.00
Ce	5557	20.69	17.42	1.30	10.60	16.80	25.80	506.00
Pr	948	6.48	7.32	0.31	1.61	3.17	9.36	67.50
Nd	4303	13.36	11.76	0.61	6.52	10.40	16.60	236.00
Sm	5103	1.94	1.42	0.03	1.08	1.63	2.34	19.90
Eu	5270	0.43	0.28	0.03	0.25	0.37	0.52	5.83
Gd	1670	2.80	2.75	0.28	1.24	1.94	3.38	39.70
Tb	4878	0.33	0.22	0.02	0.20	0.28	0.40	4.08
Dy	717	3.39	2.46	0.36	1.74	2.70	4.33	19.20
Ho	351	0.75	0.56	0.12	0.37	0.62	0.97	4.59
Er	1070	1.54	1.08	0.18	0.83	1.29	1.93	11.20
Tm	42	0.44	0.41	0.07	0.20	0.28	0.46	1.99
Yb	7222	1.02	0.68	0.07	0.56	0.87	1.28	9.27
Lu	4945	0.16	0.24	0.01	0.10	0.14	0.19	10.10

After eliminating L qualifiers from the dataset, it is apparent that the dataset changes rather significantly. Pr, Dy, Ho, Er, and Tm all show significant decrease in sample size, meaning that those particular REYs are mostly comprised of data with L qualifiers. There are also various changes in mean concentrations and sample deviations associated with eliminating data with L qualifiers, most notably a decrease in both for Tb, Ho, and Lu. Those REY parameters are hypersensitive to changes in L qualifier data.

If possible, it is always good to compare dataset to standard reference datasets that have been studied extensively. The National Coal Resources Data System (NCRDS) dataset collected by Robert Finkelman is widely considered the standard reference dataset for REY concentrations. It has been well-researched and widely referenced [60]. Below in Table 3 are some summary statistics of the NCRDS dataset found in literature.

Table 3. Summary statistics of the NCRDS database [61]

REY	count	mean (ppm)	standard deviation (ppm)	maximum (ppm)
Y	7897	8.5	6.7	170
La	6235	12	16	300
Ce	5525	21	28	700
Pr	1533	2.4	-	65
Nd	4749	9.5	-	230
Sm	5151	1.7	1.4	18
Eu	5266	0.4	0.33	4.8
Gd	2376	1.8	-	39
Tb	5024	0.3	0.23	3.9
Dy	1510	1.9	2.7	28
Ho	1130	0.35	-	4.5
Er	1792	1	1.1	11
Tm	365	0.15	-	1.9
Yb	7522	0.95	-	1.9
Lu	5006	0.14	0.1	1.8

It is apparent that the data more closely resembles that of the COALQUAL database without L qualifiers. The most notable similarity is the in the number of samples associated with Pr, Dy, Ho, Er, and Tm. In the original dataset there is no significant difference between the number of samples for each REY, but in the dataset with no L qualifiers and the NCRDS dataset there are less samples for these particular REYs. This aligns with literature regarding the rarity of REYs.

To further evaluate data quality, it is good practice to normalize data with a standard dataset for REY concentrations. Essentially what normalization does is allow the user to compare features in a dataset to established measures and visualize differences. For this work, the Upper Continental Crust (UCC) dataset, seen in Table 4, is used to normalize the three datasets above.

Table 4. The UCC mean REY concentration standard dataset [62]

REY	Mean (ppm)
Y	22
La	30
Ce	64
Pr	7.1
Nd	26
Sm	4.5
Eu	0.88
Gd	3.8
Tb	0.64
Dy	3.5
Ho	0.8
Er	2.3
Tm	0.33
Yb	2.2
Lu	0.32

When the normalized REY distributions are plotted, a smooth curve indicates reliability in the dataset whereas zig-zag fluctuations indicate unreliability. Figure 12 shows the REY distributions of each dataset normalized by the UCC.
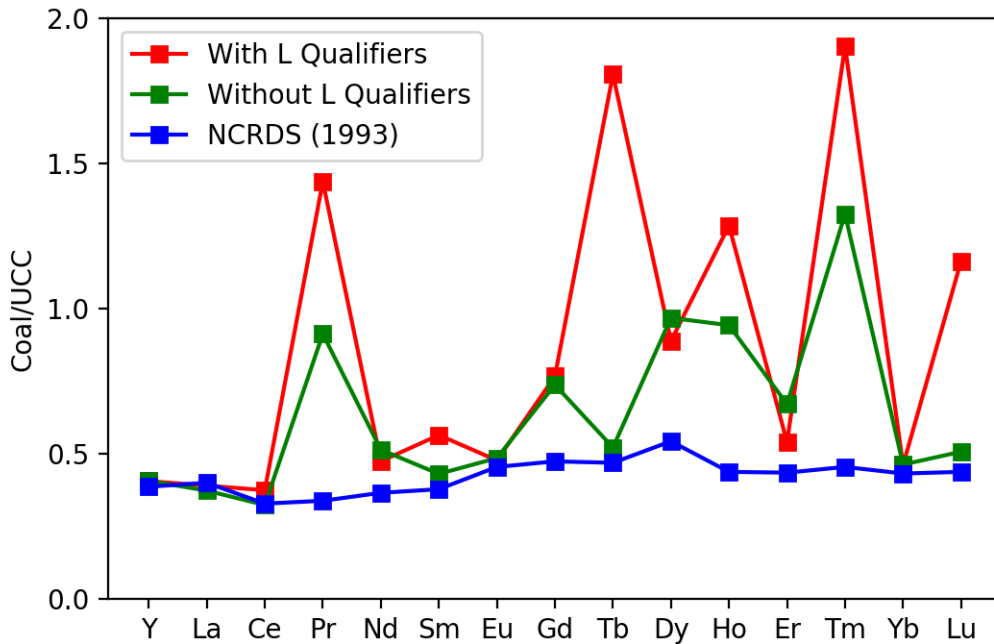
Figure 12. UCC normalized mean REY concentrations from the COALQUAL database and NCRDS database.

From the plot above, the normalized NCRDS distribution shows a very smooth curvature indicating reliability. This aligns with the NCRDS reputation as the standard reference for REY concentrations amongst researchers and scholars. The COALQUAL database with L qualifiers has some definite points of unreliability. The REYs associated with unreliability are Pr, Gd, Tb, Dy, Ho, Tm, and Lu. Most of these are REYs with a large portion of L qualifiers associated with them. The extremity of the L qualifier effect is dampened when samples with L qualifiers are eliminated, but that distribution is not even close to as reliable as the NCRDS distribution is. Therefore, it is reasonable to conclude that there is an unreliable quality associated with the COALQUAL database.

3.2. The Major Source of Unreliability in the COALQUAL Database

There are several factors that can affect the reliability of the data in the COALQUAL database. Instrumentation and analysis techniques have a large role in measuring concentrations of REYs in samples. Sophistication of technique and technology is often correlated with accuracy in measurement. Data in the COALQUAL database is measured using three techniques: ESA-1, ESA-2, and INAA. ESA stands for Emission Spectrographic Analysis. ESA-1 is a semi-quantitative 6-step emission spectrographic analysis measures the concentration of trace elements in a sample on an ash-basis in ppm. ESA-2 is also semi-quantitative but can be described as an automatic plate reading computer-assisted emission spectrographic analysis. INAA stands for Instrumental Neutron Activation Analysis and is the only purely quantitative approach out of the three. Each technique has a unique limit of detection (LOD). Generally, the lower the LOD, the higher the precision and accuracy [60]. Each REY in the COALQUAL database uses two of these methods to measure concentration depending on when the test was conducted and where the test was conducted. A REY was either measured by a combination of ESA-1 and ESA-2 for various samples or ESA-2 and INAA. The semi-quantitative techniques use a combination of coefficients and confidence intervals to account for error. Because these techniques rely on statistical computation for results, they are not as reliable as INAA (). INAA also has lower inherent detection limits than ESA-1 and ESA-2. Therefore, it can be concluded that INAA is more accurate than ESA-2 and even more accurate than ESA-1.

One potential cause of the large portion of data with L qualifiers associated with certain REYs is explained by the Oddo-Harkins effect. This principle states that as elements become heavier, they tend to sink to lower depths in the Earth's crust and therefore are found in lower crustal abundances. Due to the Oddo-Harkins effect, heavier REYs are typically present in lower concentrations, often below the LODs and therefore have large amounts of data with L qualifiers.

This could be an explanation as to why the number of samples changes drastically for the

HREYs (Dy, Ho, Er, Tm, and Lu) when samples with L qualifiers are removed from the original

dataset [62].

Table 5. Percentage of samples below detection limit (BDL) or above detection limit

(ADL) for each analysis technique for each REY.

| | Analysis Technique | | | | | |
| | ESA-1 | | ESA-2 | | INAA | |
REY	BDL	ADL	BDL	ADL	BDL	ADL
Y	0.3	32.2	0	67.5	-	-
La	-	-	7.4	16	0	76.6
Ce	-	-	8.5	7.4	0.1	84
Pr	8.1	0.5	75	16.4	-	-
Nd	9.1	5.1	18.5	67.3	-	-
Sm	-	-	8.1	1.7	0.5	89.7
Eu	-	-	6	3.1	0.3	90.6
Gd	7.4	1.1	62.8	28.7	-	-
Tb	-	-	9.1	0	4.1	86.8
Dy	7.8	0.8	79.4	12	-	-
Ho	8.4	0.1	85.3	6.2	-	-
Er	8	0.5	72.9	18.6	-	-
Tm	8.5	0	90.8	0.7	-	-
Yb	-	-	0.1	29.7	0.5	69.7
Lu	-	-	9.1	0.1	2.4	88.4

Table 5 shows that data for Y, Pr, Nd, Gd, Dy, Ho, Er, and Tm were measured by ESA-1

and ESA-2, and La, Ce, Sm, Eu, Tb, Yb, and Lu were measured by ESA-2 and INAA. For data

measured by ESA-1 and ESA-2, most of the ESA-2 samples were BDL. Pr, Gd, Dy, Ho, Er, Tm

were all subjected to unreliability from Fig. 12, and also exhibit the highest percentages of BDL.

Given that the ESA techniques are the least reliable and there are a great number of samples that

are BDL it is easy to see that unreliability stems from L qualifier data. Most of the REYs tested

by ESA-2 and INAA have a large portion of their concentrations measured by INAA and are

ADL, due in part to the lower detection limits of the INAA. These REYs had the least amount of data with L qualifiers and is therefore more reliable.

3.3. Correcting for Data with L Qualifiers and Addressing Sample Bias

It is clear that data with L qualifiers is the main source of unreliability in the COALQUAL database. This needs to be mitigated in order to continue statistical analysis and for machine learning purposes. One way to do this is to change the value of the coefficient associated with the L qualifier. As discussed earlier, when a value has a L qualifier associated with it, it is multiplied by a value of 0.7. The reasoning behind doing so is based on research performed by Connor et al. (1976). The research suggests that multiplying reported LODs by any coefficient between 0 and 1 has negligible consequences to statistics except when data with L qualifiers makes up a significant portion of samples associated with a REY [63]. For Pr, Gd, Dy, Ho, Er, and Tm, more than 70% of concentration data has an L qualifier. This means that for these REYs, too much of the data is below the LOD. One way to mitigate this is to assign a new coefficient or weight to each sample with an L qualifier. This has been done for the COALQUAL database in research performed by Lin et al. Lin et al. assigned each REY concentration with an L qualifier a new coefficient called the qualifier factor, Q [60]. Each concentration with an L qualifier was multiplied by a potential qualifier factor from 0 to 1 by 0.1 intervals (i.e. 0, 0.1, 0.2, …, 1). This yields 11 groups of mean concentrations for each REY. Then the mean, standard deviation, and relative standard deviation (RSD) of the groups was calculated for each REY. Based on those observations, a qualifier factor was chosen for each REY in order to create a smooth curve for the distributed means of each normalized REY concentration. However, qualifier factors were not chosen to align with the NCRDS directly with the data. In fact, the distribution given by Lin et al. is not as smooth as the NCRDS and

therefore is subjected to more unreliability. In this work, weights will be given to each REY concentration and the same methodology will be used to visualize the effects of the weights given. The major adjustment will be that weights will be calculated to fit the NCRDS distribution as closely as possible. The reason this is change is being made is because the NCRDS mean concentration data for REYs is the most researched and well documented work in this area. Speculative data should be held to that standard as a benchmark for further research until more reliable data is captured.

