

Glass Dataset

CST8390\_23W Assignment 3



ALGONQUIN COLLEGE

Computer Engineering Technology

Business Intelligence and Data Analytics

Hoang, John – 040896360

Kingdom, Neil – 040967309

Date: March 18, 2023

Prepared for:

Dr. Anu Thomas

Table of Contents

[List of Tables and Figures 3](#_Toc130072781)

[Introduction 4](#_Toc130072782)

[Data Understanding 5](#_Toc130072783)

[Data Preparation 6](#_Toc130072784)

[Local Outlier Factor and Isolation Forest Data Preparation 7](#_Toc130072785)

[Visualize the Data 8](#_Toc130072786)

[Preparing the ARFF File for Outlier Detection 10](#_Toc130072787)

[Modeling and Evaluation 11](#_Toc130072788)

[Clustering 12](#_Toc130072789)

[K-Mean Clustering 12](#_Toc130072790)

[Farthest First Clustering 14](#_Toc130072791)

[Outlier Detection 16](#_Toc130072792)

[Local Outlier Factor 16](#_Toc130072793)

[Isolation Forest 18](#_Toc130072794)

[Discussion of Results 19](#_Toc130072795)

[Outlier Detection Results 19](#_Toc130072796)

[Outlier Detection and Clustering 21](#_Toc130072797)

[Additional Research 22](#_Toc130072798)

[Conclusion 25](#_Toc130072799)

[Division of Work 25](#_Toc130072800)

[References 26](#_Toc130072801)

[Appendix 1 28](#_Toc130072802)

[Appendix 2 30](#_Toc130072803)

[Appendix 3 31](#_Toc130072804)

[LOF Results 31](#_Toc130072805)

[ISF Results 36](#_Toc130072806)

[Appendix 4 42](#_Toc130072807)

## List of Tables and Figures

[Table 1 Initial Glass data set Attributes 5](#_Toc130072487)

[Table 2 Glass data set Class 5](#_Toc130072488)

[Table 3 List of Compound and Description [12] 6](#_Toc130072489)

[Table 4 Class distribution 6](#_Toc130072490)

[Table 5 Summary Statistics 7](#_Toc130072491)

[Table 6 Outlier Selection 10](#_Toc130072492)

[Table 7 Simple K-Mean Summary with Seed of 1 13](#_Toc130072493)

[Table 8 K-mean 5 Clusters Analysis 14](#_Toc130072494)

[Table 9 Farthest First Clusters Analysis 15](#_Toc130072495)

[Table 10 Farthest First Summary 15](#_Toc130072496)

[Table 11 LOF Accuracy Table 17](#_Toc130072497)

[Table 12 ISF Accuracy Table 18](#_Toc130072498)

[Table 13 Ensemble Accuracy Table 21](#_Toc130072499)

[Table 14 SimpleKMeans Table of k Clusters Seed of 10 24](#_Toc130072500)

[Figure 1 Weka Visualizing Chart - Id to Calcium 8](#_Toc130072501)

[Figure 2 Weka Visualizing Chart - Mg to Ba 9](#_Toc130072502)

[Figure 3 Weka Visualizing Chart - Al to K 9](#_Toc130072503)

[Figure 4 List of Selected Outliers By Id 11](#_Toc130072504)

[Figure 5 Elbow Chart 13](#_Toc130072505)

[Figure 6 LOF Prediction Margin 16](#_Toc130072506)

[Figure 7 LOF Confusion Matrix 17](#_Toc130072507)

[Figure 8 LOF Prediction Weka Summary 17](#_Toc130072508)

[Figure 9 ISF Confusion Matrix 18](#_Toc130072509)

[Figure 10 ISF Prediction Weka Summary 18](#_Toc130072510)

[Figure 11 Type 2 Attribute Na Normalized Chart 19](#_Toc130072511)

[Figure 12 Type 2 Attribute Na Standardized Chart 19](#_Toc130072512)

[Figure 13 Ensemble LOF and ISF Type 2 By Largest Value of Na – Instance ID 71 19](#_Toc130072513)

[Figure 14 Ensemble Results Correctly Classified Outliers 20](#_Toc130072514)

[Figure 15 Ensemble - Correctly Classified Outlier On Either LOF or ISF 20](#_Toc130072515)

[Figure 16 K-mean k=9 21](#_Toc130072516)

[Figure 17 K-mean k=10 21](#_Toc130072517)

[Figure 18 K-mean k=11 22](#_Toc130072518)

[Figure 19 K-mean k=16 22](#_Toc130072519)

[Figure 20 Weka Cluster With Simple K Mean Seed of 1 23](#_Toc130072520)

[Figure 21 Elbow Chart on Weka With Seed of 10 24](#_Toc130072521)

[Figure 22 Ensemble - Incorrectly Classified by Either LOF or ISF 42](#_Toc130072522)

[Figure 23 Ensemble - Outliers Not Identified by LOF or ISF 42](#_Toc130072523)

# Introduction

In this report, the glass data set is processed for unsupervised learning to alleviate the classification problem from the data set, supplied by UCI Machine Learning Repository, by assessing the available data and identifying outliers [1]. The properties of glass and specific composition of glass is crafted based on the use case. The glass data set obtained from Machine Learning Repository(MLR), contains 7 class label information identifying properties of glass that differ in use cases and processing. Float glass is processed by floating liquid glass on a bed of molten metal producing a flat sheet of high visual quality. For this, float glass is highly considered for use on building windows and vehicle windshields but occasionally windows and windshields are processed by plate glass or sheet glass. Other glass objects are often crafted using the glassblowing technique for ease of molding the glass to the desired shape as in the case with vehicle headlights, containers and tableware [2].

The objective is to solve a classification problem by discovering new information from the data set that are not apparent and to find meanings in the newly acquired information. A quick approach is to implement clustering on the glass data set by grouping the data by it’s closeness to other data, in other words the regional density of the data. Weka and Excel are the tools considered for building the unsupervised model. Excel is used to prep and format the data for compatibility with Weka’s ARFF file type to perform unsupervised learning on [3].

Weka’s built-in algorithm for clustering, k-mean and farthest first algorithm is considered for assessing the data set. The K-mean algorithm is a greedy algorithm for partitioning the samples into k clusters by minimizing the sum of squared distances to the clusters centre and is a quick and simple approach to analyze data. Weka’s simple k-mean replicate the K-mean algorithm and provides an option to selecting the distance measuring formula to use on the sample. Some distance measuring methods are: Euclidean distance (in Weka uses the mean of the data set) and Manhattan distance (in Weka uses the median of the data set) [4]. In this report the Euclidean distance measurement is used. K-mean has an efficiency that is define as having a time complexity of O of number of clusters, times the number of iteration, times the number of instances, times the number of dimensions (O(k\*i\*n\*d)). Weka’s farthest first algorithm follows farthest first traversal algorithm of Hochbaum & Shmoys (1985) to distinguish the min-max diameter of the clustering problem and the metric k-center problem [5]. Compared to k-mean, farthest first algorithm is faster since its time complexity is O(k\*n). The centroid is designated based on the algorithm’s metric space for the first centroid then sets the next cluster centroid at the first farthest point within the region of the current centroid.

The significance of detecting and handling outliers is crucial to improving the quality of the data set and constructing a good predictive model hence solving the classification problem. Therefore, the local outlier factor method (LOF) and isolation forest method (ISF) for outlier detection are considered. Weka’s built-in algorithms does not come with LOF and ISF algorithm, however, Weka has a built-in package manager that allows for installation of LOF and ISF packages [6] [7]. In LOF, the algorithm builds a profile of the data point individually and assigns the data an outlier score. Outlier score is determined as follow: outlier score of approximately 1 indicate the data has similar density as its neighbours, outlier score of less than 1 indicate the data has higher density than its neighbours (inlier) and outlier score more than 1 indicate the data has lower density than its neighbours (outlier) [8]. In ISF, the algorithm isolates each data point individually and assigns the data an outlier or isolation score. Isolation score is determined as follow: isolation score of approximately 1 indicates the data is an outlier, isolation score of less than 0.5 indicate normal observations and an entire data set with isolation score close to 0.5 indicate the entire data set does not have any distinct anomalies [9]. An ensemble evaluation is then conducted on the combined results of LOF and ISF to the glass data set.

At the end of the report, an evaluation of the results of the unsupervised model is provided with explanations. Evaluation and explanation include comparison of results between k-mean and farthest first algorithm, comparison of results between LOF and ISF, and the comparison of results between clustering methods and outlier detection methods.

# Data Understanding

The glass data set contains 214 instances with no missing data. There is 1 class of 7 types, the 4th type of Vehicle non-float processed does not have any records in the data set, and 9 attributes described in Table 1 and 2 below.

Table 1 Initial Glass data set Attributes

|  |  |
| --- | --- |
| **Attribute/Feature** | **Type** |
| Id | Numeric Discrete |
| RI Refractive Index | Numeric Continuous |
| Na Sodium | Numeric Continuous |
| Mg Magnesium | Numeric Continuous |
| Al Aluminum | Numeric Continuous |
| Si Silicon | Numeric Continuous |
| K Potassium | Numeric Continuous |
| Ca Calcium | Numeric Continuous |
| Ba Barium | Numeric Continuous |
| Fe Iron | Numeric Continuous |

Table 2 Glass data set Class

|  |  |
| --- | --- |
| **Class** | **Type** |
| Glass Type | 1 (Window Building Float-Processed) |
| 2 (Window Building Non-Float-Processed) |
| 3 (Window Vehicle Float-Processed) |
| 4 (Window Vehicle Non-Float-Processed) |
| 5 (Container Glass) |
|  | 6 (Tableware Glass) |
|  | 7 (Headlamps Glass) |

Refraction index of a material is a measure of the level of light refracting in and out of the material. It is also a measure of the speed of light travelling in and out of the material and is calculated by the angle of a beam of light changes as it enters the material. The angle of incidence is also known as the reflection and is represented by i. The angle of refraction is the angle measured as the beam of light exit the material from the surface and is represented by r. The relation between the speed of light and the angles is the refractive index and is given by sin I / sin r = v1/v2. For example, a refractive index close to 0 indicate that light travels at the speed of light when traveling through the material where as air has a refractive index of 1 indicates that light is traveling at near maximum speed [10] [11]. The range of refraction index for the glass type discussed in this report are in the range of 1.47 to 1.53.