For each REY concentration, a new weight will be assigned to it simply called the L weight (W). The same range of W's are assigned to each REY concentration as Lin's work. The mean, standard deviation, and relative standard deviation are called in the same manner. The effects of which are shown below in Fig. 13.
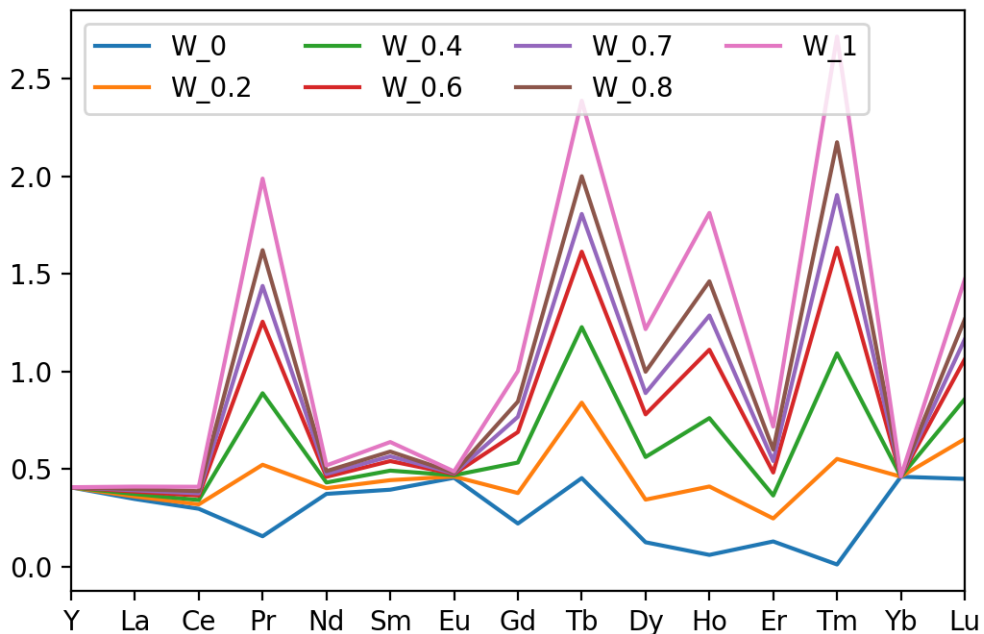
Figure 13. The effect of various values of W on the REY mean concentration distribution
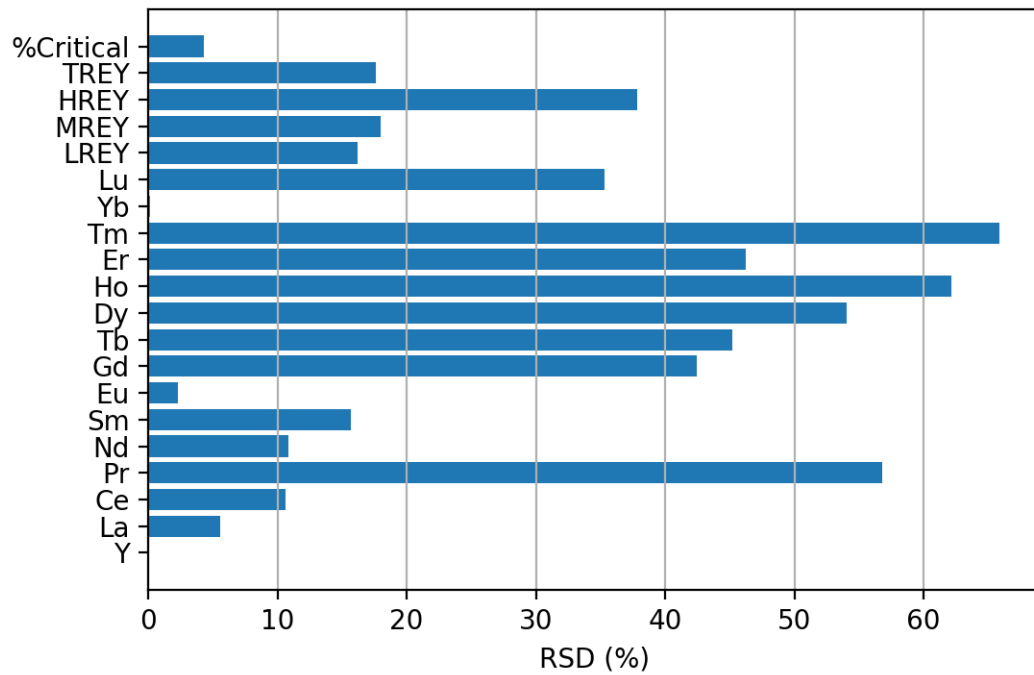
Figure 14. RSD percentage for each REY and specified REY groupings

The RSD percentage is the ratio of the standard deviation to the mean of a population. It is a good measure of variance for each REY or REY grouping. Fig. 14 shows that HREYs are more affected by the values of W than MREYs or LREYs. This is likely due to their naturally lower concentration relative to the lighter REYs as discussed earlier, i.e. the Otto-Harkins effect. The COALQUAL database has a lot of HREY samples that are below LOD. Fig. 15 shows this representation with respect to RSD.
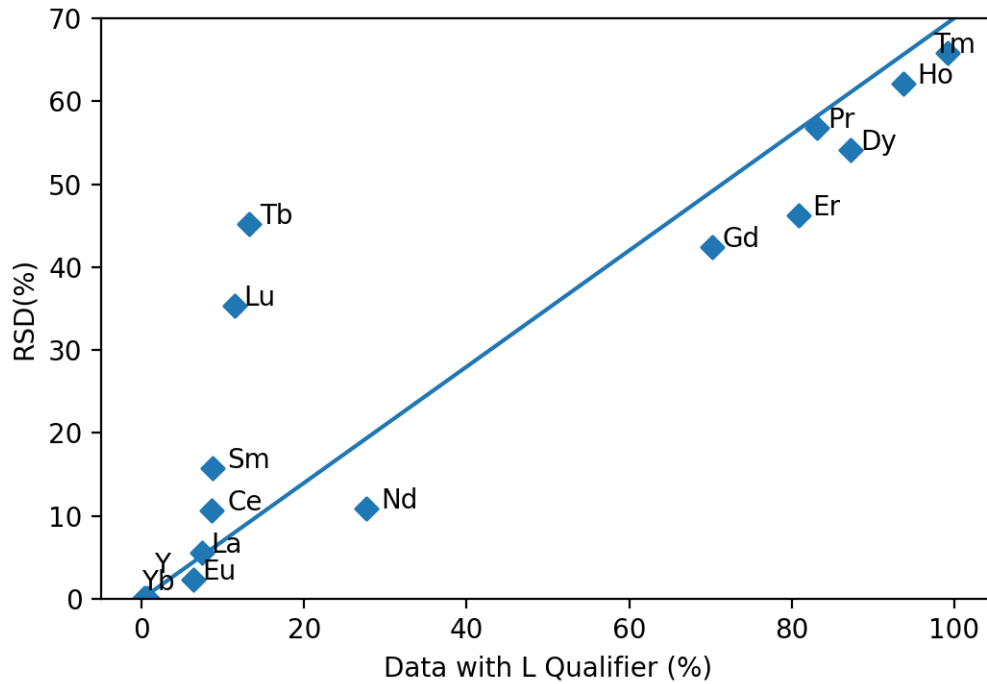
Figure 15. The effect that percent of L data for each REY has on each respective RSD.

Based on the relationship in Fig. 15, the standardized variance of Pr, Gd, Dy, Ho, Er, and Tm is strongly correlated with an increase in the percentage of data with an L qualifier. If the ratio of the mean of concentrations with an L qualifier to the ratio of the mean of concentrations without an L qualifier is observed in relation to RSD, a similar effect is observed. See Fig. 16 below.
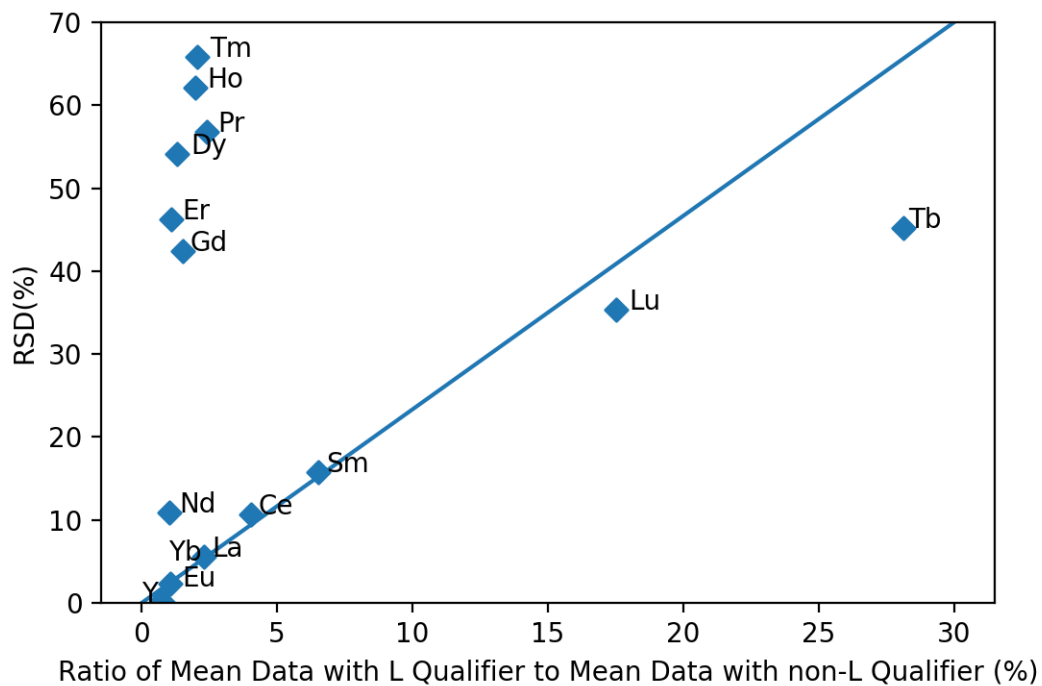
Figure 16. The effect that the ratio of mean concentrations with L qualifiers to mean concentrations without L qualifiers for each REY has on each respective RSD.

When this relationship is viewed, the same outliers emerge. This is evidence to suggest that weighting each sample with an L qualifier with a specified weight for each REY is a viable option for standardizing the REY distribution with the NCRDS REY mean concentration distribution. This process is done iteratively, selecting different weights for each REY and observing how the distribution changes. Weights were chosen that aligned the distribution with the NCRDS distribution. The result is seen below in Fig. 17.
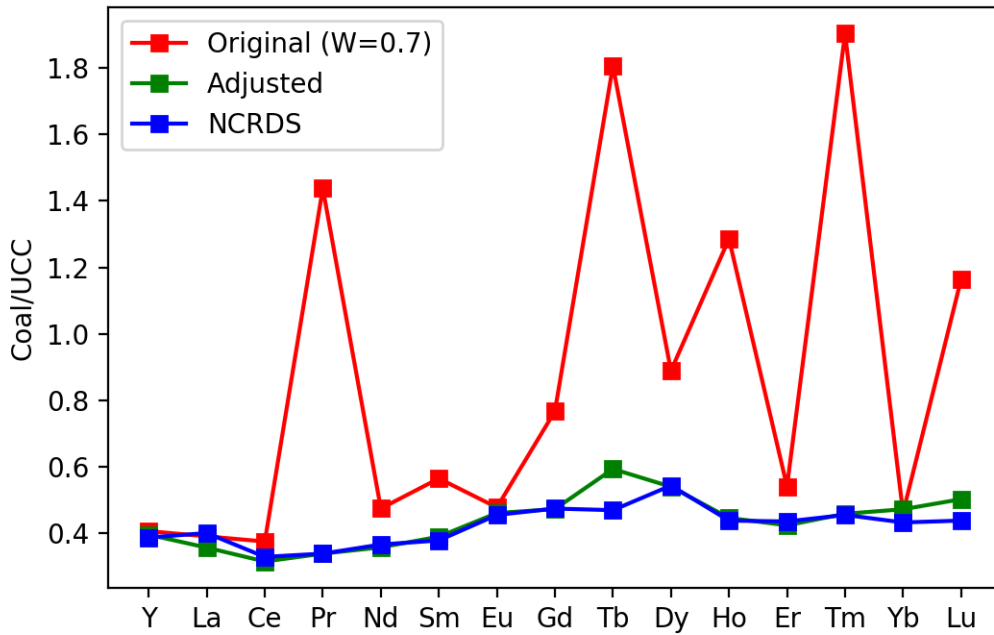
Figure 17. The results of assigning each specified values of W to each REY

This distribution more closely aligns with the NCRDS distribution and is therefore more reliable. The values of W and error percentage calculated by taking the difference between the original and adjusted means of each REY by the adjusted mean of the TREY concentration is given in Table 6 below.

Table 6. Selected values of W for each REY and calculated error percentage

Element	W	REY Original Means	REY Adjusted Means	Error%
Y	0.7	8.9	8.7	0.3
La	0.5	11.7	10.7	1.7
Ce	0.4	24	20.1	6.5
Pr	0.11	10.2	2.4	12.9
Nd	0.2	12.3	9.3	5
Sm	0.2	2.5	1.8	1.2
Eu	0.7	0.4	0.4	0
Gd	0.5	2.9	1.8	1.8
Tb	0.07	1.2	0.4	1.3
Dy	0.5	3.1	1.9	2
Ho	0.3	1	0.4	1
Er	0.7	1.2	1	0.3
Tm	0.19	0.6	0.2	0.7
Yb	0.7	1	1	0
Lu	0.05	0.4	0.2	0.3
LREY		60.7	44.3	27.2
MREY		16.5	13.2	5.5
HREY		4.3	2.8	2.5
TREY		81.5	60.3	35.2

These W values are much different than the values of the L qualifier. Table 6 shows the relationship of those W values and the effects they have on REY mean concentrations. As mentioned previously the value of the L qualifier is 0.7 in the COALQUAL database. As predicted, the error percentage associated with each REY mean concentration is not significant. The most error present with these L weights are in the Total REY (TREY) category. There is a 34.5% difference between the original and adjusted concentrations. The TREY percent error for the work of Lin et al was 24.5%, and therefore, there is significant difference between the qualifier factors from Lin et al. and the L weights calculated in this work.