Table 3 shows a list of the compounds in the glass with a brief description is provided and Table 4 shows the class distribution.

Table 3 List of Compound and Description [12]

|  |  |
| --- | --- |
| **Elements** | **Description** |
| Na Sodium | Soft metal, reactive |
| Mg Magnesium | Lightweight metal |
| Al Aluminum | Lightweight  non-corroding metal |
| Si Silicon | Hard metalloid |
| K Potassium | Soft metal, reactive |
| Ca Calcium | Soft metal |
| Ba Barium | Soft metal absorbs  X-rays |
| Fe Iron | Medium-hard metal,  magnetic |

Table 4 Class distribution

|  |  |  |
| --- | --- | --- |
| **Type** | **Count** | **Pct** |
| 1 (Window Building Float-Processed) | 70 | 32.7% |
| 2 (Window Building Non-Float-Processed) | 76 | 35.5% |
| 3 (Window Vehicle Float-Processed) | 17 | 7.9% |
| 4 (Window Vehicle Non-Float-Processed) | 0 | 0 |
| 5 (Container Glass) | 13 | 6.1% |
| 6 (Tableware Glass) | 9 | 4% |
| 7 (Headlamps Glass) | 29 | 13.6% |

# Data Preparation

On preparing the data set for clustering, normalizing and standardizing the data is considered to compare the data when identifying outliers. Normalizing or Standardizing the data set entirely or partially produces the same result when clustering on Weka and has minimum improvement on the time complexity of the data set, therefore, the original data is used in the unsupervised learning process. It is recommended to normalized data when different units are used on attribute’s values in the data set. In the glass data set, elements are measured using the same units except for the refractive index which is a calculated value, however, it’s value is consistent across all data. It may be reasoned that the data set can be fine tuned so that data has a visual consistency across the attributes but it seems rather confusing when viewing and comparing atomic elements measured in the same unit of scale. In cases where it is highly recommended to normalized is when working with data that consists of different types, for example age, height and weight, which are all measured by different scaling unit. Normalizing would assist the analysist ability to gauge and compare the contribution of the attributes by it’s distance measurement weight [13] [3].

Table 5 Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute** | **Min** | **Max** | **Mean** | **SD** | **Correlation with Class** |
| RI | 1.5112 | 1.5339 | 1.5184 | 0.0030 | -0.1642 |
| Na | 10.73 | 17.38 | 13.4079 | 0.8166 | 0.5030 |
| Mg | 0 | 4.49 | 2.6845 | 1.4424 | -0.7447 |
| Al | 0.29 | 3.5 | 1.4449 | 0.4993 | 0.5988 |
| Si | 69.81 | 75.41 | 72.6509 | 0.7745 | 0.1515 |
| K | 0 | 6.21 | 0.4971 | 0.6522 | -0.0100 |
| Ca | 5.43 | 16.19 | 8.9570 | 1.4232 | 0.0007 |
| Ba | 0 | 3.15 | 0.1750 | 0.4972 | 0.5751 |
| Fe | 0 | 0.51 | 0.0570 | 0.0974 | -0.1879 |

Table 5 shows a summary of the glass data set attributes statistics and is useful in identifying outliers. The min, max, standard deviation and correlation shows the dispersion of data and the mean provides a view of the central tendency.

## Local Outlier Factor and Isolation Forest Data Preparation

Weka’s LOF and ISF algorithm requires a nominal binary attribute to apply outlier detection on. Outliers are provided with ordinal value of “yes” for outliers and value of “no” for instances that are normal [6] [7].

There are many ways to identify outliers all of which does not guarantee that the data under investigation is an outlier. Although there are no conclusive rules to identifying outliers, there are general rules in identifying outliers. The most basic method is by visually assessing the charts and graphs from the data set and further narrowing down and comparing the instances by its attribute’s values. Another method is by assessing the data set by groups. In Excel, filtering the data by type then by each of the attributes values ordered by highest to lowest and vice versa makes it easy to identify outliers. The idea is to locate values that are separated from the rest of the data values. Most of the time the min and max values of an attribute are considered outliers but it is not always the case. On assessing the data, there are cases where application of a combination of methods is required to identify outliers. For example, comparing the instance to it’s statistical summary, shown in Appendix 1, or by comparing the attribute value to it’s normalized and standardized statistics. Identifying outliers manually is about visualizing the grouping of the data and then make a judgement on whether the data is an irregular or a normal value [14].

### Visualize the Data

The id column contains charts that describe the attribute visually and is available in Weka’s visualize tab. With id on the x axis, the selected y attribute is scaled and colour coded by class label. Figure 1 shows the comparison of classes to the quantity of calcium. Since the x-axis is an id number, it is used to spread the data point out so that each individual points are easier to compare. In consideration of the calcium properties, imagine that the x-axis is removed, the resulting chart is a 1-dimensional chart that shows the classes that are inseparable by measurement of calcium, such as, class 1, 3 and 7. Class 2, non-float processed windows for building, class 5, container glass, and class 6, tableware glass, shows similarities in quantity of calcium where a considerable amount of records has values above 10.81.



Figure 1 Weka Visualizing Chart - Id to Calcium

Figure 2 is a chart that compare the data set magnesium to barium properties. It shows that there is a clear separation between class 7, headlamp glass, with the other classes except for some similarities with class 5, container glass. Classes 1, 2, 3 and 6 all shows clusters that are overlapping and indistinguishable.

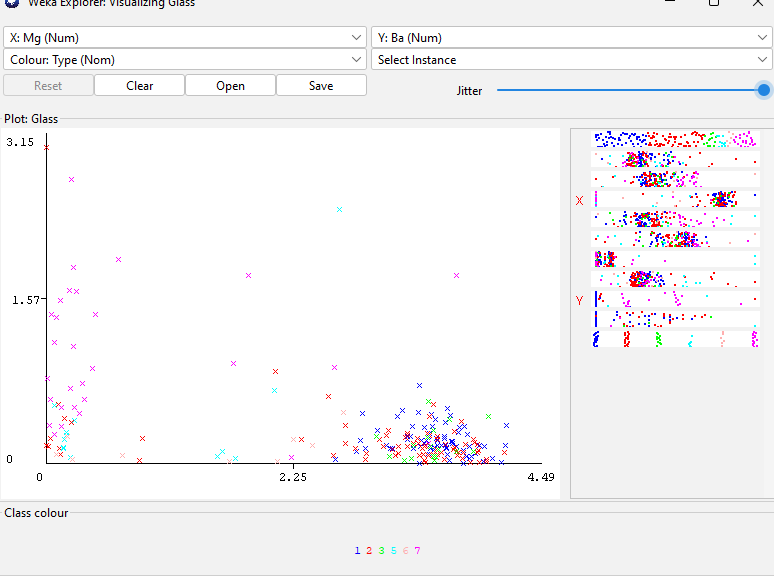


Figure 2 Weka Visualizing Chart - Mg to Ba

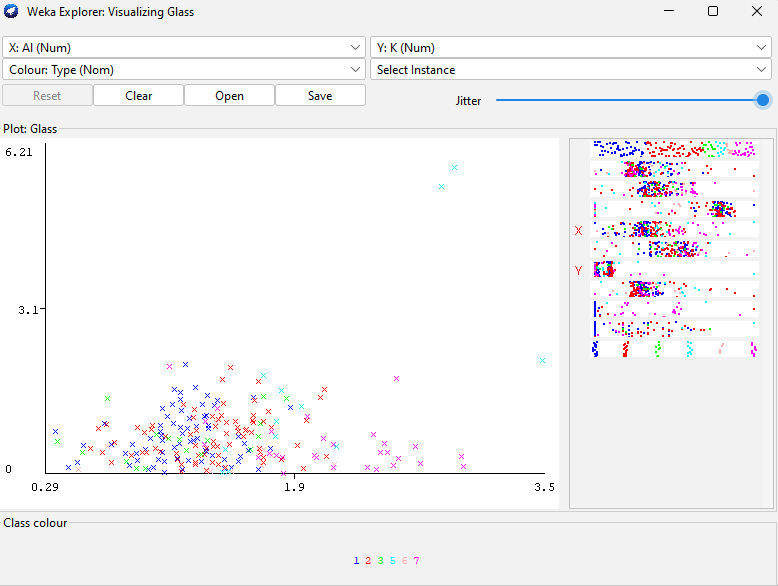


Figure 3 Weka Visualizing Chart - Al to K

Figure 3 is a comparison of the data set aluminum to potassium properties. It shows clusters of class 1, 2 and 7 (float processed windows for building, non-float processed windows for building and headlamp glass, respectively). On further examination, class 3 (float processed vehicle windshield glass) is inseparable from class 1 and 2 for both float and non float processed windows. Class 5 seems like a cluster above class 1 and 2 with a positive correlation. Class 6 cannot be identified since there are not enough instances of the class.

### Preparing the ARFF File for Outlier Detection

Table 6 shows the final consideration of the outliers from manual selection. There are some other data that qualifies as outliers but were excluded, on assessment explained in evaluation, to keep the count of outlier low. It makes sense to keep the percentage of outlier low to about 20%, otherwise, the sample is tainted [15]. Since class label 6, tableware glass, only has 9 instances it is considered an outlier class as shown in Figure 3 class 6 could not be identified. All the data considered as outliers were processed manually by examining the original value, the statistical values, the class label, the normalized statistic values and the standardized statistic values (min, max, mean and standard deviation).