The subject of bias in relation to coal in the United States needs to be addressed. Sample bias can be defined as

$$\text{Sample Bias} = \frac{\left[\dfrac{(Number\ of\ coal\ samples\ from\ a\ state)}{(Total\ number\ of\ coal\ samples\ in\ COALQUAL)}\right]}{\left[\dfrac{Coal\ reserve\ in\ a\ state}{The\ total\ reserve\ in\ the\ US}\right]} \qquad Eq.15$$

The coal reserve data is obtained from the EIA (2017) [64]. In Figure 18, the Sample Bias of
REY can be seen for each respective state.
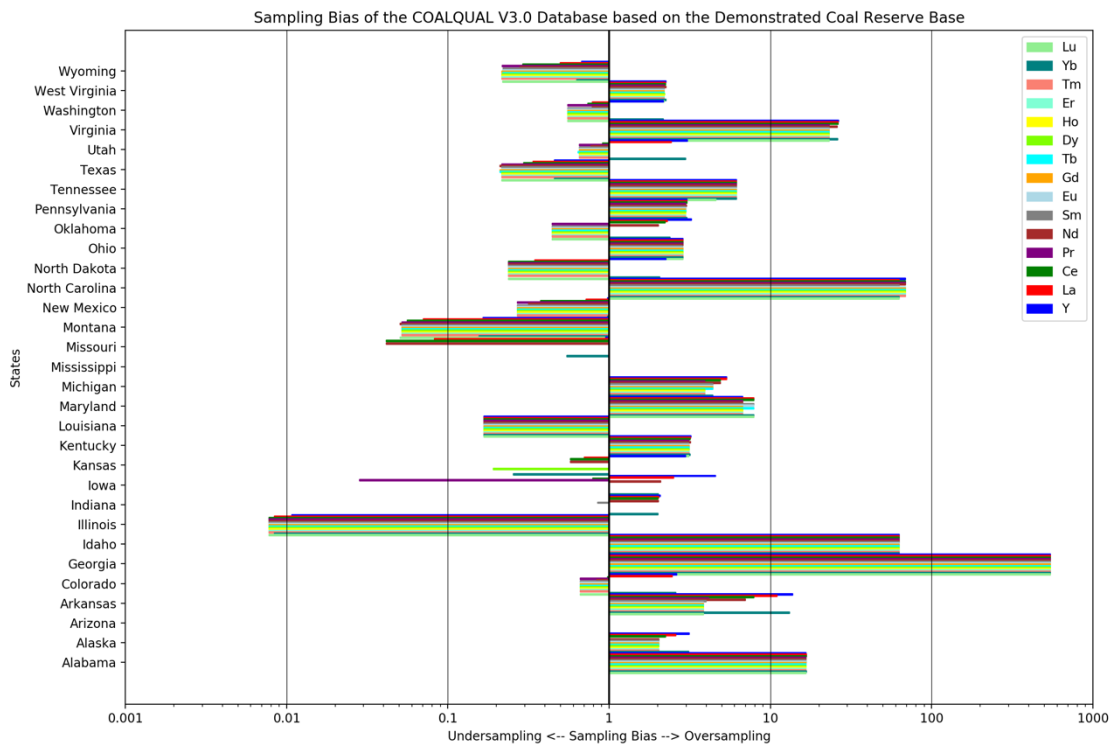
Figure 18. Sample Bias for each REY in each state.

From the visualization, most states are oversampled, especially Georgia, Idaho, Alabama, North

Carolina, and Utah. Illinois is severely under-sampled along with Missouri, Wyoming, and

Texas. It will be important to keep this sample bias in mind as this work proceeds. Bias in the

COALQUAL database is almost certainly due to most mining activity at the time of collection being dedicated in the Appalachian region of the United States.

3.4 Classifying Coal Samples as Promising or Unpromising

It is apparent from utilizing L weights to smooth out COALQUAL's normalized mean REY concentration distribution, that the adjusted COALQUAL data can be further used to reliably classify coal samples from geographical and geological categories. Unequal sample size must be addressed, then COALQUAL must be broken down in geographical and geological categories, and criteria for classification must be established.

One factor that is of great significance is sample size for each REY concentration. For further calculation, only samples with complete REY profiles need to be considered. For each REY, there are some samples with some missing data. The standard approach to addressing this problem is to eliminate all samples with missing REY concentrations and compare to the distribution with missing samples to verify the reliability of each. When missing values are removed, the sample size is reduced to 5485 samples. First, L weights are going to be eliminated in this process so a comparison must be made between the original means of REY concentrations and the adjusted means with L weights added. Table 7 shows this comparison.

Table 7. Original and adjusted means for each REY and percentage of samples with L weights for the equal sample size.

REY	Percent W	Original Mean	Adjusted Mean
Y	0.1	8.7	8.7
La	0.1	10.7	10.7
Ce	2.7	20	20.1
Pr	83.2	10.3	2.4
Nd	26.4	11.5	9.3
Sm	8.6	2.5	1.8
Eu	6.4	0.4	0.4
Gd	70	2.9	1.8
Tb	12.9	1.2	0.4
Dy	87.1	3.1	1.9
Ho	93.7	1	0.4
Er	81	1.2	1
Tm	99.3	0.6	0.2
Yb	0.5	1	1
Lu	11.4	0.4	0.2
TREY	-	75.5	60.3

From comparing Table 7 with Table 6, there is no large difference between the original and adjusted means for each REY when samples with missing data are removed. The distribution for each REY can be seen in Fig. 19. There is no large difference between the distribution with missing values and the distribution without missing values.
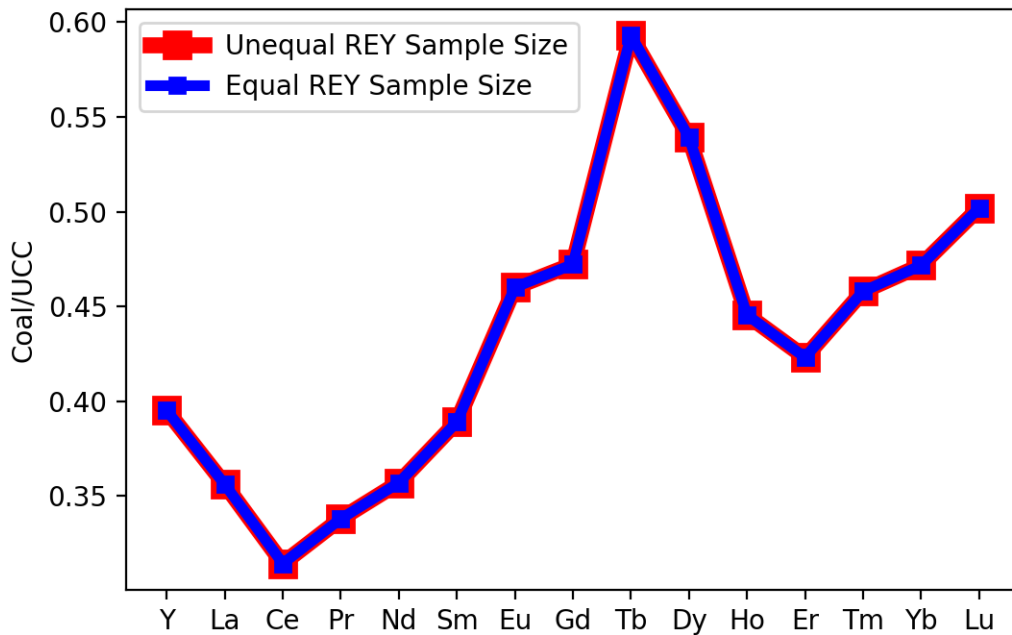
Figure 19. Normalized distributions of the unequal sample size and the equal
sample size.

Next, all samples in COALQUAL must be organized into geographical and geological
categories. Coal provinces, regions, states, rank, and geological age were chosen as the
categories that would yield the most insight. To view the results for categories involving
the original samples view Tables 8, 10, 12, 14, and 16. To view the results for categories
involving the adjusted samples view Tables 9, 11, 13, 15, and 17.

Table 8. Coal provinces with original samples.

Coal Province	No. of samples	Mean ash yield, dry (wt%)	LREY original mean	MREY original mean	HREY original mean	TREY original mean	%Dev original
Alaska	82	15.55	52.15	19.87	4.48	76.50	27.36
Eastern	4358	12.16	54.21	15.12	3.99	73.32	22.07
Gulf	70	21.91	108.32	44.11	10.47	162.90	171.20
Interior	366	11.81	48.05	15.26	3.90	67.22	11.91
Northern Great Plains	224	11.04	48.61	18.89	5.58	73.07	21.66
Pacific Coast	12	36.51	178.11	70.88	20.48	269.47	348.63
Rocky Mountain	373	14.26	60.26	22.08	6.10	88.44	47.24
Total Mean	-	17.61	78.53	29.46	7.86	115.85	-
Total RSD (%)	-	51.73	62.05	70.61	76.42	65.09	-

Table 9. Coal provinces with adjusted samples.

Coal Province	No. of samples	Mean ash yield, dry (wt%)	LREY adjusted mean	MREY adjusted mean	HREY adjusted mean	TREY adjusted mean	%Dev adjusted mean
Alaska	82	15.55	40.04	15.31	2.41	57.76	3.84
Eastern	4358	12.16	45.93	12.90	2.69	61.52	2.42
Gulf	70	21.91	72.08	33.40	5.50	110.98	84.76
Interior	366	11.81	35.97	12.26	2.30	50.53	15.87
Northern Great Plains	224	11.04	27.18	10.57	2.12	39.86	33.64
Pacific Coast	12	36.51	96.71	39.50	9.53	145.74	142.63
Rocky Mountain	373	14.26	36.66	13.51	2.59	52.76	12.16
Total Mean	-	17.61	50.65	19.64	3.88	74.16	-
Total RSD (%)	-	51.73	48.88	59.62	70.97	52.51	-

Table 10. Coal Regions with original samples.

Coal Region	No. of samples	Mean ash yield, dry (wt%)	LREY original mean	MREY original mean	HREY original mean	TREY original mean	%Dev original mean
Central Appalachia	1633	9.80	49.53	13.24	3.63	66.39	10.53
Eastern	347	11.55	45.91	14.44	3.64	63.99	6.53
Fort Union	64	12.19	50.89	22.23	6.60	79.72	32.72
Green River	150	10.59	39.01	13.84	3.45	56.29	6.28
Northern Appalachia	1674	13.37	53.71	15.99	4.10	73.80	22.86
Powder River	160	10.58	47.70	17.55	5.17	70.42	17.23
San Juan River	62	19.57	78.97	26.51	6.40	111.88	86.26
Southern Appalachia	991	13.71	61.63	16.53	4.35	82.51	37.37
Uinta	89	14.08	62.54	23.35	7.34	93.22	55.19
Others	285	17.78	79.48	29.78	7.58	116.84	94.53
Mean	-	13.32	56.94	19.34	5.23	81.51	-
RSD (%)	-	23.93	23.98	29.77	30.96	24.76	-

Table 11. Coal Regions with adjusted samples.

Coal Region	No. of samples	Mean ash yield, dry (wt%)	LREY adjusted mean	MREY adjusted mean	HREY adjusted mean	TREY adjusted mean	%Dev adjusted mean
Central Appalachia	1633	9.80	42.74	11.50	2.54	56.78	5.46
Eastern	347	11.55	34.94	11.95	2.24	49.12	18.23
Fort Union	64	12.19	23.78	11.57	2.25	37.60	37.40
Green River	150	10.59	27.29	9.09	1.50	37.88	36.93
Northern Appalachia	1674	13.37	44.69	13.57	2.71	60.97	1.50
Powder River	160	10.58	28.53	10.17	2.06	40.76	32.14
San Juan River	62	19.57	55.44	18.41	3.29	77.14	28.43
Southern Appalachia	991	13.71	52.49	13.92	2.87	69.28	15.33
Uinta	89	14.08	32.68	13.24	3.03	48.95	18.51
Others	285	17.78	53.15	20.83	3.68	77.66	29.29
Mean	-	-	13.32	39.57	13.42	2.62	55.61	-
RSD (%)	-	-	23.93	29.55	27.06	24.38	27.55	-

Table 12. States with original samples

State	No. of samples	Mean ash yield, dry (wt%)	LREY original mean	MREY original mean	HREY original mean	TREY original mean	%Dev original mean
Alabama	953	13.84	61.84	16.60	4.38	82.81	37.87
Alaska	82	15.55	52.15	19.87	4.48	76.50	27.36
Colorado	167	13.54	56.31	20.90	5.97	83.19	38.49
Indiana	115	10.77	47.10	13.72	3.80	64.62	7.58
Kentucky	954	10.64	50.32	13.91	3.81	68.05	13.29
Maryland	55	16.01	60.42	15.29	4.56	80.27	33.64
Montana	94	12.04	57.77	22.73	7.14	87.65	45.92
Ohio	679	12.77	47.06	15.36	4.07	66.49	10.70
Pennsylvania	825	14.29	61.69	17.48	4.20	83.37	38.80
Virginia	446	9.86	47.91	12.22	3.19	63.32	5.41
West Virginia	577	9.60	47.91	13.53	3.69	65.13	8.43
Wyoming	198	11.36	51.04	19.50	5.35	75.89	26.35
Others	340	16.44	71.84	25.81	6.84	104.50	73.98
Mean	-	-	12.83	54.87	17.46	4.73	77.06	-
RSD (%)	-	-	18.19	13.77	23.31	26.06	15.29	-

Table 13. States with adjusted samples

State	No. of samples	Mean ash yield, dry (wt%)	LREY adjusted mean	MREY adjusted mean	HREY adjusted mean	TREY adjusted mean	%Dev adjusted mean
Alabama	953	13.84	52.56	13.96	2.88	69.40	15.54
Alaska	82	15.55	40.04	15.31	2.41	57.76	3.84
Colorado	167	13.54	32.23	11.87	2.32	46.42	22.72
Indiana	115	10.77	36.50	11.49	2.41	50.40	16.10
Kentucky	954	10.64	42.94	11.93	2.65	57.51	4.25
Maryland	55	16.01	48.80	12.45	2.86	64.11	6.73
Montana	94	12.04	27.45	11.16	2.46	41.07	31.63
Ohio	679	12.77	36.89	13.03	2.57	52.49	12.61
Pennsylvania	825	14.29	53.82	14.92	2.92	71.66	19.30
Virginia	446	9.86	40.80	10.39	2.15	53.34	11.20
West Virginia	577	9.60	39.88	11.80	2.48	54.16	9.83
Wyoming	198	11.36	31.91	11.90	2.15	45.97	23.47
Others	340	16.44	48.38	18.35	3.54	70.26	16.97
Mean	-	-	12.83	40.94	12.97	2.60	56.50	-
RSD (%)	-	-	18.19	19.94	16.71	14.55	17.44	-

Table 14. Coal ranks with original samples

Coal Rank	No. of samples	Mean ash yield, dry (wt%)	LREY original mean	MREY original mean	HREY original mean	TREY original mean	%Dev original
Bituminous	4846	12.11	53.65	15.40	4.07	73.11	21.72
Lignite	155	13.11	52.64	19.31	5.24	77.19	28.51
Subbituminous	398	18.45	86.17	34.96	9.11	130.24	116.82
Others	84	18.20	72.55	18.71	5.62	96.88	61.29
Mean	-	15.47	66.25	22.09	6.01	94.36	-
RSD (%)	-	21.51	24.34	39.59	36.10	27.63	-