Table 6 Outlier Selection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Categorical** | **Type** | | | | |
| Outlier | Categorical Ordinal | Class | Count | Percent | Total |
| yes | 1 – Windows FP | 16 | 7.34% | Count |
| 2 – Windows NFP | 10 | 4.59% | 48 |
| 3 – Vehicle WS FP | 1 | 0.46% |
| 5 – Container Glass | 5 | 2.29% | Percent |
| 6 – Tableware Glass | 9 | 4.13% | 22.02% |
| 7 – Headlamp | 7 | 3.21% |
| no | 1 – Windows FP | 54 | 24.77% | Count |
| 2 – Windows NFP | 66 | 30.28% | 170 |
| 3 – Vehicle WS FP | 16 | 7.34% |
| 5 – Container Glass | 8 | 3.67% | Percent |
| 6 – Tableware Glass | 0 | 0% | 77.98% |
| 7 – Headlamp | 22 | 10.09% |

Below is a query of the glass data set using MariaDb for visualization

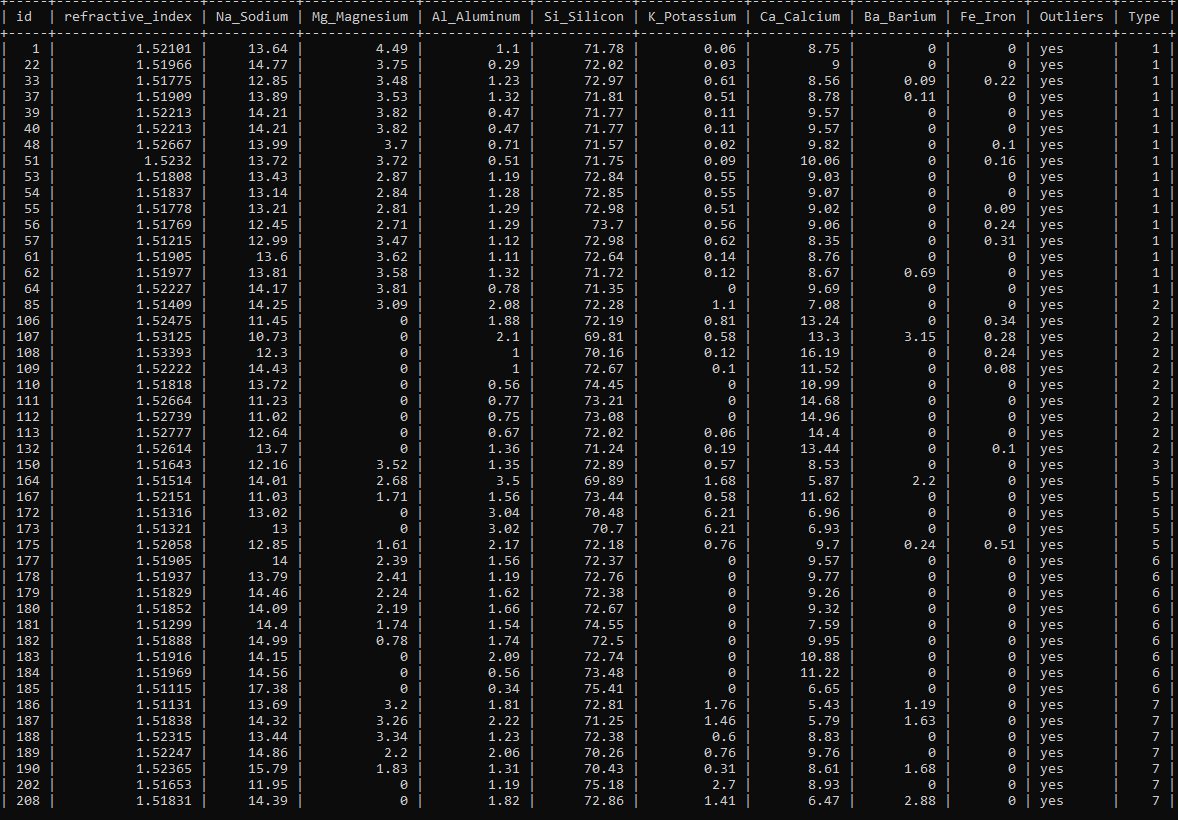


Figure 4 List of Selected Outliers By Id

# Modeling and Evaluation

Two different approach are considered for modeling and evaluation, clustering and outlier detection. Weka’s simple k-mean algorithm and Weka’s farthest first algorithm is processed on the data without the binary outlier categorical attribute. Local Outlier Factor and Isolation Forest package add-on is used for outlier detection. LOF and ISF algorithm is used on the processed data with the binary outlier categorical attribute.

## Clustering

Clustering is performed on continuous data for both simple k-mean and farthest first algorithm. Since the Id attribute does not have any relevance in classifying the clusters the Id attribute is ignored using Weka’s clustering feature “Ignore attributes” by selecting the Id item. Weka’s “Ignore attributes” features allows for multiple item selection. It is not necessary to classify by class type so the label class is ignored by holding the CTRL key and selecting the type item from the “Ignore attributes” feature. Same with the outlier attribute. The “Use training set” is selected for the cluster mode. Both k-mean and farthest first clustering method is provided with a seed of 1, a starting number of k clusters of 2 and for k-mean the Euclidean distance formula is used. All other settings are left as its default settings.

### K-Mean Clustering

On running k-mean with 2 clusters, Weka’s simple k mean selects 2 random point using the seed of 1 with an initial centroid for cluster 0 at 1.51763,12.8,3.66,1.27,73.01,0.6,8.56,0,0 and cluster 1 at 1.51977,13.81,3.58,1.32,71.72,0.12,8.67,0.69,0, ordered in respect to the list of attributes. The divergence took 7 iterations with the final iteration centroid listed in Table 7. The process is repeated by incrementing the number of clusters and running the algorithm until an elbow point is observed. Excel is used to plot the elbow chart with the number of clusters on the x axis and the “within cluster sum of squared errors” on the y axis. Figure 5 shows the elbow point at the 5th cluster. This makes sense since class label 6 is considered an outlier and class label 4 is not available in the data set. The classes that are considered are class label 1, 2, 3,5 and 7 for a total of 5 classes.

Table 8 is considered on examination of the elbow point cluster. Cluster 0 and 1 contain class label 1,2,3,5 and 1,2,3,7, respectively, which makes sense considering Figure 1, 2 and 3 showing clusters of these classes in the same area. In both clusters 0 and 1, the odd class label, 5 and 7, shows that it is 100% outlier, see Table 8. Cluster 2 contain classes 5,6 and 7, this is interesting considering Figure 2, class 5 and 7 are separable from the other classes. Cluster 3 contain all the classes and seems to be the catch-all cluster at the centre of the data where classes 6 and 7 are 100% outliers. Cluster 4 seems to be the outlier cluster where 58% of the instances are outliers the majority of which is class 2 with 82% outliers and class 6 with 100% outlier with 1 instance.

The elbow point is also referred to as the optimal number of cluster point. In Figure 5 elbow chart, the elbow point is located by looking at where the line shows the steepest drop, the point at the end of the steep is the elbow point. This is confirmed by the “within cluster sum of squares errors” average difference to the elbow point is 4.35324 while the average difference of the subsequent points is 1.05647. Conforming with the elbow methodology that the “within clusters sum of squared errors’ are smaller in subsequent number of k clusters [16]. There are other methods that can be used to find the optimal number of clusters point such as, the Silhouette Method, the Bayesian Information Criterion (BIC) and the Gaussian Mixture Model (GMM) but they are not discussed in this report [17].

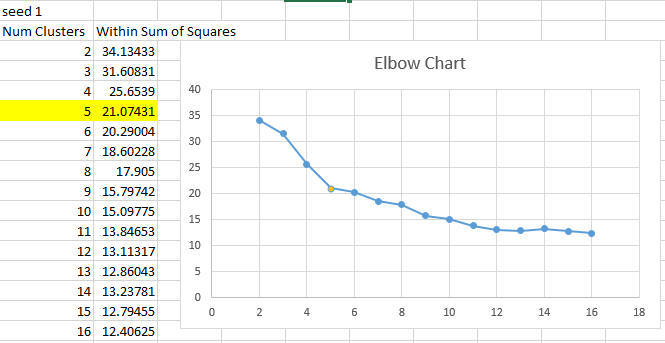


Figure 5 Elbow Chart

Table 7 Simple K-Mean Summary with Seed of 1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Simple K-Mean Weka’s Algorithm | | | | | | | | | | |
| k | Instances | Final Centroid (chronological order per row starting at centroid 0,1,2 …) | | | | | | | | |
| RI | Na | Mg | Al | Si | K | Ca | Ba | Fe |
| 2 | 162(76%) | 1.5181 | 13.2811 | 3.4541 | 1.3104 | 72.6122 | 0.4957 | 8.624 | 0.028 | 0.061 |
| 52 (24%) | 1.5191 | 13.8027 | 0.2871 | 1.864 | 72.7715 | 0.5013 | 9.9954 | 0.6333 | 0.0462 |
| 3 | 135(63%) | 1.5184 | 13.1864 | 3.4963 | 1.3689 | 72.7666 | 0.5602 | 8.3804 | 0.0316 | 0.0601 |
| 39 (18%) | 15.227 | 13.4321 | 2.4851 | 1.1231 | 71.9046 | 0.2262 | 10.6233 | 0.0562 | 0.0782 |
| 40 (19%) | 1.5175 | 14.1317 | 0.1393 | 2.0152 | 72.9882 | 0.548 | 9.2782 | 0.7752 | 0.0258 |
| 4 | 133(62%) | 1.5173 | 13.1798 | 3.4384 | 1.3905 | 72.7974 | 0.5616 | 8.3949 | 0.028 | 0.0623 |
| 29 (14%) | 1.522 | 13.8917 | 3.5241 | 0.9341 | 71.679 | 0.1879 | 9.5762 | 0.0855 | 0.0521 |
| 27 (13%) | 1.5161 | 14.5252 | 0.1652 | 2.283 | 72.9696 | 0.5904 | 8.3448 | 1.0319 | 0.0144 |
| 25 (12%) | 1.5224 | 12.8532 | 0.4208 | 1.4216 | 72.6548 | 0.4116 | 11.89 | 0.1356 | 0.0804 |
| 5 | 40 (19%) | 1.5181 | 13.0777 | 3.3975 | 1.2957 | 72.706 | 0.5248 | 8.7345 | 0.0253 | 0.2178 |
| 24 (11%) | 1.5222 | 13.9863 | 3.5525 | 0.9213 | 71.6083 | 0.1717 | 9.55 | 0.1033 | 0.0196 |
| 31 (14%) | 1.5163 | 14.3913 | 0.169 | 2.2094 | 73.0477 | 0.6326 | 8.5561 | 0.8987 | 0.0126 |
| 100 (47%) | 1.5172 | 13.2257 | 3.4181 | 1.4212 | 72.7878 | 0.5639 | 8.3544 | 0.0296 | 0.0113 |
| 19 (9%) | 1.5236 | 12.7263 | 0.3305 | 1.2979 | 72.4842 | 0.2768 | 12.5016 | 0.1658 | 0.0789 |
| 6 | 40 (19%) | 1.5181 | 13.0777 | 3.3975 | 1.2957 | 72.706 | 0.5248 | 8.7345 | 0.0253 | 0.2178 |
| 22 (10%) | 1.5224 | 13.9945 | 3.5455 | 0.8914 | 71.5759 | 0.1614 | 9.615 | 0.1077 | 0.0214 |
| 30 (14%) | 1.5163 | 14.4083 | 0.1153 | 2.2233 | 73.046 | 0.6537 | 8.5523 | 0.9033 | 0.013 |
| 16 (7%) | 1.5171 | 13.5856 | 2.5769 | 1.6162 | 72.7831 | 0.535 | 8.5156 | 0.2325 | 0.0056 |
| 19 (9%) | 1.5236 | 12.7263 | 0.3305 | 1.2979 | 72.4842 | 0.2768 | 12.5016 | 0.1658 | 0.0789 |
| 87 (41%) | 1.5173 | 13.1824 | 3.5589 | 1.3856 | 72.7733 | 0.5563 | 8.3394 | 0.0013 | 0.012 |

Table 8 K-mean 5 Clusters Analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Simple K-mean – 5 Clusters, seed = 1 | | | | | | | |
| Cluster | Class Label | Count | Percent | Number of Outlier | | Total | Total Outlier |
| 0 | 1 | 15 | 37.5% | 3 | 20% | Count | |
| 2 | 21 | 52.5% | 0 | 0% | 40 | 4 |
| 3 | 3 | 7.5% | 0 | 0% | Percent | |
| 5 | 1 | 3% | 1 | 100% | 19% | 10% |
| 1 | 1 | 17 | 71% | 8 | 47% | Count | |
| 2 | 2 | 8% | 0 | 0% | 24 | 11 |
| 3 | 2 | 8% | 0 | 0% | Percent | |
| 7 | 3 | 13% | 3 | 100% | 11% | 46% |
| 2 | 5 | 4 | 13% | 3 | 75% | Count | |
| 6 | 3 | 10% | 3 | 100% | 31 | 8 |
| 7 | 24 | 77% | 2 | 8% | Percent | |
| N/A | N/A | N/A | N/A | N/A | 14% | 26% |
| 3 | 1 | 38 | 38% | 4 | 11% | Count | |
| 2 | 42 | 42% | 1 | 2% | 100 | 13 |
| 3 | 12 | 12% | 1 | 8% |
| 5 | 1 | 1% | 0 | 0% | Percent | |
| 6 | 5 | 5% | 5 | 100% | 47% | 13% |
| 7 | 2 | 2% | 2 | 100% |
| 4 | 2 | 11 | 58% | 9 | 82% | Count | |
| 5 | 7 | 37% | 1 | 14% | 19 | 11 |
| 6 | 1 | 5% | 1 | 100% | Percent | |
| N/A | N/A | N/A | N/A | N/A | 9% | 58% |

### Farthest First Clustering

Farthest first clustering is considered when there are concerns on computational power and time to quickly detect outliers from a large data set. Table 9 is prepared to observe the count of outliers per class label for each clusters in k and the percentage of instances per cluster that are outliers. Table 10 is a summary of clusters in k including the designation of the final cluster’s centroid values. Farthest first algorithm randomly designates the first farthest point of current centroid for its next centroid, any prior centroids retains its position on the axis on increments to the number of k clusters when fed with the same seed value. In short, on increment of k number of clusters the new added point is observed, k-center metric re-establishes a diameter for the centroid and data points are updated based on k-center metric [5].

Table 9 shows that on incrementing k, new clusters are formed with 100% of the clusters being outliers and cluster 0 consisting of the majority of the data set instances. On closer examination, the percentage of outliers in cluster 0 decreases as k is increased. Recall Figure 2 and reflect on cluster 3 in Table 9, the cluster picked up classes 5 and 7 data point because the 2 classes shows to be inseparable. Observe cluster 5 in k 6 and reflect on Figure 3, class label 5 shows a positive correlation with class label 1 and 2 but is closer to class label 2.

Table 9 Farthest First Clusters Analysis

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **k** | **Cluster** | **Class Label And Count** | | | | | | **Outlier** | |
| **1** | **2** | **3** | **5** | **6** | **7** | **Outlier Count** | **Percent** |
| 2 | 0 | 70 | 73 | 17 | 13 | 9 | 29 | 45 | 21% |
| 1 |  | 3 |  |  |  |  | 3 | 100% |
| 3 | 0 | 70 | 73 | 17 | 10 | 9 | 29 | 42 | 19% |
| 1 |  | 3 |  |  |  |  | 3 | 100% |
| 2 |  |  |  | 3 |  |  | 3 | 100% |
| 4 | 0 | 70 | 73 | 17 | 10 | 9 | 6 | 40 | 18% |
| 1 |  | 3 |  |  |  |  | 3 | 100% |
| 2 |  |  |  | 2 |  |  | 2 | 100% |
| 3 |  |  |  | 1 |  | 23 | 3 | 100% |
| 5 | 0 | 70 | 72 | 17 | 10 | 7 | 6 | 37 | 17% |
| 1 |  | 3 |  |  |  |  | 3 | 100% |
| 2 |  |  |  | 2 |  |  | 2 | 100% |
| 3 |  |  |  | 1 |  | 23 | 3 | 13% |
| 4 |  | 1 |  |  | 2 |  | 3 | 100% |
| 6 | 0 | 70 | 69 | 16 | 8 | 7 | 6 | 35 | 16% |
| 1 |  | 2 |  |  |  |  | 2 | 100% |
| 2 |  |  |  | 2 |  |  | 2 | 100% |
| 3 |  |  |  | 1 |  | 23 | 3 | 13% |
| 4 |  | 1 |  |  | 2 |  | 3 | 100% |
| 5 |  | 4 | 1 | 2 |  |  | 3 | 43% |

Table 10 Farthest First Summary

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Farthest First Weka’s Algorithm | | | | | | | | | | |
| k | Instances | Final Centroids (chronological order per row starting at centroid 0,1,2 …) | | | | | | | | |
| RI | Na | Mg | Al | Si | K | Ca | Ba | Fe |
| 2 | 211 (99%) | 1.51763 | 12.8 | 3.66 | 1.27 | 73.01 | 0.6 | 8.53 | 0.0 | 0.0 |
| 3 (1%) | 1.53125 | 10.73 | 0.0 | 2.1 | 69.81 | 0.58 | 13.3 | 3.15 | 0.28 |
| 3 | 208 (97%) | 1.51763 | 12.8 | 3.66 | 1.27 | 73.01 | 0.6 | 8.56 | 0.0 | 0.0 |
| 3 (1%) | 1.53125 | 10.73 | 0.0 | 2.1 | 69.81 | 0.58 | 13.3 | 3.15 | 0.28 |
| 3 (1%) | 1.51316 | 13.02 | 0.0 | 3.04 | 70.48 | 6.21 | 6.96 | 0.0 | 0.0 |
| 4 | 185 (86%) | 1.51763 | 12.8 | 3.66 | 1.27 | 73.01 | 0.6 | 8.56 | 0.0 | 0.0 |
| 3 (1%) | 1.53125 | 10.73 | 0.0 | 2.1 | 69.81 | 0.58 | 13.3 | 3.15 | 0.28 |
| 2 (1%) | 1.51316 | 13.02 | 0.0 | 3.04 | 70.48 | 6.21 | 6.96 | 0.0 | 0.0 |
| 24 (11%) | 1.51831 | 14.39 | 0.0 | 1.82 | 72.86 | 1.41 | 6.47 | 2.88 | 0.0 |
| 5 | 182 (85%) | 1.51763 | 12.8 | 3.66 | 1.27 | 73.01 | 0.6 | 8.56 | 0.0 | 0.0 |
| 3 (1%) | 1.53125 | 10.73 | 0.0 | 2.1 | 69.81 | 0.58 | 13.3 | 3.15 | 0.28 |
| 2 (1%) | 1.51316 | 13.02 | 0.0 | 3.04 | 70.48 | 6.21 | 6.96 | 0.0 | 0.0 |
| 24 (11%) | 1.51831 | 14.39 | 0.0 | 1.82 | 72.86 | 1.41 | 6.47 | 2.88 | 0.0 |
| 2 (1%) | 1.51115 | 17.38 | 0.0 | 0.34 | 75.41 | 0.0 | 6.65 | 0.0 | 0.0 |
| 6 | 176 (82%) | 1.51763 | 12.8 | 3.66 | 1.27 | 73.01 | 0.6 | 8.56 | 0.0 | 0.0 |
| 2 (1%) | 1.53125 | 10.73 | 0.0 | 2.1 | 69.81 | 0.58 | 13.3 | 3.15 | 0.28 |
| 2 (1%) | 1.51316 | 13.02 | 0.0 | 3.04 | 70.48 | 6.21 | 6.96 | 0.0 | 0.0 |
| 24 (11%) | 1.51831 | 14.39 | 0.0 | 1.82 | 72.86 | 1.41 | 6.47 | 2.88 | 0.0 |
| 3 (1%) | 1.51115 | 17.38 | 0.0 | 0.34 | 75.41 | 0.0 | 6.65 | 0.0 | 0.0 |
| 7 (3%) | 1.52058 | 12.85 | 1.61 | 2.17 | 72.18 | 0.76 | 9.7 | 0.24 | 0.51 |