Table 15. Coal ranks with adjusted samples

Coal Rank	No. of samples	Mean ash yield, dry (wt%)	LREY adjusted mean	MREY adjusted mean	HREY adjusted mean	TREY adjusted mean	%Dev adjusted
Bituminous	4846	12.11	44.48	12.84	2.65	59.96	0.18
Lignite	155	13.11	33.90	12.26	2.27	48.43	19.38
Subbituminous	398	18.45	52.90	23.62	4.19	80.71	34.36
Others	84	18.20	55.90	14.95	3.46	74.32	23.72
Mean	-	15.47	46.79	15.92	3.14	65.85	-
RSD (%)	-	21.51	21.08	33.08	27.32	22.02	-

Table 16. Geological Age with original samples

Geological Age	No. of samples	Mean ash yield, dry (wt%)	LREY original mean	MREY original mean	HREY original mean	TREY original mean	%Dev original
Cretaceous	351	13.09	53.90	19.89	5.36	79.15	2.92
Pennsylvanian	4631	11.96	53.26	14.98	3.94	72.18	11.47
Tertiary	406	15.36	67.03	25.99	7.00	100.02	22.67
Others	33	22.48	84.48	26.42	7.48	118.38	45.19
Mean	-	15.72	64.67	21.82	5.95	92.43	-
RSD (%)	-	30.02	22.66	24.98	27.15	22.67	-

Table 17. Geological Age with adjusted samples

Geological Age	No. of samples	Mean ash yield, dry (wt%)	LREY adjusted mean	MREY adjusted mean	HREY adjusted mean	TREY adjusted mean	%Dev adjusted
Cretaceous	351	13.09	33.43	12.59	2.38	48.41	40.63
Pennsylvanian	4631	11.96	44.88	12.74	2.63	60.25	26.11
Tertiary	406	15.36	41.91	16.77	3.03	61.71	24.32
Others	33	22.48	60.54	19.86	4.39	84.80	4.00
Mean	-	15.72	45.19	15.49	3.11	63.79	-
RSD (%)	-	30.02	25.06	22.59	28.87	23.86	-

Based on research from Seredin (Seredin and Dai, 2012), there are two main criteria that are used to classify coal samples in terms of their REY promise. The first is total rare earth oxide (TREO) concentration on an ash basis. Coal samples in the COALQUAL database have trace element concentrations on a whole coal basis. This must be converted to an ash basis by Eq. 16.

$$REY_{ASH} = \frac{REY_{Coal}}{\left(\frac{Dry\ Ash\ \%}{100}\right)}$$ *Eq.16*

For coal sample classification purposes, the REY concentration for each REY of each sample must be converted to an ash basis, and then converted to rare earth oxide (REO) concentration for each sample. These REY concentrations on an ash basis must be converted for each REY and then summed to calculate the TREO concentration.

$$REO\ ppm, ash\ basis = REY\ ppm, ash\ basis * \left(\frac{2 * REY\ MW + 3 * O\ MW}{2 * REY\ MW} \right) \qquad Eq.17$$

The criterion for a promising sample is a TREO ash basis concentration of 1000 ppm or greater. The next criterion is the outlook coefficient (C_{outl}). C_{outl} is the ratio of critical REY concentration to excess REY concentration for each sample. A C_{outl} < 0.7 is considered unpromising and a $C_{outl} \geq 0.7$ is promising. Both criteria must be met for a sample to be considered promising; otherwise the sample is considered to be unpromising [65]. Original and adjusted samples are plotted with the result shown in Fig. 20. Table 18 shows percentages of original and adjusted samples that are promising for ranks, coal provinces, coal regions, and states. Figs. 21, 22, and 23 show the probabilities of uncovering promising samples amongst all coal samples collected in the COALQUAL database for ranks, coal regions, and states respectively. To show how these Fig. 20 translates geographically, in Fig. 24, promising and unpromising coal samples are plotted on a map of the U.S. and the intensity of the color of each respective class is based on the magnitude of the TREO concentration of each sample.
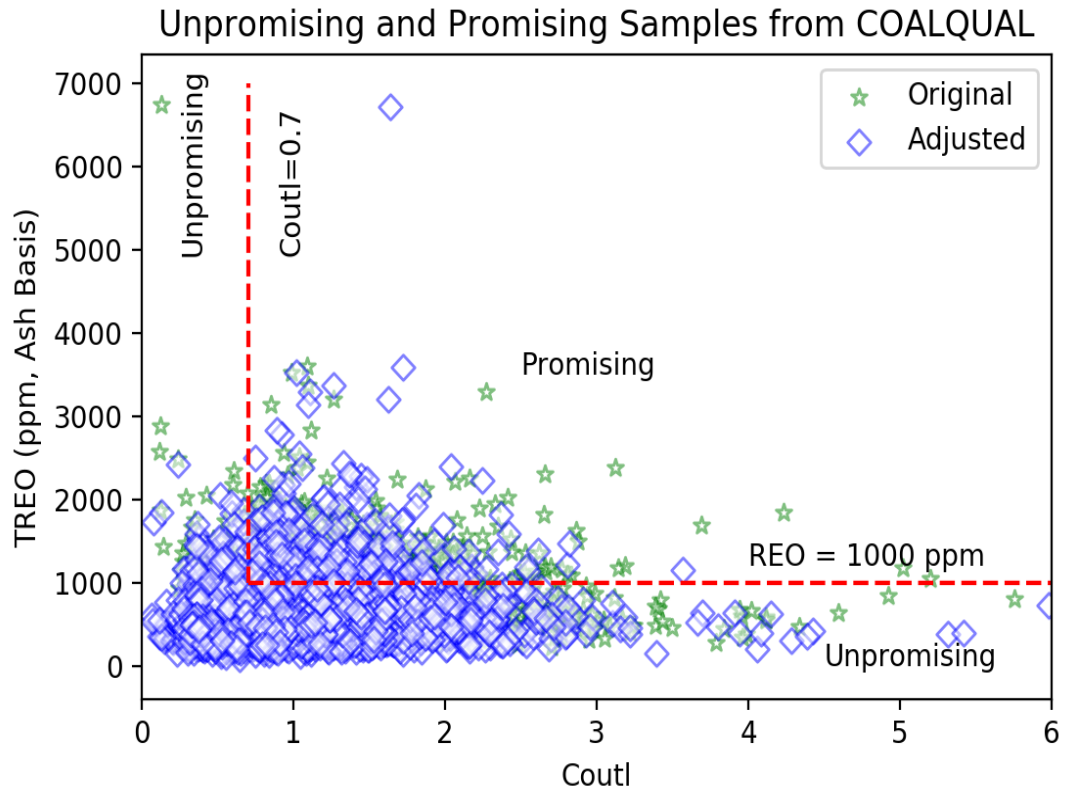
Figure 20. Unpromising and promising samples from COALQUAL based on TREO and C_{outl}.

Table 18. Percentages of original and adjusted samples that are promising

Categories	No. of original samples	Original percent	No. of adjusted samples	Adjusted percent
Rank	-	-	-	-
Bituminous	691	91.64	454	97.01
Subbituminous	37	4.91	6	1.28
Lignite	21	2.79	6	1.28
Others	4	0.53	2	0.43
Province	-	-	-	-
Eastern	626	82.91	431	91.90
Interior	40	5.30	21	4.48
Rocky Mountain	38	5.03	4	0.85
Northern Great Plains	33	4.37	6	1.28
Gulf	12	1.59	6	1.28
Region	-	-	-	-
Central Appalachian	350	46.60	239	51.07
Southern Appalachian	150	19.97	105	22.44
Northern Appalachian	123	16.38	85	18.16
Eastern	35	4.66	21	4.49
Powder River	27	3.60	6	1.28
State	-	-	-	-
Kentucky	186	24.64	136	29.00
Alabama	139	18.41	97	20.68
West Virginia	111	14.70	59	12.58
Pennsylvania	82	10.86	57	12.15
Virginia	77	10.20	58	12.37
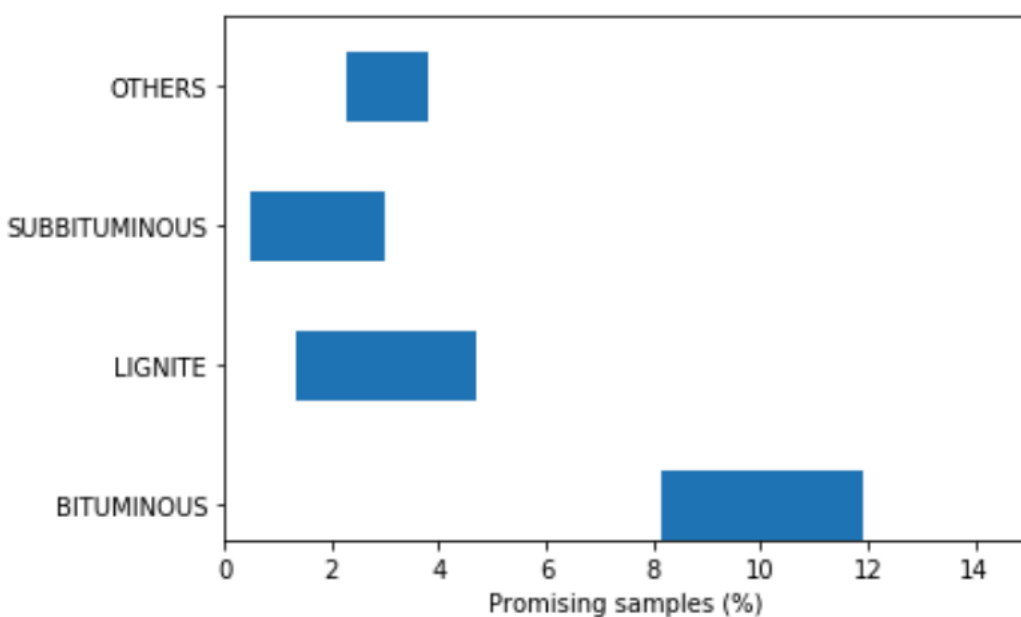
Figure 21. Probabilities of finding promising samples amongst coal ranks in

COALQUAL
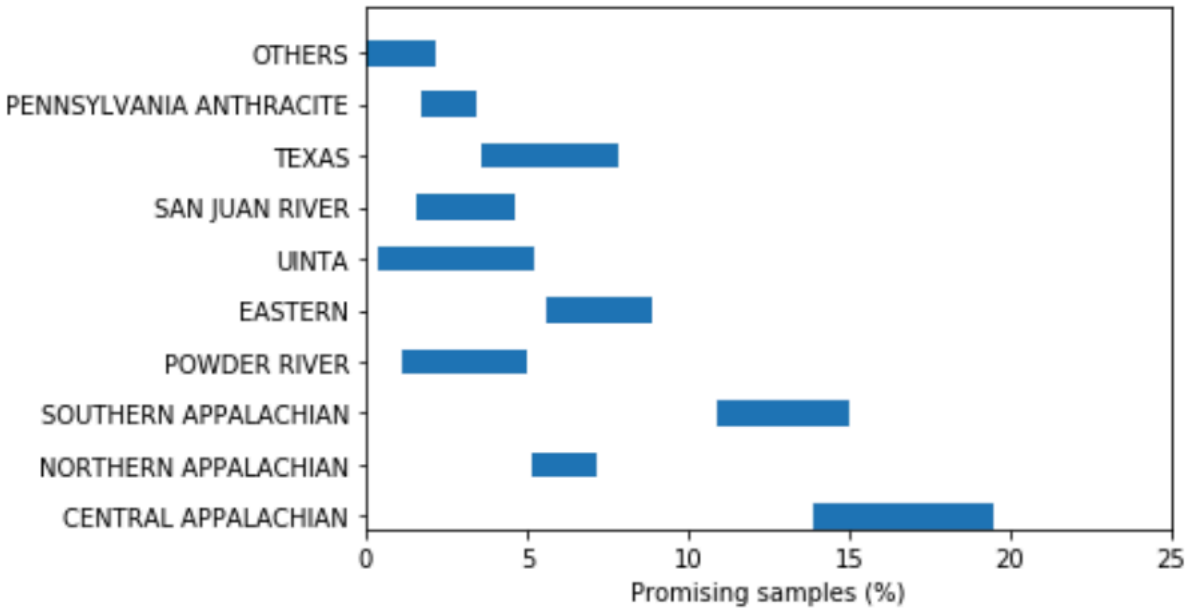
Figure 22. Probabilities of finding promising samples amongst coal regions in