## Outlier Detection

LOF and ISF both require a binary nominal attribute to determine outliers. LOF can computer categorical attributes and ISF can only be performed on numerical data with the exception of the binary nominal attribute. The setting on Weka is provided on LOF with Euclidian distance, both algorithms are provided with a seed of 1, the test options is set to 10-fold cross-validation and the nominal binary attribute is selected as the outlier comparison attribute. Since the model is not built for classification of class label, the class type is included in the outlier predictor model. However, for ISF a filter nominal to binary filter is applied to the class type to meet the requirement factor of the algorithm.

### Local Outlier Factor

Weka’s LOF algorithm utilizes a prediction margin to detect outliers and it appear that Weka assigns false positive/negative prediction of outliers a negative predictive margin value as in Figure 6 with the complete table provided in Appendix 3. The left side of Figure 6 shows the negative prediction margin values with the last record a positive instance while the right side shows the positive prediction margin.

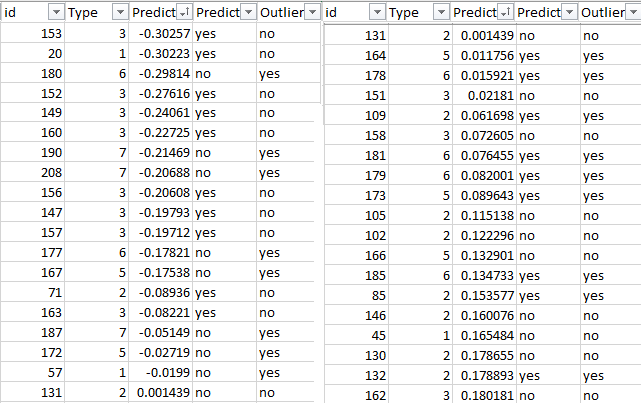


Figure 6 LOF Prediction Margin

Weka’s LOF algorithm successfully predicted outliers with 80.37% correctly classified instances for 172 instances and 42 incorrectly classified instances or 19.63% misclassification. On examining the results further, class label 3 has 81.25% of the misclassification cases with the remaining from class label 2 and 1. This is particularly interesting because recall Figure 2 and 3, class label 3 is overlapped by class label 1 and 2. Since class label 3 has merely 17 instances in a large grouping of data, it is foreseeable that class label 3 can be seen as the outlier within that cluster. Class labels that were correctly classified as outliers are 2, 3,5, 6 and 7 with the majority of correctly classified class label 2 with 9 instances.

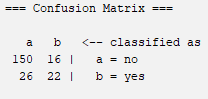


Figure 7 LOF Confusion Matrix

Table 11 LOF Accuracy Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outlier** | **True Positive/Negative Rate** | **False Positive/Negative Rate** | **#Misclassification** | **Total** |
| **a (no)** | 0.904 | 0.096 | 16 | 156 |
| **b (yes)** | 0.458 | 0.542 | 26 | 48 |

Figure 8 is a summary of the LOF outlier detection results. Out of 214 instances 38 were classified as outliers. The bar graphs show the correctly classified instances in blue.

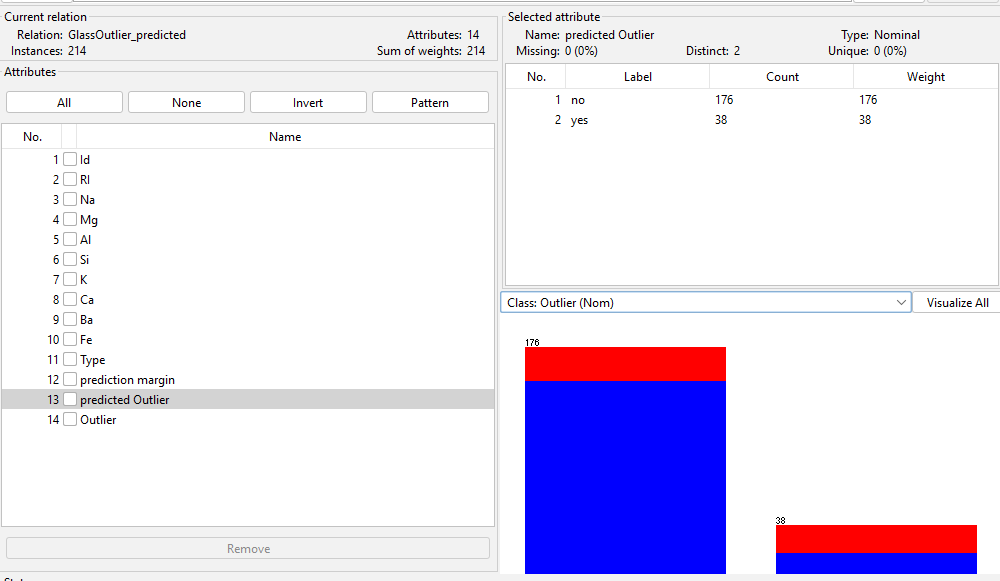


Figure 8 LOF Prediction Weka Summary

### Isolation Forest

The complete ISF results can be found in Appendix 3. It appears that the ISF predictive margin utilizes the same method as the LOF in assigning false positive/negative results a negative predictive margin. Weka’s ISF algorithm successfully predicted outliers with 83.18% correctly classified instances for 178 instances and on 16.82% of the cases 36 instances were misclassified. Unlike LOF, ISF has misclassification of outliers on all class label with 44.4% misclassification of class label 1 outlier as the majority of the cases. The misclassification of outliers is as follow: 16 instances of class label 1, 3 class label 2, 5 class label 3, 7 class label 5, 4 class label 6, and 1 class label 7. The correctly classified outliers are class label 2,5,6 and 7 with class label 2 making up 33.3% of the correctly classified outliers. Classes 5, 6 and 7 are moderately the same at around 21% of the correctly classified outliers.

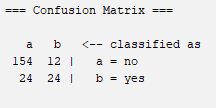


Figure 9 ISF Confusion Matrix

Table 12 ISF Accuracy Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outlier** | **True Positive/Negative Rate** | **False Positive/Negative Rate** | **#Misclassification** | **Total** |
| **a (no)** | 0.928 | 0.072 | 12 | 156 |
| **b (yes)** | 0.500 | 0.500 | 24 | 48 |

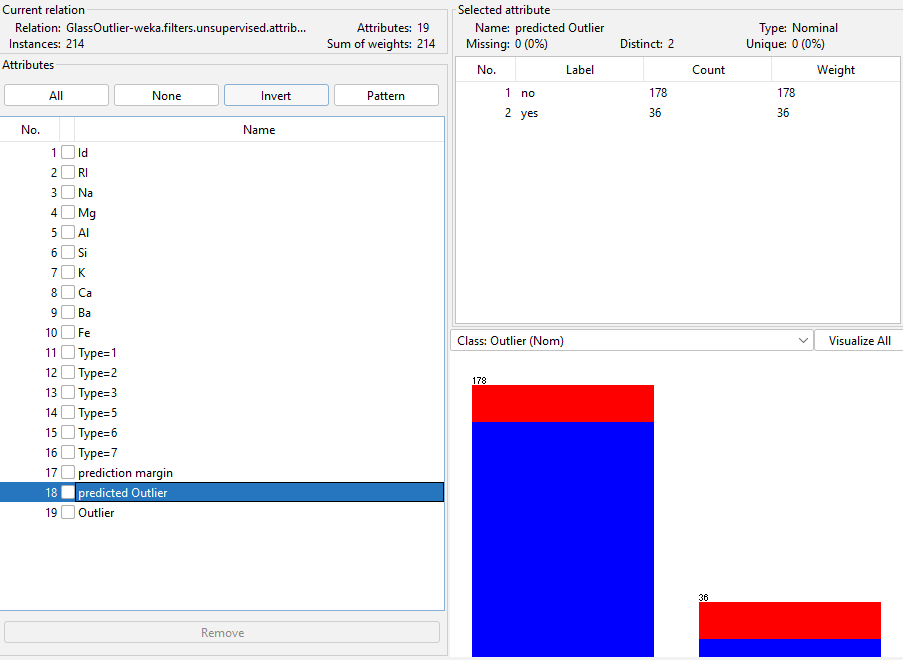


Figure 10 ISF Prediction Weka Summary

# Discussion of Results

In this section the results are gathered and discussed in detail. Results from outlier detection methods are combined and compared on an ensemble basis. The complete ensemble result is provided in Appendix 2. Furthermore, a deep evaluation of the comparison between outlier detection and clustering method is discussed.

## Outlier Detection Results

As mentioned, the selection of outliers in data preparation is evaluated and explained in this section. Below is the normalized chart and standardized chart of Na, Sodium, attribute, Figure 11 and 12 respectively. Figure 13 is a snap of the ensemble results prepared for evaluation. The ensemble results are filtered by class label 2, Na Sodium, and ordered by largest value of Sodium. Although instance id 71 has the highest sodium it was not selected as an outlier due to its behaviour on the standardized chart, Figure 12. It was decided that because it falls between -2 and 2 it can be considered normal. There are a few instances that were considered on this basis to keep the count of outlier low.