COALQUAL
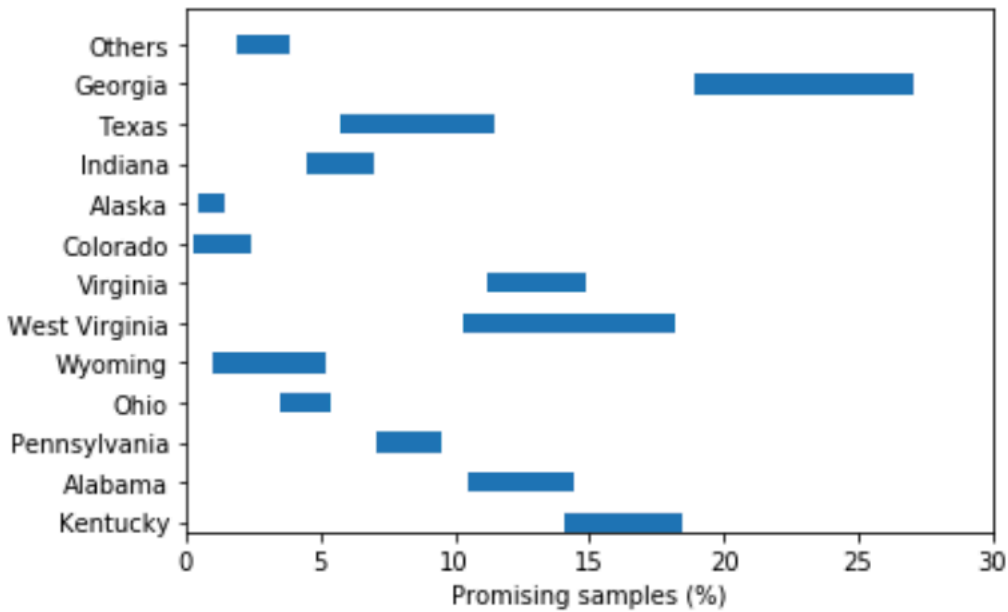
Figure 23. Probabilities of finding promising samples amongst states in COALQUAL
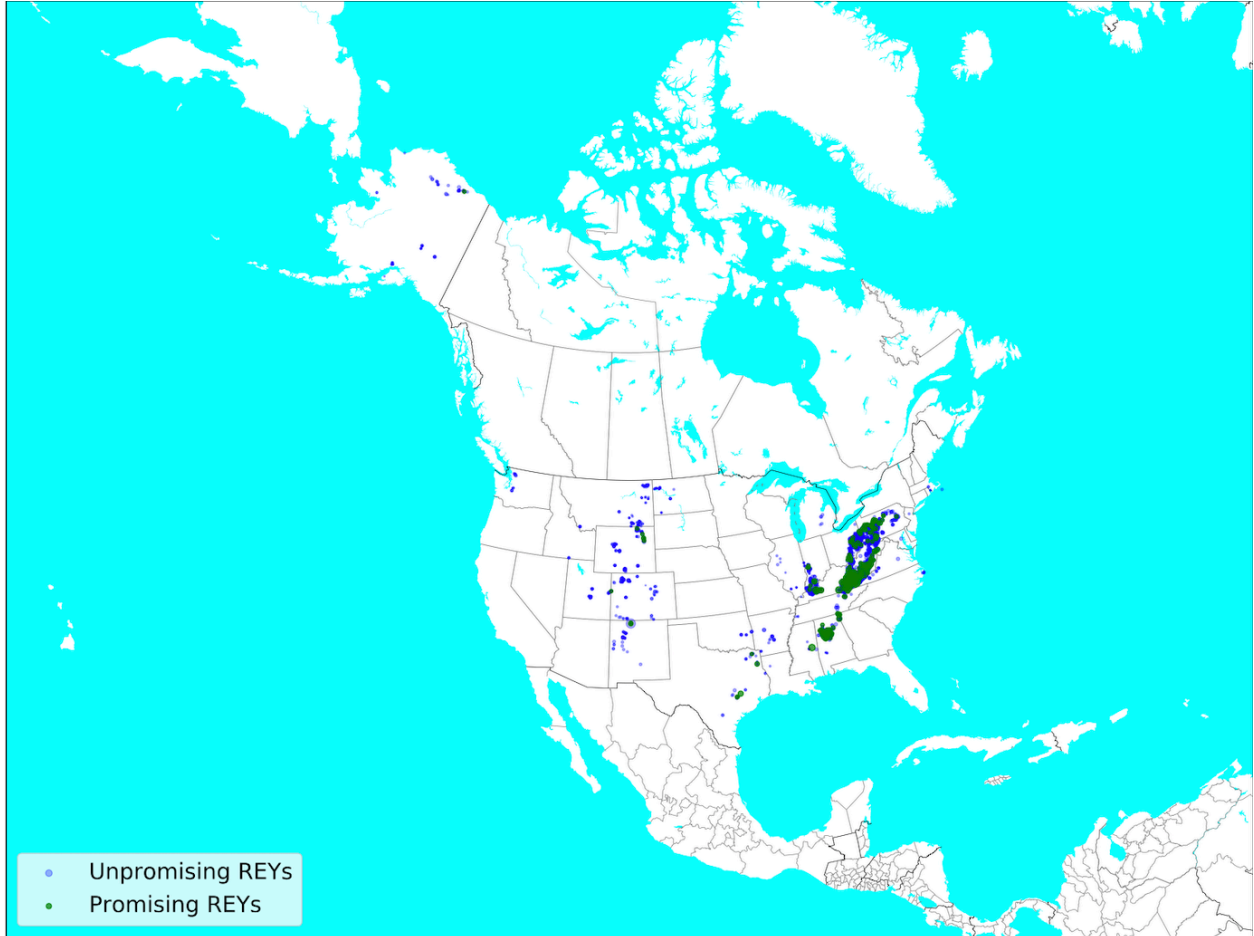
Figure 24. Map of the United States showing locations and concentration density of unpromising and promising coal samples

3.5 Feature selection for machine learning purposes

For the application purposes of this work, very few features need to be used in the machine learning algorithms. These features need to be easy to measure and obtain either from field collection or in the lab. All features were compared to each other in a correlation matrix to determine their relationships to each other. The features that are the most correlated to REY concentrations will be most useful in terms of classification because the classes are based on REY concentrations. The Pearson correlation coefficient (R) is defined in Eq. 18 where x is the value of one variable and y is the value of the other variable [42]. R needs to be calculated for

each attribute with respect to REYs. A R ≥ 0.5 shows that two attributes are significantly positively correlated.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{Covariance(X,Y)}{\sqrt{Variance(X)Variance(Y)}} \qquad Eq.18$$

Table 19. Attributes from COALQUAL that correlate most to REY concentrations

Feature	Correlation															
	Ce	Dy	Er	Eu	Gd	Ho	La	Lu	Nd	Pr	Sm	Tb	Tm	Y	Yb	Sum
Re Concentration (ppm)	0.582	0.684	0.578	0.478	0.644	0.751	0.571	0.621	0.564	0.891	0.713	0.635	0.931	0.508	0.548	9.701
Al Concentration (ppm)	0.786	0.783	0.677	0.657	0.637	0.572	0.795	0.267	0.64	0.738	0.408	0.212	0.836	0.52	0.661	9.189
Si Concentration (ppm)	0.693	0.784	0.684	0.566	0.655	0.62	0.695	0.36	0.586	0.77	0.454	0.308	0.89	0.51	0.6	9.174
United States Geological Survey (GS) Ash Percent	0.684	0.789	0.647	0.596	0.705	0.596	0.688	0.304	0.621	0.761	0.404	0.266	0.842	0.5	0.613	9.015
GS Ash Concentration (ppm)	0.681	0.781	0.644	0.591	0.692	0.603	0.685	0.321	0.596	0.753	0.416	0.282	0.84	0.509	0.613	9.009
Ni Concentration (ppm)	0.515	0.613	0.606	0.398	0.462	0.729	0.5	0.61	0.465	0.788	0.728	0.653	0.883	0.439	0.465	8.854
Au Concentration (ppm)	0.632	0.848	0.787	0.516	0.523	0.624	0.632	0.261	0.55	0.715	0.387	0.248	0.871	0.456	0.544	8.595
Standard Ash Percent	0.678	0.728	0.59	0.592	0.656	0.54	0.678	0.249	0.579	0.702	0.377	0.214	0.787	0.48	0.607	8.458
U Concentration (ppm)	0.694	0.628	0.561	0.643	0.575	0.479	0.687	0.238	0.705	0.566	0.406	0.229	0.642	0.618	0.65	8.323
Mb Concentration (ppm)	0.546	0.509	0.464	0.386	0.498	0.589	0.55	0.472	0.579	0.623	0.6	0.504	0.692	0.637	0.528	8.187

From the Pearson correlation results in Table 19, ash percentage stands out as good option for a feature that correlates to REY concentrations. Trace element concentrations such as for Re, Al, Si, Ni, Au, U, and Mb have good correlations as well, but are not good choices for the machine learning model because measuring trace element concentrations is a more extensive laboratory analysis than measuring ash percentages requiring more time and financial resources.
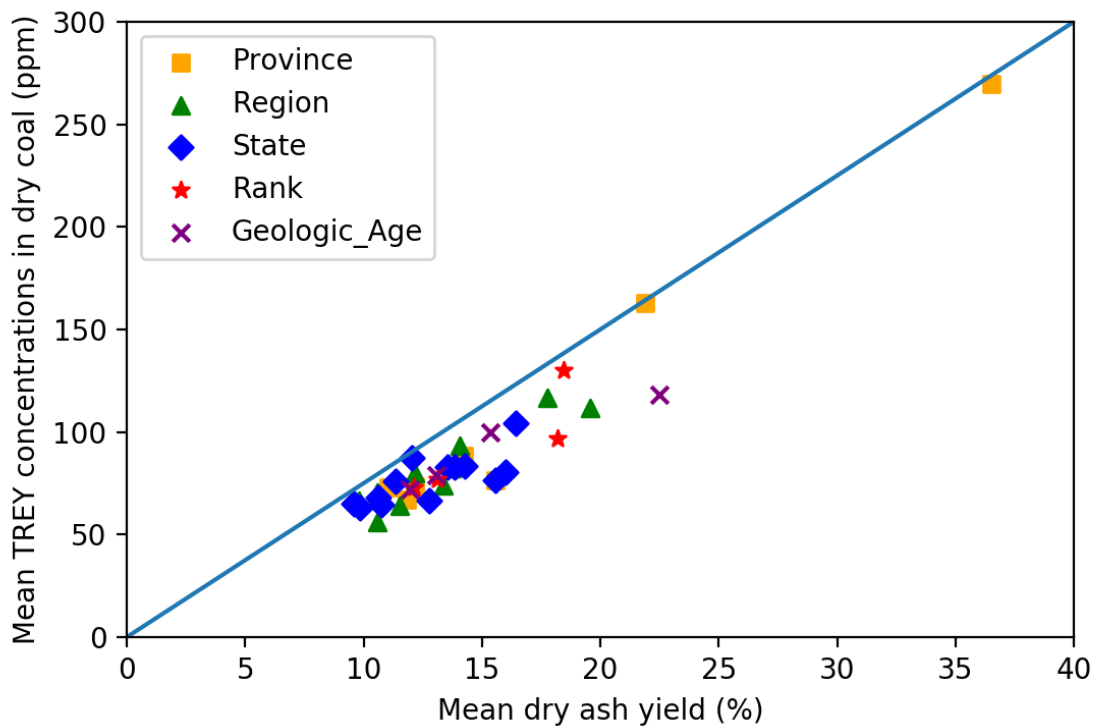
Figure 25. Mean TREY concentration with respect to mean dry ash percentage

Fig. 25 shows that as TREY concentration increases, ash percentage increases. Because of this relationship, it follows that in the COALQUAL database, REYs are associated primarily with the inorganic matrix of coal. This probably represented more in this dataset because it is skewed toward bituminous coals characterized larger inorganic matrices.

Pearson correlation is the preferred way to determine the best attributes for machine learning; however, there are many ways to do this. Another way of doing this is by using the feature importance application that is a by-product of the random forest algorithm in python's scikit-learn library. After a training set is fit to a random forest model, it can determine the most important features utilized in fitting the model. This is a fairly reliable way of determining which variables are the most important for machine learning purposes. Based on a training set

consisting of 80% of the samples from COALQUAL that have complete REY profiles, the

following features are the most important.
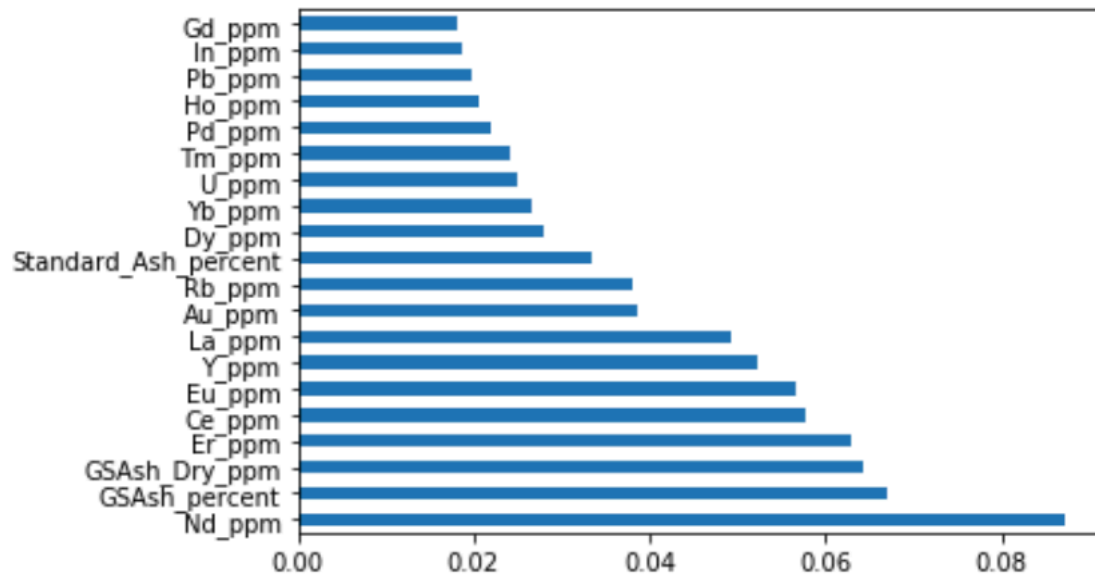
Figure 25. Random forest feature importance for COAQUAL database samples with

complete REY profiles

Random forest feature importance, shown in Fig. 25, supports the results shown from taking

Pearson correlation coefficient calculations. All ash related metrics are near the top in terms of

feature importance.

There are no other features in the COALQUAL database that meet the profile of being

useful in terms of the goals of the machine learning model besides ash-based attributes. Trace

element concentrations require analysis techniques that require time and financial resources.

Also, if trace element concentrations were used then it renders this machine learning model

irrelevant because coal samples could directly have their REY concentrations analyzed and

traditional analysis could be conducted. Field-based metrics would be great to have for the

machine learning model, but there were too many samples with missing values in terms of

various locational information (i.e. coal region, province, field, mine, etc.). Geological

information would have been wonderful to have as well, but the same problem missing values problem is present. Proximate analysis measurements are good to use for this model because they are numerical data types that more easily and more commonly measured in the lab than trace element concentrations. However, with the exception of ash, no other proximate showed promising correlation to REY concentrations. Ash metrics are easily fit to machine learning algorithms and will provide a readily available low-cost feature to add to the dataset if more samples are provided. Standard ash percent will be used as the only feature in the machine learning model because that ash metric is the measured using the American Society for Testing and Materials International (ASTM) standard for measuring ash content in coal samples. This is the international standard for ash measurement and is therefore more commonly used. GS Ash percent and concentration are measured using the USGS standard which is used less in practice. Having only one feature may seem too simplistic, but as discussed earlier some algorithms perform better given fewer features to classify on. This model is designed to be a tool to save time and money in regard to testing and analyzing coal samples for their REY concentration. As long as the model performs fairly reliably and minimizes false positives, the model is useful. Given more feature collection from coal samples, more features can be added to the model. More comprehensive analysis needs to be conducted on coal samples. This will yield a larger and more robust dataset that can be feed into machine learning algorithms. This methodology will be discussed later.