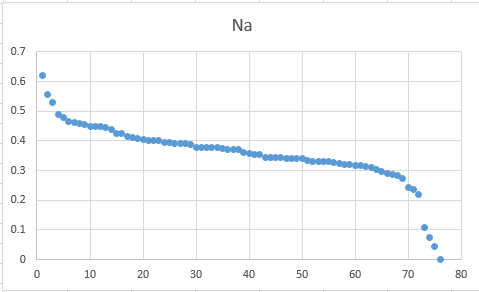


Figure 11 Type 2 Attribute Na Normalized Chart

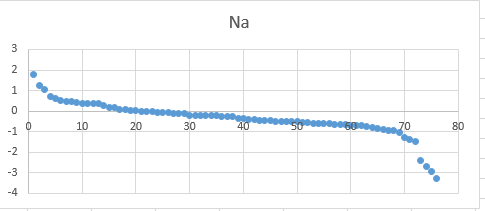


Figure 12 Type 2 Attribute Na Standardized Chart

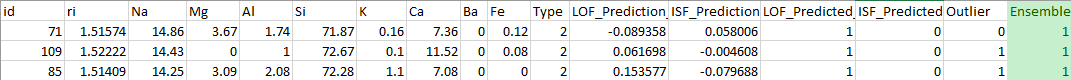


Figure 13 Ensemble LOF and ISF Type 2 By Largest Value of Na – Instance ID 71

On examining the ensemble results, 17 instances were correctly classified and 4 instances were incorrectly classified in the ensemble results shown in Figure 14. 12 instances were correctly classified by either LOF or ISF shown in Figure 15. Additional analysis on the ensemble results include: 4 instances were incorrectly classified as outliers by both LOF and ISF, 20 instances were incorrectly classified as outlier by either LOF or ISF, and 19 instances of outliers were not detected by either LOF or ISF shown in Figure 16 and Figure 17, respectively, in Appendix 4.

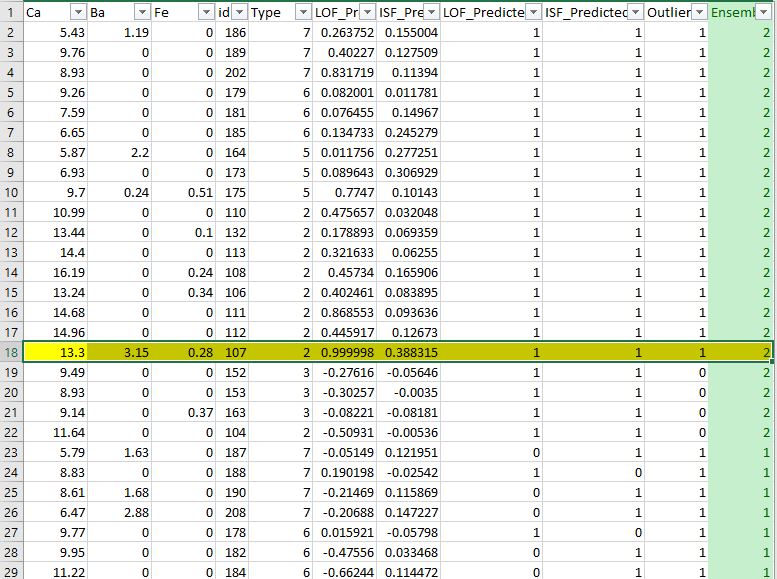


Figure 14 Ensemble Results Correctly Classified Outliers

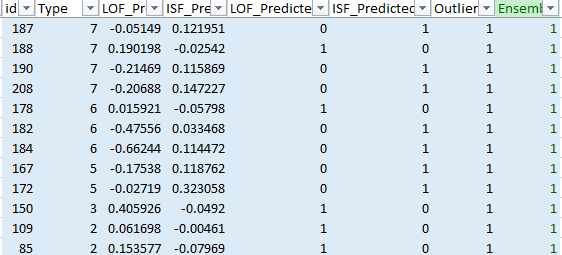


Figure 15 Ensemble - Correctly Classified Outlier On Either LOF or ISF

Table 13 Ensemble Accuracy Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outlier** | **True Positive/Negative Rate** | **False Positive/Negative Rate** | **#Misclassification** | **Total** |
| **a (no)** | 0.976 | 0.0235 | 4 | 166 |
| **b (yes)** | 0.354 | 0.646 | 31 | 48 |

## Outlier Detection and Clustering

Since Weka’s k-mean clustering at the elbow point did not yield any clusters with outlier k is incremented. Figure 18 shows when k-mean is processed with 9 clusters it produces 2 clusters with outliers. Cluster 1 and cluster 4 with cluster 1 having mostly class label 7 outliers and cluster 4 purely class label 2 outliers. Incrementing k-mean with 10 and 11 clusters does not seem to produce better clustering of outliers as shown in Figure 19 and Figure 20. However, setting k-mean clusters at 16 produce 3 clusters of outliers, namely cluster 1, 2 and 6 with cluster 5 short of 3 instances for it to be an outlier cluster.

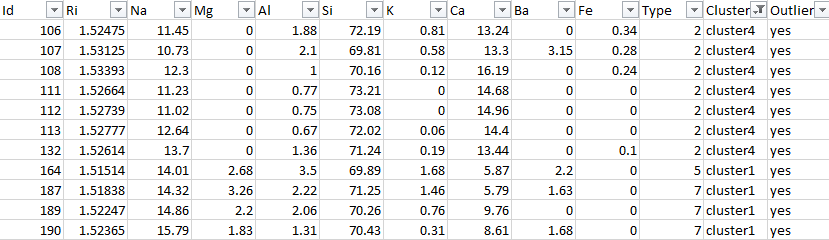


Figure 16 K-mean k=9

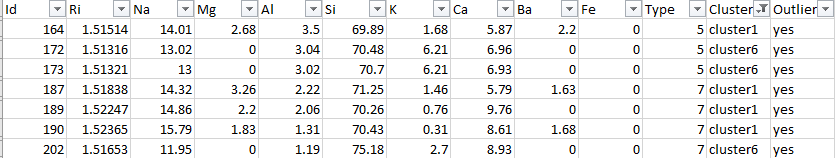


Figure 17 K-mean k=10

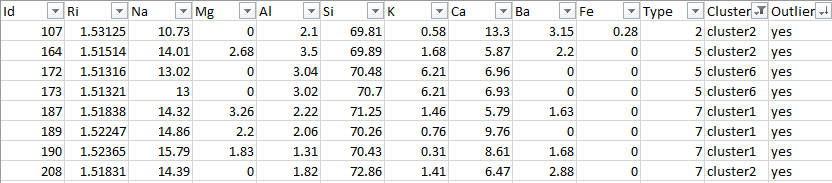


Figure 18 K-mean k=11

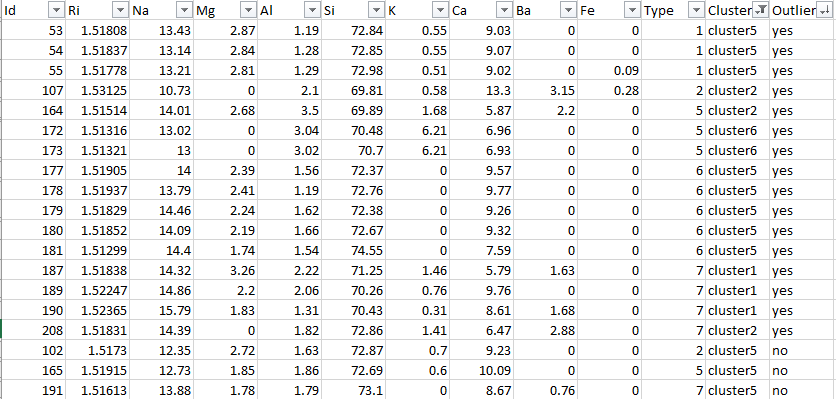


Figure 19 K-mean k=16

## Additional Research

On previous research on finding the optimal k number of clusters a seed of 1 was used on Weka’s simple k-mean clustering to be consistent with the seed used for the other models. In this section, the elbow method for locating the optimal k is revisited because of an interesting note where the “within clusters sum of square error” is slightly less.

Prior to clustering the data, the optimal number of k-means must be determined. The *elbow* method is a popular determinant for the idealized ‘k’ value in regard to k-means. K-means (the number of groups that the data set is split into evenly, where ‘k’ represents a mathematical constant denoting the number of groups) differs from the number of clusters, which relate to how many centroids are used in the during clustering. In Weka, under the “Classify” tab, the *SimpleKMeans* clustered is selected to find the optimal value of k. *Euclidian Distance* is selected as the underlying algorithm that will be used to calculate the distances between each attribute in the dataset and their nearest centroid. Due to the fact that Euclidian Distance is calculated using the Pythagorean Theorem, which requires taking the square root, only Numeric types are suitable for this algorithm. Since all attributes aside from the class value are Numeric, no additional formatting of the data is required. The seed value for SimpleKMeans is left as 10, as well as the default algorithm, which is Euclidian Distance. An initial value of 4 is selected for k. Optionally, displayStdDevs may be set to true in order to see how much the estimated averages of each attribute differ from the actual average. For the *Cluster Mode*, “Use training set” is selected, with class type and id attribute ignored using Weka’s *Ignore attributes*.