3.6 Machine Learning and Feature Preprocessing

The first step in preparing the data for machine learning purposes is to encode the samples determined to be promising or unpromising as a binomial. Adjusted unpromising and promising samples from the COALQUAL database with complete REY profiles are assigned a

class of 0 and 1 respectively. Then this dataset containing 5485 samples is randomly split into a training set and a test set. 80% of samples are assigned to the training set and 20% are assigned to the test using a stratified method. This way the test set has the same proportion of class values as the training set. Upon doing so, it is obvious that there are a lot of more unpromising samples and promising samples in both sets. This bias must be accounted for in order for any given machine learning algorithm to classify the samples correctly. Synthetic Minority Over-Sampling Technique (SMOTE) is an algorithm that over-samples the minority class by implanting synthetic examples to each sample with the minority class along the line segments joining any/all of the minority class nearest neighbors, k. Based on how much over-sampling is conducted, neighbors from the k nearest neighbors are randomly selected. The synthetic samples are created by calculating the difference between the sample feature under consideration and its nearest neighbor in data space. The difference is multiplied by a random number between 0 and 1 and it is added to the considered sample feature [66]. The full algorithm is detailed in Fig. 26. SMOTE will create enough synthetic samples to even up the number of class values in a dataset. When the original distribution of data is compared to the new distribution with synthetic samples, not much difference is observed. The next step in the preprocessing phase is to impute the data so there are no missing values in the dataset. Imputing replaces missing values with the median for each respective feature. Since standard ash percent is the only feature used, the median for standard ash percent in all samples in the test set is used to fill in any missing values for standard ash percent. With the training and test sets having no missing values and having added synthetic samples to account for class bias, and both datasets containing standard ash percent as the only feature and unpromising and promising classifications for each sample, the datasets are ready for machine learning implementation.

Algorithm $SMOTE$(T, N, k)

Input: Number of minority class samples T; Amount of SMOTE $N\%$; Number of nearest neighbors k

Output: $(N/100) * T$ synthetic minority class samples

1. (∗ *If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd.* ∗)
2. **if** $N < 100$
3.   **then** Randomize the T minority class samples
4. $T = (N/100) * T$
5. $N = 100$
6. **endif**
7. $N = (int)(N/100)$ (∗ *The amount of SMOTE is assumed to be in integral multiples of 100.* ∗)
8. $k =$ Number of nearest neighbors
9. $numattrs =$ Number of attributes
10. $Sample[\,][\,]$: array for original minority class samples
11. $newindex$: keeps a count of number of synthetic samples generated, initialized to 0
12. $Synthetic[\,][\,]$: array for synthetic samples
   (∗ *Compute k nearest neighbors for each minority class sample only.* ∗)
13. **for** $i \leftarrow 1$ **to** T
14.   Compute k nearest neighbors for i, and save the indices in the $nnarray$
15.   Populate(N, i, $nnarray$)
16. **endfor**

   $Populate(N,\ i,\ nnarray)$ (∗ *Function to generate the synthetic samples.* ∗)
17. **while** $N \neq 0$
18.   Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
19.   **for** $attr \leftarrow 1$ **to** $numattrs$
20.   Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
21.   Compute: $gap =$ random number between 0 and 1
22. $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
23.   **endfor**
24. $newindex++$
25. $N = N - 1$
26. **endwhile**
27. **return** (∗ *End of Populate.* ∗)
   End of Pseudo-Code.

Figure 26. The SMOTE pseudo-algorithm [67]

3.7 Machine Learning Algorithm Implementation

The five machine learning algorithms trained and tested in this work are k-nearest neighbors, logistic regression, CART decision tree classifier, random forest, and adaboost classifier. The two major performance measures that each algorithm is tested on is classification accuracy and percentage of false positives.

3.7.1 k-Nearest Neighbors

The only hyperparameter that needed to be tuned for k-nearest neighbors is k, number of nearest neighbors in the majority of the voting process. In order to tune this parameter, possible values of k from 1 to 100 were iteratively tested to the fitted algorithm. As mentioned previously, there is no need to "train" the algorithm. The entire dataset is used for classification purposes.

The main performance indicator used to evaluate the algorithm was accuracy, which is the ratio of correct classifications to total classifications. Accuracy was measured for all values of k from 1 to 100, and a k = 40 showed a near peak in accuracy with a zig zag relationship between accuracy and k value throughout all possible values of k. A visualization of the results are shown in Fig. 27. Mean squared error (MSE) was chosen as the parameter to eliminate misclassification error related to k. The optimal k value would minimize MSE. A display of this is shown in Fig. 28. A k ≥ 40 did not show significant decrease in MSE and therefore k = 40 was

chosen as the optimal k value.
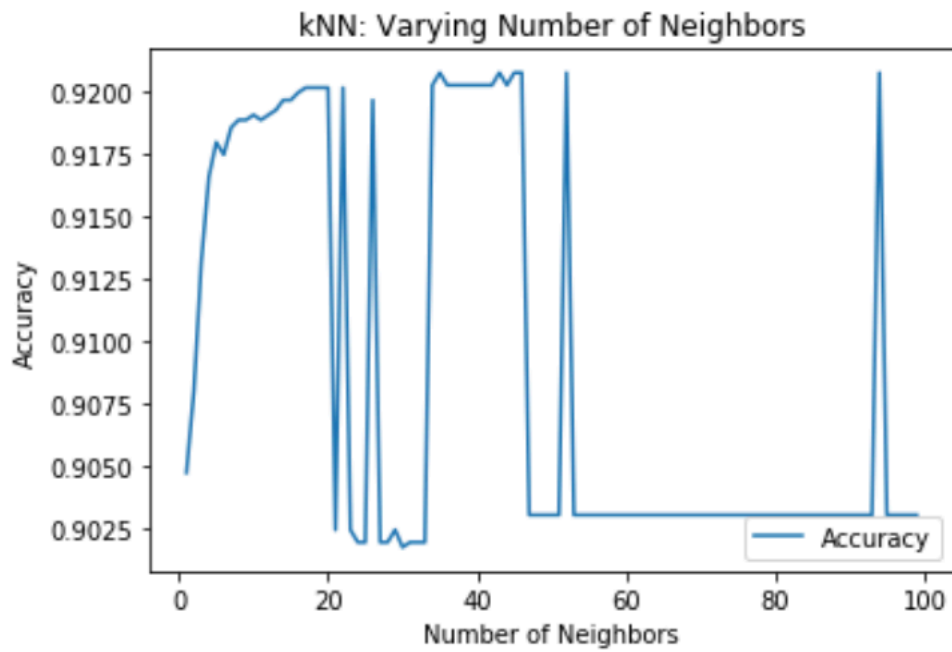
Figure 27. Accuracy versus possible k values
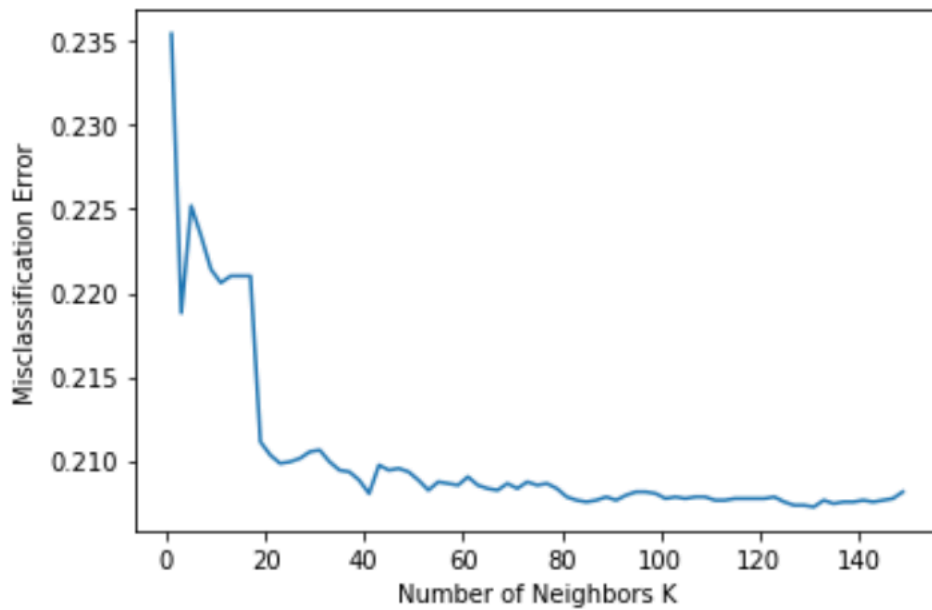
Figure 28. MSE vs possible k values

The algorithm performed very well. K-nearest neighbors performed with an accuracy of about 92% and only about 2% false positives. See Fig. 30 for all performance metrics. The

confusion matrix is a very useful tool to define classification error. A diagram explaining the

confusion matrix is shown in Fig. 29.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 29. Diagram of a confusion matrix [68]

```
Accuracy: 92.03 %
Confusion Matrix:
 [[4864  152]
 [ 648 4368]]
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      5016
           1       0.97      0.87      0.92      5016

    accuracy                           0.92     10032
   macro avg       0.92      0.92      0.92     10032
weighted avg       0.92      0.92      0.92     10032
```

Figure 30. Performance results of k-nearest neighbors' algorithm

There were only 152 false positives out of 10032 samples which equates to a percentage of about

2%.

3.7.2 Logistic Regression

For logistic regression, the main parameter that must be tuned is C, the inverse regularization parameter. The regularization parameter, λ, is added to the cost function in order to account for weighting in the algorithm. λ is proportional to the magnitude of the sum of the weights in the cost function. C is a penalty term added to the cost function to discourage overfitting [69].

In order to determine the proper value of C, the gridsearchcv function is utilized in sklearn. This function exhaustively searches for parameters that will maximize algorithm performance after the algorithm is fitted to a dataset [70]. Gridsearchcv runs five cross validations of the desired algorithm and uses a specified range for any given parameters that need to be tested. The function aims to maximize accuracy. Given that C is the inverse of a term in the cost function, the optimal C will minimize the cost function. Fig. 31 shows the optimal C value after tuning. With a minimized cost function, and a maximized accuracy, the best result is chosen.

```
#Tuning hyperparameters using GridSearchCV

c_space = np.logspace(-5, 8, 15)
param_grid = {'C': c_space}

logreg = LogisticRegression(solver='lbfgs')

logreg_cv = GridSearchCV(logreg, param_grid, cv=5)
logreg_cv.fit(Coalqual_train_StdAsh_per_X_SM, Coalqual_train_y_SM)

print("Tuned Logistic Regression Parameters: {}".format(logreg_cv.best_params_))
print("Best score is {}".format(logreg_cv.best_score_))
Tuned Logistic Regression Parameters: {'C': 0.05179474679231213}
Best score is 0.6725641664590082
```

Figure 31. Optimal C value and best accuracy with the COALQUAL database fitted to the logistic regression algorithm

The optimal C = 0.052 yields an accuracy of about 67% utilizing the training data. The

chosen C parameter must be subjected to the test data after being trained.

```
Confusion Matrix:
 [[640 363]
 [ 27  67]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.64      0.77      1003
           1       0.16      0.71      0.26        94

    accuracy                           0.64      1097
   macro avg       0.56      0.68      0.51      1097
weighted avg       0.89      0.64      0.72      1097
```
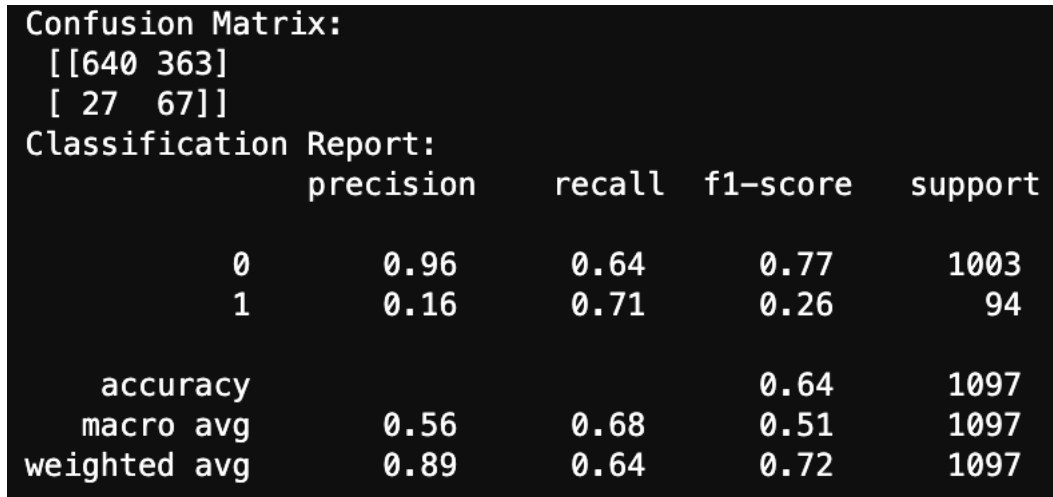
Figure 32. Confusion matrix and classification report for logistic regression

When the test set is processed by the algorithm, the accuracy decreases to 64%. The percent of

false positives is about 33%. All performance results are shown in Fig. 32.