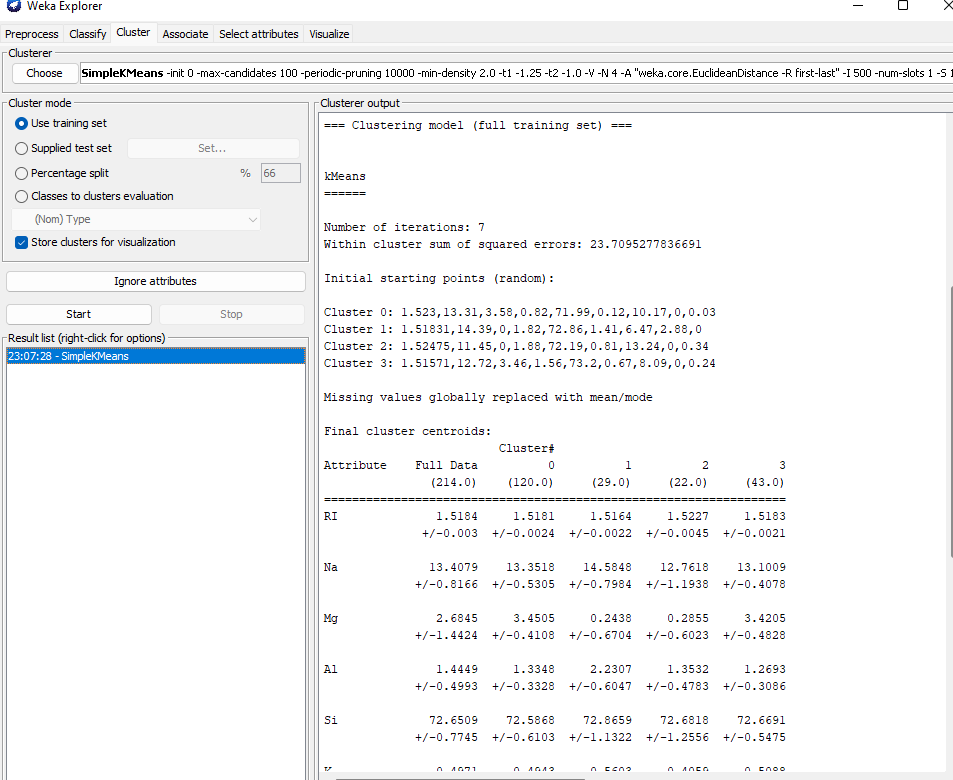


Figure 20 Weka Cluster With Simple K Mean Seed of 1

In order to find the optimal value for k, a particular interest is taken to the value of “Within cluster sum of squared errors” (which from here onwards will be abbreviated as WCSSE). This value represents the sum of Squared Error (SSE) of each attribute in the dataset. SSE is a commonly used algorithm for calculating accuracy within a dataset. It takes the average of the difference of *Absolute Errors* squared, where the Absolute Error (AE) is defined as the estimated value subtracted from the actual value. The formula for this can be seen below:

The value of n represents the number of instances in the attribute column. X sub i represents the position of the value of attribute x for instance i and upper-case X represents the position of the value for the same attribute at the nearest centroid. In the case of SimpleKMeans, taking the Absolute Error is identical to finding the Euclidian distance between each attribute and its nearest centroid.

As the value of k increases, the value of WCSSE decreases. This is due to the fact that average distortion increases since there are fewer instances in the cluster, which means that they will be closer to the centroid. The point at which distortion decreases the most is the point at which the elbow lies. This must be visualized in line graph or scatter plot. For this report, SimpleKMeans was ran a total of 6 times with values of k ranging between 4 and 9, inclusive. The results are shown below:

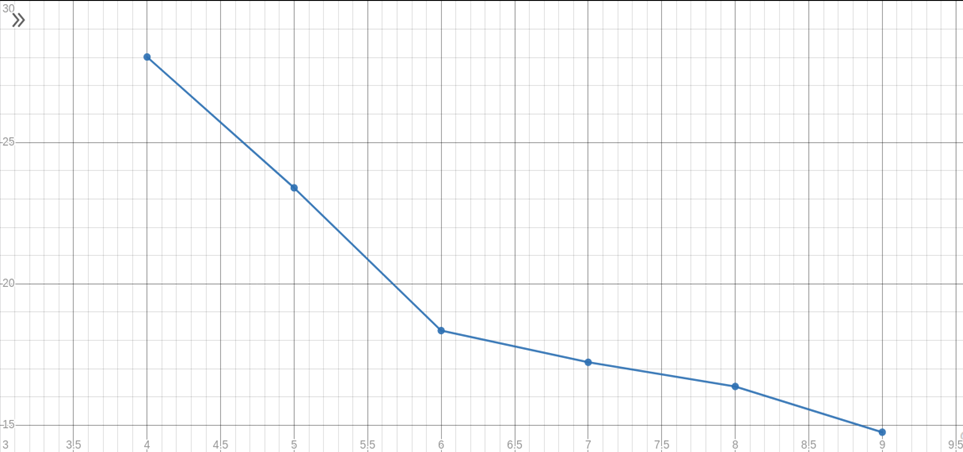


Figure 21 Elbow Chart on Weka With Seed of 10

Table 14 shows slightly better WCSSE when using a seed of 10 rather than 1.

Table 14 SimpleKMeans Table of k Clusters Seed of 10

|  |  |  |  |
| --- | --- | --- | --- |
| Results of SimpleKMeans With Different Values of K | | | |
| Value of K | **Within Cluster Sum of Squared Errors** | Number of Iterations | Incorrectly Clustered Instances |
| 4 | 23.7095 | 7 | 109 (50.93%) |
| 5 | 21.9491 | 9 | 108 (50.47%) |
| 6 | 19.4035 | 9 | 22 (57.01%) |
| 7 | 18.6 | 9 | 127 (59.35%) |
| 8 | 18.5845 | 7 | 128 (59.81%) |
| 9 | 18.1854 | 7 | 128 (59.81%) |

# Conclusion

In spite of the above evaluations, k-mean cluster seems to have outlier clusters when increasing k number of clusters beyond the number of class label. Since class label 6 is considered an outlier there are 5 classes 1,2,3,5 and 7. The elbow point as shown in Figure 5 shows that by using the elbow method the elbow can be observed at the 5th cluster. As shown in Figure 18, 9 clusters in k-mean produce 2 outlier clusters that are 100% outliers with a total of 11 outliers. Farthest first clustering algorithm produces outlier clusters early in the assigning of numbers of k clusters but the count of outliers is few. As k is increased the new cluster may or may not be an outlier cluster. A method is required to identify whether the new clusters is an outlier cluster as well as reviewing previous clusters, although clusters may contain very few instances. In utilizing LOF or ISF, the algorithm successfully computed a result of 80 to 83% of correctly classified instances. However, the false negative rate of both algorithm is roughly 50 to 55% and false positive rate of about 7 to 10% as shown in Figure 7 and Table 12 on LOF classification and Figure 9 and Table 13 on ISF classification. Combining the results of LOF and ISF yielded the best results for identifying outliers shown in Figure 14. The combined algorithm produce 21 instances classified as outliers with 17 instances that are correctly classified and 4 instances incorrectly classified. Table 13 shows that the false positive rate of the ensemble is around 2% and although the false negative rate is higher over 65% it is simply not considered. The ensemble method allows for the highest number of outliers detected with minimum errors while clustering with k-mean produce outlier clusters with less outliers but more accurate. Where as, the farthest first algorithm is much faster and uses less computing time but yields very few outliers in outlier clusters. Farthest first method may require a method to process the new cluster and a review of any previous clusters, except for cluster 0, to identify outlier clusters.

## Division of Work

|  |  |
| --- | --- |
| **Section/Topic** | **Author** |
| Title Page | John |
| Introduction | John |
| Data Understanding | John and Neil |
| Data Preparation | John |
| Modeling | John and Neil |
| Evaluation | John and Neil |
| Discussion of Results | John and Neil |
| Additional Research | Neil |
| Conclusion | John |
| Excel Files, ARFF files, etc. | John and Neil |
| Tables and Figures | John and Neil |