3.7.3 CART Decision Tree Classifier

The CART decision tree classifier has many hyperparameters to tune. The maximum

depth or the maximum levels of the tree, the minimum number of samples required to split a

node, the minimum number of samples required at each leaf node, how the weights are

associated with the classes, the function used to measure split quality, the way the tree splits at

each node, and whether or not the data should be presorted before beginning the algorithm all

need to tuned to the best possible outcome that minimizes cost and increases accuracy.

```
#Tuning hyperparameters for Decision Tree Classifier

dt = DecisionTreeClassifier()

# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
#class_weight
class_weight = ['balanced', None]
#criterion
criterion = ['gini', 'entropy']
#splitter
splitter = ['best', 'random']
#presort
presort = [True, False]
```

Figure 32. Hyperparameters that need to be tuned for the CART algorithm

```
# Create the random grid
random_grid = {'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'class_weight': class_weight,
               'criterion': criterion,
               'splitter': splitter,
               'presort': presort}

dt_random = RandomizedSearchCV(estimator = dt, param_distributions = random_grid, n_iter = 100, cv = 3, verbose=2, random_state=42, n_jobs = -1)
dt_random.fit(Coalqual_train_StdAsh_per_X_SM, Coalqual_train_y_SM)
dt_random.best_params_
```

Figure 33. Randomized search procedure for selecting optimal hyperparameters

The sci-kit learn function, RandomizedSearchCV, is used to tune hyperparameters for this

algorithm as well as other decision tree classifiers. It is faster at processing larger combinations

of hyperparameters than gridsearchcv [71]. This process is shown in Figs 32, 33, and 34.

```
Best parameters are {'splitter': 'random', 'presort': True, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': None, 'criterion': 'entropy',
'class_weight': 'balanced'}
```

Figure 34. Best parameters for the COALQUAL database fitted to the CART algorithm

Using the best parameters, the algorithm is trained and tested. The results from the test are below.

```
Confusion Matrix:
 [[789 214]
 [ 54  40]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.79      0.85      1003
           1       0.16      0.43      0.23        94

    accuracy                           0.76      1097
   macro avg       0.55      0.61      0.54      1097
weighted avg       0.87      0.76      0.80      1097

Stored 'dt' (DecisionTreeClassifier)
```
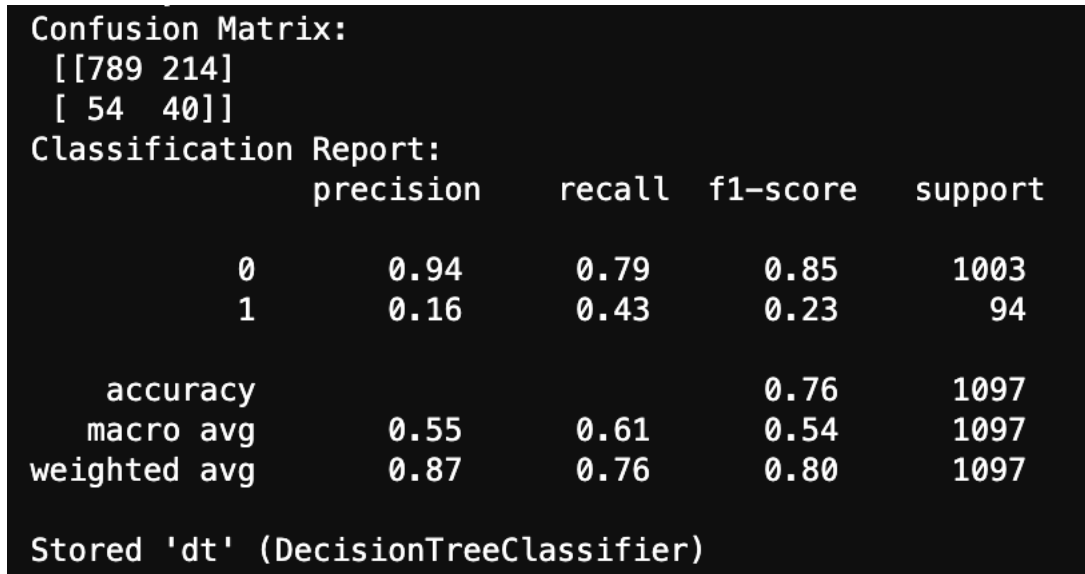
Figure 35. Confusion matrix and classification report for CART algorithm

Fig. 35 shows that accuracy is 76% and about 20% of classifications were false positives.

3.7.4 Random Forest

Since random forest is an ensemble method using decision tree classifiers, its application is conducted very similarly to CART's. There are a couple of different hyperparameters that need to be tuned for ensemble learning purposes. This process is detailed in Figs 36, 37, and 38 below.

```
#Tuning hyperparameters for Random Forest Classifier

rnd_clf = RandomForestClassifier()

# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
#class_weight
class_weight = ['balanced', 'balanced_subsample', None]
#criterion
criterion = ['gini', 'entropy']
#oob_score
oob_score = [True, False]
#warm_start
warm_start = [True, False]
```

Figure 36. Hyperparameters that need to be tuned for the random forest algorithm

The main difference in hyperparameters is the number of trees in the random forest or the

number of estimators used in the ensemble.

```
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'class_weight': class_weight,
               'criterion': criterion,
               'oob_score': oob_score,
               'warm_start': warm_start}

rf_random = RandomizedSearchCV(estimator = rnd_clf, param_distributions = random_grid, n_iter = 100, cv = 3, verbose=2, random_state=42, n_jobs = -1)
rf_random.fit(Coalqual_train_StdAsh_per_X_SM, Coalqual_train_y_SM)
print("Best parameters are {}".format(rf_random.best_params_))
```

Figure 37. Randomized search procedure for selecting optimal hyperparameters for

random forest algorithm

```
Best parameters are {'warm_start': False, 'oob_score': True, 'n_estimators': 1400, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_depth': 60, 'criterion':
'entropy', 'class_weight': 'balanced'}
```

Figure 38. Best parameters for the COALQUAL database fitted to the random forest

algorithm

The optimum number of trees for the COALQUAL database is 1400 in the random forest

algorithm with a depth of 60 nodes.

```
Confusion Matrix:
 [[825 178]
 [ 52  42]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.82      0.88      1003
           1       0.19      0.45      0.27        94

    accuracy                           0.79      1097
   macro avg       0.57      0.63      0.57      1097
weighted avg       0.88      0.79      0.83      1097
```

Figure 39. Confusion matrix and classification report for random forest algorithm

Fig. 39 shows that accuracy is 79% and about 16% of classifications are false positives for the

random forest algorithm.

3.7.5 Adaboost

Like random forest, adaboost is an ensemble algorithm and it follows a process of tuning

that is very similar to random forest. The process is detailed in Figs. 40, 41, and 42.

```python
#Tuning hyperparameters for Adaboost Classifier

AB_clf = AdaBoostClassifier(base_estimator = dt, random_state=42)

n_estimators = [int(x) for x in np.linspace(start = 50, stop = 2000, num = 10)]
learning_rate = [0.001, 0.01, 0.1, 1]
algorithm = ['SAMME', 'SAMME.R']

# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'learning_rate': learning_rate,
               'algorithm': algorithm}
```

Figure 40. Hyperparameters that need to be tuned for the adaboost algorithm

71

```
rf_random = RandomizedSearchCV(estimator = AB_clf, param_distributions = random_grid, n_iter = 100, cv = 3, verbose=2, random_state=42, n_jobs = -1)
rf_random.fit(Coalqual_train_StdAsh_per_X_SM, Coalqual_train_y_SM)
print("Best parameters are {}".format(rf_random.best_params_))
```

Figure 41. Randomized search procedure for selecting optimal hyperparameters for the

adaboost algorithm

```
Best parameters are {'n_estimators': 1566, 'learning_rate': 0.001, 'algorithm': 'SAMME.R'}
```

Figure 42. Best parameters for the COALQUAL database fitted to the adaboost algorithm

```
Confusion Matrix:
 [[793 210]
 [ 54  40]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.79      0.86      1003
           1       0.16      0.43      0.23        94

    accuracy                           0.76      1097
   macro avg       0.55      0.61      0.54      1097
weighted avg       0.87      0.76      0.80      1097
```

Figure 43. Confusion matrix and classification report for the adaboost algorithm

See Fig. 43. The accuracy for adaboost is 76% and about 19% of classifications are false

positives for the adaboost algorithm.

CHAPTER 4: RESULTS

The main objective of this work was to create a machine learning model that would be useful in classifying coal samples as promising or unpromising in terms of their REY concentrations and economic outlook. This model would be useful because it would act as a filter for coal samples and would eliminate samples with low likelihood of being promising from the pool of samples that would be promising. Having to test fewer samples would reduce cost and time associated with analytically measuring trace element concentrations.

The machine learning algorithms tested were k-nearest neighbors, logistic regression, the CART decision tree classifier, random forest, and adaboost. Table 20 below shows the performance results for all machine learning algorithms tested.

Table 20. Performance results for all machine learning algorithms.

Performance Results		
Machine Learning Algorithm	Accuracy (%)	False Positive (%)
k-nearest neighbors	92	2
logistic regression	64	33
CART decision tree	76	20
random forest	79	16
adaboost	76	19

Each model was tested based on the accuracy associated in with classification, and the percentage of false positives that each model classified. According to the results obtain from testing, k-nearest neighbors outperformed all of the other algorithms. With an accuracy score of about 92% and about 2% false positives, this algorithm outperformed all other algorithms significantly.

There are several reasons k-nearest neighbors likely outperformed all other algorithms. One of the primary reasons is that k-nearest neighbors performs very well when very few features are used. As mentioned previously, with increasing dimensionality, data points increase their distance from one another which lowers the performance of the algorithm. This k-nearest neighbors model only has one feature inputted into it and therefore does not have to work as hard to create discrete neighborhoods for accurate classification. For most models, the more data instances inputted into a machine learning algorithm, the more accurate the algorithm will perform in terms of classification. With k-nearest neighbors, the dataset does not need to be split into a training and test set because it is a "lazy learner" (the target function is approximated locally). Lazy learners have the ability to solve a multitude of computations and deal with fluctuations in the domain of the problem, simultaneously. All instances can be entered into the algorithm and the more unbiased instances that are placed into the algorithm, the better it will perform. Since there are a total of 10032 real and synthetic samples from SMOTE utilized in the machine learning algorithms, there are quite a bit of samples that k-nearest neighbors can utilize. The other machine learning algorithms split the dataset into a training and test set, 80% and 20% respectively. This means that when they are tested, they only test 1097 samples which lowers performance.

Logistic regression was the worst performer primarily because logistic regression is very much a parametric statistical learning method. It is dependent on the inputted data being linear, otherwise the algorithm does not perform accurately. Both the random forest and adaboost algorithms performed decent. If more features would have been added to those algorithms they would have performed better. Due to not having many important features without missing data and the need to simplify the model in terms of the number of features, decision tree classifiers

and ensemble learning algorithms will not perform as well as they potentially could. This is will be addressed in the conclusion.

CHAPTER 5: CONCLUSION

The results show that machine learning models can definitely provide benefit in reducing cost and time in testing coal samples for REY concentrations. All machine learning algorithms were able to classify coal samples at an accuracy of over 60 percent and a percentage of false positives below 35 percent. The best of these algorithms performed exceptionally well, 92% accurate classifications with 2% false positives. This is was accomplished using a dataset that was not collected for the purposes of machine learning. This methodology is extremely effective if reliable data is used. In practice, the data quality utilized in this work could be improved. First of all, assumptions about coal samples have to be made in order for this particular model and methodology to be useful in terms of reliable classification. Soil from anywhere on the surface of the Earth could be combusted and an ash content could be used in this machine learning model that could yield a promising result. Obviously, this is not representative of reality. In order for samples to be used in this particular model, they must come from a location where, historically, promising coal samples have been discovered. The model can still be useful in terms of further researching coal samples from areas of known promise, but in terms of exploration, this limits the usefulness of the model. The methodology should be changed in accordance to further usage if exploration is part of the project scope.

5.1 Recommendations

Research groups across the country have been researching coal samples for the purposes of REY exploration and extraction. The methodology presented in this thesis could be extremely useful if data collection is structured correctly. When collecting samples from the field for

classification purposes, ample field data must also be collected to give the most powerful results in terms of classification. It is recommended that geological information such as stratigraphic formation, coal group, coal bed, the relationship between non-coal material and the coal bed, depth from the surface of the earth to the top of the sample, thickness of the sample, the sample's geological age, seam, rank, and other geological information be recorded for each sample. Completeness in capturing geological information as well as locational information will give the best results in terms of classification. If 5000 coal samples with complete geological, locational, proximate, and REY trace element concentration features are collected, then the five machine learning algorithms can be utilized in the same methodology is as in this work. Categorical data will need to be encoded, but every other step will be the same. The major difference is that the user can include a collection of features in the machine learning algorithms instead of focusing on just ash content. Having a combination of reliable features without missing data will provide better classification power rather than simply using ash content. For example, ensemble methods perform best given a large number of features because they perform better with randomness. Having a large number of samples with a wide array of features without missing data points is the most important factor in machine learning success. After the algorithms are trained using the REY concentrations measured for each sample, then new samples with complete geological, locational, and proximate data can be utilized without measuring REY concentrations to accurately classify coal samples. Promising samples can then be analyzed against all the features collected to discover exploration insights.

From a cost and time standpoint, the model has the potential to be very useful. For INAA analysis, each sample costs $340 [72]. The majority of samples tested will be unpromising and therefore will not need to be tested, but most samples collected from the field will likely be

unpromising. If arbitrarily, 1000 samples are collected, and 100 are promising and 900 are

unpromising; the model could potentially save upwards of $300,000 from testing fewer samples

for trace element concentrations.

CHAPTER 8: REFERENCES

[1]   B. Mason and C. B. Moore, Principles of geochemistry, 4th ed., New York: Wiley, 1982.

[2]   W. D. Jackson and G. Christiansen, "International Strategic Minerals Inventory summary report - Rare earth oxides," U.S. Geological Survey Circular 930-N, p. 68, 1993.

[3]   J. L. Sabot and P. Maestro, Lanthanides, in Kroschwitz, J.I., and Grant, M.H., eds., Kirk-Othmer encyclopedia of chemical technology, 4th ed., New York: John Wiley, 1995.

[4]   K. H. Wedepohl, "The composition of the continental crust," Geochimica et Cosmochimica Acta, vol. 59, no. 7, pp. 1217-1232, 1995.

[5]   D. R. Lide, "Abundance of elements in the earth's crust and sea," in CRC handbook of physics and chemistry, 78th edition, Boca Raton, Florida, CRC Press, 1997, p. 14.

[6]   I. McGill, "Rare earth metals," in Habashi, F., ed., Handbook of extractive metallurgy, Weinheim, New York, Wiley-VCH, 1997, pp. 1695-1741.

[7]   Tutor Circle, "Lanthanide Contraction," 2017. [Online]. Available: http://chemistry.tutorcircle.com/inorganic-chemistry/lanthanide-contraction.html. [Accessed 13 October 2019].

[8]   V. K. Pecharsky and K. A. Gschneidner, "Rare-earth element," Encyclopedia Britannica, 14 March 2014. [Online]. Available: https://www.britannica.com/science/rare-earth-element. [Accessed 13 October 2019].

[9]   Economics & Statistics Department - American Chemistry Council, "The Economic Benefits of the North American Rare Earths Industry," Rare Earth Technology Alliance, 2014.

[10] "U.S. Department of Energy National Energy Technology Laboratory," 2019. [Online]. Available: https://www.netl.doe.gov/research/coal/rare-earth-elements/overview. [Accessed 13 October 2019].

[11] D. Wise, "Security and supply chains and the role of rare earth elements," The Cipher, 2019. [Online]. Available: https://www.thecipherbrief.com/column/expert-view/security-and-supply-chains-role-rare-earth-elements-1091?utm_source=Aggregators&utm_campaign=facc7d4a3a-217 EMAIL_CAMPAIGN_2017_04_09&utm_medium=email&utm_term=0_b02a5f1344-facc7d4a3a-122460921. [Accessed 10 November 2019].

[12] Y. Kanazawa and M. Kamitani, "Rare Earth Minerals and Resources in the World," Journal of Alloys and Compounds, Vols. 408-412, pp. 1339-1343, 2005.

[13] A. Golev, M. Scott, P. D. Erskine, S. H. Ali and G. R. Ballantyne, "Rare earths supply chains: Current status, constraints and opportunities," Resources Policy, vol. 41, pp. 52-59, 2014.

[14] C. K. Gupta and N. Krishnamurthy, Extractive Metallurgy of Rare Earths, CRC Press, 2005.

[15] T. Bank, E. Roth, B. Howard and E. Granite, "U.S. Department of Energy National Energy Technology Laboratory: Geology of Rare Earth Deposits," 2016. [Online]. Available: https://www.netl.doe.gov/research/coal/rare-earth-elements/publications. [Accessed 13 October 2019].

[16] B. Seal, P. Verplanck, B. Van Gosen and A. Grosz, Geologic and Environmental Characteristics of Rare Earth Element Deposit Types Found in the United States, United States Geological Survey, 2012.

[17]   U.S. Department of Energy, "Critical Materials Strategy," U.S. Department of Energy, 2011.

[18]   V. Seredin, "A new method for primary evaluation of the outlook for rare earth element ores," Geology of Ore Deposits, vol. 52, pp. 428-433, 2010.

[19]   U.S. Department of the Interior; U.S. Geological Survey, "Mineral Commodity Summaries 2017," USGS, 2017.

[20]   G. Barakos, H. Mischo and J. Gutzmer, "The AusIMM Bulletin - An outlook on the rare earth elements mining industry," April 2016. [Online]. Available: https://www.ausimmbulletin.com/feature/an-outlook-on-the-rare-earth-elements-mining-industry/. [Accessed 14 October 2019].

[21]   "REE - Rare earth elements and their uses," Geoscience News and Information - Geology.com, 2017. [Online]. Available: http://geology.com/articles/rare-earth-elements/. [Accessed 17 October 2019].

[22]   X. J. Yang, A. Lin, L. Xiao-Liang, Y. Wu, W. Zhou and Z. Chen, "China's ion-adsorption rare earth resources, mining consequences and preservation," Environmental Development, vol. 8, pp. 131-136, 2013.

[23]   N. Foley, "Rare earth element accumulation processes resulting in high-value metal enrichments in regolith," United States Geological Survey, December 2016. [Online]. Available: https://minerals.usgs.gov/science/residual/index.html. [Accessed 10 November 2019].

[24]   V. Zepf, A new approach to the nexus of supply, demand and use: exemplified along the use of neodymium in permanent magnets, Berlin: Springer-Verlag, 2013.

[25] V. V. Seredin and R. B. Finkelman, "Metalliferous coals: A review of the main genetic and geochemical types," International Journal of Coal Geology, vol. 76, pp. 253-289, 2008.

[26] V. V. Seredin and S. Dai, "Coal deposits as potential alternative sources for lanthanides and yttrium," International Journal of Coal Geology, vol. 94, pp. 67-93, 2012.

[27] K. H. Johanneson and X. Zhou, "Geochemistry of the rare earth element in natural terrestrial waters: a review of what is currently known," Chinese Journal of Geochemistry, vol. 16, pp. 20-42, 1997.

[28] V. V. Seredin and M. Y. Shpirt, "Rare earth elements in the humic substance of metalliferous coals," Lithology and Mineral Resources, vol. 52, pp. 428-433, 1999.

[29] V. Seredin, "Major regularities of the REE distribution in coal," Doklady Earth Sciences, vol. 377, pp. 250-253, 2001.

[30] S. I. Arbuzov and V. V. Ershov, Geochemistry of rare elements in coals of Siberia, Tomsk: D-Print, 2007.

[31] S. I. Arbruzov, V. V. Potseluev and L. P. Rikhvanov, Rare elements in coals of the Kuznetsk Basin, Kemerovo, 2000.

[32] S. Dai, D. Li, C. -L. Chou, L. Zhao, Y. Zhang, D. Ren, Y. Ma and Y. Sun, "Mineralogy and geochemistry of boehmite-rich coals: new insights from the Haerwusu Surface Mine, Jungar Coalfield, Inner Mongolia, China," International Journal of Coal Geology, vol. 74, pp. 185-202, 2008.

[33] V. Ershov, "Rare earth elements in the coals of Kizelovsk basin," Geochimia, vol. 3, pp. 274-276, 1961.

[34]   G. M. Eskenazy, "Rare earth elements and yttrium in lithotypes of Bulgarian coals," Organic Geochemistry, vol. 11, pp. 83-89, 1987.

[35]   V. Seredin, "Rare earth element-bearing coals from the Russian Far East deposits," International Journal of Coal Geology, vol. 30, pp. 101-129, 1996.

[36]   V. Seredin, "Metalliferous coals: formation conditions and outlooks for development," in Resources of Russia, vol. VI, Moscow, Geoinformmark, 2004, pp. 452-519.

[37]   P. Zubovic, T. Stadnichenko and N. Sheffey, "Geochemistry of minor elements in coals of the Northern Great Plains Coal Province.," U.S. Geol. Surv. Bull., 1117-A, p. 58, 1961.

[38]   G. M. Eskenazy, "Aspects of the geochemistry of rare earth elements in coal: an experimental approach," International Journal of Coal Geology, vol. 38, pp. 285-295, 1999.

[39]   A. Szalay, "Cation exchange properties of humic acids and their importance in the geochemical enrichment of UO2 and other cations," Geochimica et Cosmochimica Acta, vol. 28, pp. 1605-1614, 1964.

[40]   R. Finkelman, "The origin, occurrence, and distribution of the inorganic constituents in low-rank coals.," in Basic Coal Science Workshop, Houston, TX, USA, 1981.

[41]   Asiri, Sidath. "Machine Learning Classifiers." Medium, Towards Data Science, 11 June 2018, towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623.

[42]   Kvam, Paul H. Nonparametric Statistics with Applications to Science and Engineering. John Wiley, 2017.

[43]   Grant, Peter. "k-Nearest Neighbors and the Curse of Dimensionality." Medium, Towards

         Data Science, 24 July 2019, towardsdatascience.com/k-nearest-neighbors-and-the-curse-

         of-dimensionality-e39d10a6105d.

[44]   Brownlee, Jason. "Logistic Regression for Machine Learning." Machine Learning

         Mastery, 12 Aug. 2019, machinelearningmastery.com/logistic-regression-for-machine-

         learning/. [Accessed 5 December 2019].

[45]   "What Is Sigmoid Function and Explain in Detail?" i2tutorials, 25 Sept. 2019,

         www.i2tutorials.com/top-machine-learning-interview-questions-and-answers/what-is-

         sigmoid-function-and-explain-in-detail/. [Accessed 13 December 2019].

[46]   Shalizi, Cosma. Logistic Regression.

         https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

[47]   Gupta, Shailaja. "Pros and Cons of Various Classification ML Algorithms." Medium,

         Towards Data Science, 28 Feb. 2019, towardsdatascience.com/pros-and-cons-of-various-

         classification-ml-algorithms-3b5bfb3c87d6. [Accessed 5 February 2020].

[48]   Gron, Aurlien. Hands-On Machine Learning with Scikit-Learn and TensorFlow:

         Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc. April

         2017.

[49]   Hershy, Andrew. "Gini Index vs Information Entropy." Medium, Towards Data Science,

         10 Jul. 2019, towardsdatascience.com/gini-index-vs-information-entropy-7a7e4fed3fcb.

         [Accessed 10 February 2020].

[50]   K, Dhiraj. "Top 5 Advantages and Disadvantages of Decision Tree Algorithm." Medium,
       Medium, 26 May 2019, medium.com/@dhiraj8899/top-5-advantages-and-
       disadvantages-of-decision-tree-algorithm-428ebd199d9a. [Accessed 7 February 2020].

[51]   Yiu, Tony. "Understanding Random Forest." Medium, Towards Data Science, 14 August
       2019, towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed
       10 February 2020].

[52]   Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to
       Statistical Learning: with Applications in R. New York: Springer, 2013.

[53]   Kawerk, Elie. Bagging. Campus.datacamp.com. PowerPoint Presentation.

[54]   Schott, Madison. "Random Forest Algorithm for Machine Learning." Medium, Capital
       One Tech, 25 April 2019, medium.com/capital-one-tech/random-forest-algorithm-for-
       machine-learning-c4b2c8cc9feb. [Accessed 10 February 2020].

[55]   "Many Heads Are Better Than One: The Case For Ensemble Learning." KDnuggets, Sept
       2019, www.kdnuggets.com/2019/09/ensemble-learning.html. [Accessed 10 February
       2020].

[56]   Desarda, Akash. "Understanding AdaBoost." Medium, Towards Data Science, 17 Jan.
       2019, towardsdatascience.com/understanding-adaboost-2f94f22d5bfe. [Accessed 11
       February 2020].

[57]   Kawerk, Elie. Adaboost. Campus.datacamp.com. PowerPoint Presentation.

[58]   Emer, Eric. Boosting (Adaboost Algorithm)
       http://math.mit.edu/~rothvoss/18.304.3PM/Presentations/. [Accessed 11
       February 2020]. PowerPoint Presentation.

[59]   Palmer, C.A., Oman, C.L., Park, A.J., Luppens, J.A., 2015. The U.S. Geological Survey

        Coal Quality (COALQUAL) Database Version 3.0. Data Series, Reston, VA.

[60]   Lin, R., Soong, Y., Granite, E.J., 2018. Evaluation of Trace Elements in U.S. Coals Using

        the USGS COALQUAL Database Version 3.0. Part I: . Rare earth elements and yttrium

        (REY). International Journal of Coal Geology In press.

[61]   Finkelman, R.B., 1993. Trace and minor elements in coal. Org. Geochem. 593–607

        (Springer).

[62]   Taylor, S.R., McLennan, S.M., 1995. The geochemical evolution of the continental crust.

        Rev. Geophys. 33, 241–265.

[63]   Connor, J.J., Keith, J.R., Anderson, B.M., 1976. Trace-metal variation in soils and

        Sagebrush in the Powder River basin Wyoming and Montana. J. Res. US Geol. Survey 4,

        49–59.

[64]   "Annual Coal Report, 2017." U.S. Energy Information Administration. pg 68.

        https://www.eia.gov/coal/annual/archive/05842017.pdf

[65]   Gluskoter, H.J., Ruch, R.R., Miller, W.G., Cahill, R.A., Dreher, G.B., Kuhn, J.K., 1977.

        Trace Elements in Coal: Occurrence and Distribution. Circular no. 499.

[66]   Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE:

        Synthetic Minority Over-sampling Technique. 2002.

[67]   Chawla, Nitesh. "SMOTE." Carnegie Mellon University, School of Computer Science. 2

        Jun. 2002. https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-

        html/node6.html. [Accessed 16 February 2020].

[68]   Draelos, Rachel, et al. "Measuring Performance: The Confusion Matrix." Glass Box, 18

         May 2019, glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-

         matrix/.

[69]   "Regularization in Logistic Regression: Better Fit and Better Generalization?"

         KDnuggets, Jun. 2016. www.kdnuggets.com/2016/06/regularization-logistic-

         regression.html. [Accessed 11 February 2020]

[70]   "Sklearn.model_selection.GridSearchCV." Scikit, scikit-

         learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

         [Accessed 12 February 2020].

[71]   "Sklearn.model_selection.RandomizedSearchCV." Scikit, scikit-

         learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.

         [Accessed 12 February 2020].

[72]   INAA Services at Washington University,

         http://epsc.wustl.edu/~rlk/papers/inaa/inaaservices.htm. [Accessed 12 February 2020].

Group-Mineral	Formula	Carbonatite	Alkaline Intrusion-Related	Placer	Phosphorite
Oxides					
Aeschynite	$(Ln,Ca,Fe)(Ti,Nb)_2(O,OH)_6$		X		
Euxenite	$(Y,Ln,Ca)(Nb,Ta,Ti)_2(O,OH)_6$		X	X	
Fergusonite	$YNbO_4$		X		
Carbonates					
Bastnäsite	$(Ln,Y)CO_3F$	X	X		
Parisite	$Ca(Ln)_2(CO_3)_3F_2$	X	X		
Synchisite	$Ca(Ln,Y)(CO_3)_2F$	X	X		
Tengerite	$Y_2(CO_3)_3 \bullet n(H_2O)$		X		
Phosphates					
Apatite	$(Ca,Ln)_5(PO_4)_3(OH,F,Cl)$	X	X		X
Monazite	$(Ln,Th)PO_4$	X	X	X	
Xenotime	YPO_4		X	X	
Silicates					
Allanite	$(Ln,Y,Ca)_2(Al,Fe^{3+})_2(SiO_4)_3(OH)$		X		
Eudialyte	$Na_4(Ca,Ce)_2(Fe^{2+},Mn^{2+},Y)ZrSi_8O_{22}(OH,Cl)_2$		X		
Thalenite	$Y_2Si_2O_7$		X		
Zircon	$(Zr,Ln)SiO_4$		X	X	

Ln: Lanthanide (a.k.a. REE)

Table A-1. REY-bearing mineral deposits classification table [16].