# References

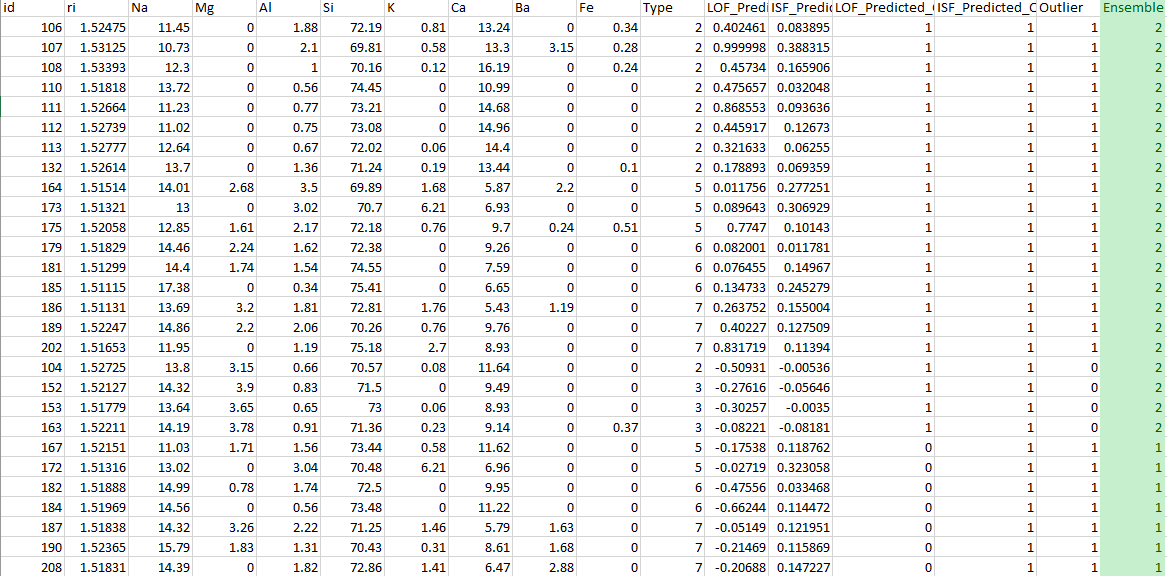
|  |  |
| --- | --- |
| [1] | Center for Machine Learning and Intelligent Systems, "Glass Identification Data Set," UCI Machine Learning Repository, 1 September 1987. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Glass+Identification. [Accessed 11 March 2023]. |
| [2] | L. Hill, "Float Glass Vs. Flat Glass Vs. Plate Glass Vs. Sheet Glass," Glass.com, 2019 August 2019. [Online]. Available: https://info.glass.com/float-flat-plate-sheet-glass/. [Accessed 10 March 2023]. |
| [3] | S. Mishra, "Unsupervised Learning and Data Clustering," Towards Data Science, 19 May 2017. [Online]. Available: https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a. [Accessed 18 March 2023]. |
| [4] | R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald and D. Scuse, "WEKA Manual for Version 3-7-8," The University of Waikato, 21 January 2013. [Online]. Available: https://statweb.stanford.edu/~lpekelis/13\_datafest\_cart/WekaManual-3-7-8.pdf. [Accessed 20 February 2023]. |
| [5] | M. K. Sharmila, "An Optimized Farthest First Clustering Algorithm," The Technological Institute of Textile and Sciences, November 2013. [Online]. Available: https://www.researchgate.net/publication/262979934\_An\_optimized\_farthest\_first\_clustering\_algorithm. [Accessed 18 March 2023]. |
| [6] | E. Frank, "isolation Forest: Class for building and using a classifier built on the Isolation Forest anomaly detection algorithm," Weka Team, 21 January 2016. [Online]. Available: https://weka.sourceforge.io/packageMetaData/isolationForest/index.html. [Accessed 10 February 2023]. |
| [7] | M. Hall, "localOutlierFactor: Filter implementing the Local Outlier Factor anomaly detection algorithm," Weka Team, 8 December 2015. [Online]. Available: https://weka.sourceforge.io/packageMetaData/localOutlierFactor/index.html. [Accessed 10 February 2023]. |
| [8] | akshayvarma, "Local Outlier Factor," GeeksforGeeks, 6 May 2022. [Online]. Available: https://www.geeksforgeeks.org/local-outlier-factor/. [Accessed 18 March 2023]. |
| [9] | A. Krishnamurthy, "Anomaly detection using Isolation Forest – A Complete Guide," Analytics Vidhya, 26 July 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide/. [Accessed 18 March 2023]. |
| [10] | R.Q. Thompson, Oberlin College, "Identification of Glass Samples by Their Refractive Index," R.Q. Thompson, Oberlin College, 10 March 2023. [Online]. Available: https://www.asdlib.org/onlineArticles/elabware/thompson/Glass/Glass(RI)PFaculty.pdf. [Accessed 10 March 2023]. |
| [11] | Addoptics, "What is refractive index?," Addoptics, 21 April 2021. [Online]. Available: https://www.addoptics.com/optics-explained/refractive-index/. [Accessed 18 March 2023]. |
| [12] | K. Enevoldsen, "The Elements — Descriptions, Uses and Occurrences," elements.wlonk.com, 10 March 2023. [Online]. Available: https://elements.wlonk.com/ElementUses.htm. [Accessed 10 March 2023]. |
| [13] | Frontline Solvers, "USING K-MEANS CLUSTERING," FrontLine Solvers, 18 March 2023. [Online]. Available: https://www.solver.com/xlminer/help/using-k-means-clustering. [Accessed 23 March 2023]. |
| [14] | A. Rajbangshi, "Outliers in data and ways to detect them.," Medium, 26 April 2020. [Online]. Available: https://medium.com/analytics-vidhya/outliers-in-data-and-ways-to-detect-them-1c3a5f2c6b1e. [Accessed 18 March 2023]. |
| [15] | K. Ferguson, "When Should You Delete Outliers from a Data Set?," Humans of data, 6 March 2018. [Online]. Available: https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/. [Accessed 10 March 2023]. |
| [16] | C. Shi, B. Wei, S. Wei, W. Wang, H. Liu and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," EURASIP Journal, 21 February 2021. [Online]. Available: https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-021-01910-w. [Accessed 18 March 2023]. |
| [17] | T. Sarkar, "Clustering metrics better than the elbow-method," Towards Data Science, 6 September 2019. [Online]. Available: https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6. [Accessed 18 March 2023]. |
| [18] | The Editors of Encyclopaedia Britannica, "Refractive Index," Britannica, 10 March 2023. [Online]. Available: https://www.britannica.com/science/atmospheric-refraction. [Accessed 10 March 2023]. |
| [19] | Analytical Sciences Digital Library, "Typical Results for Refractive Index Determination of Glass," LibreTexts Chemistry, 9 June 2020. [Online]. Available: https://chem.libretexts.org/Ancillary\_Materials/Laboratory\_Experiments/Wet\_Lab\_Experiments/Analytical\_Chemistry\_Labs/ASDL\_Labware/Forensic\_Science\_Laboratory/Glass/Typical\_Results\_for\_Refractive\_Index\_Determination\_of\_Glass. [Accessed 10 March 2023]. |

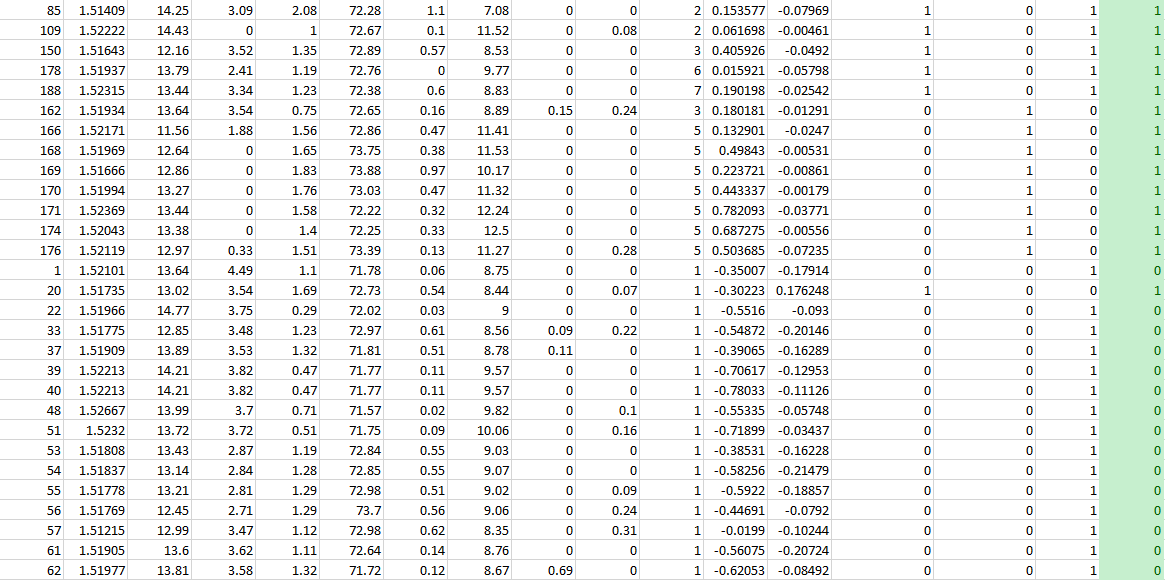
# Appendix 1

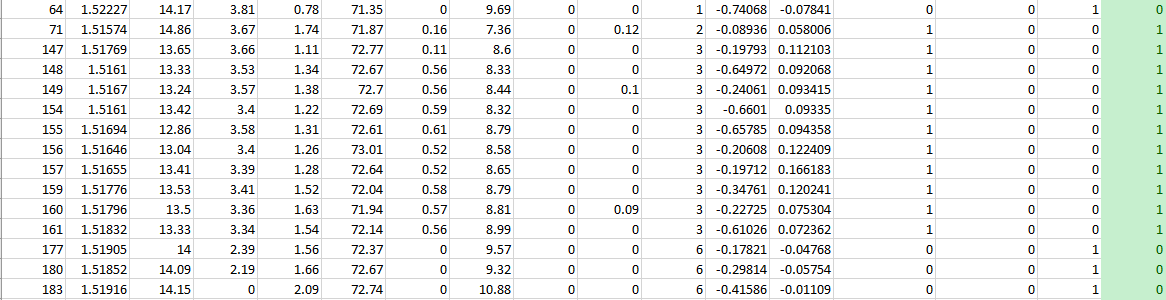
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Attribute** | **Min** | **Max** | **Mean** | **SD** |
| Building  Windows  Float  Processed (1) | RI | 1.51215 | 1.52667 | 1.518718 | 0.002252 |
| Na | 12.45 | 14.77 | 13.24229 | 0.495722 |
| Mg | 2.71 | 4.49 | 3.552429 | 0.245272 |
| Al | 0.29 | 1.69 | 1.163857 | 0.2712 |
| Si | 71.35 | 73.7 | 72.61914 | 0.565402 |
| K | 0 | 0.69 | 0.447429 | 0.213339 |
| Ca | 7.78 | 10.17 | 8.797286 | 0.570686 |
| Ba | 0 | 0.69 | 0.012714 | 0.083237 |
| Fe | 0 | 0.31 | 0.057 | 0.088436 |
| **Type**  Building  Windows Non-Float Processed (2) | **Attribute** | **Min** | **Max** | **Mean** | **SD** |
| RI | 1.51409 | 1.53393 | 1.518619 | 0.003777 |
| Na | 10.73 | 14.86 | 13.11171 | 0.659775 |
| Mg | 0 | 3.98 | 3.002105 | 1.207637 |
| Al | 0.56 | 2.12 | 1.408158 | 0.316239 |
| Si | 69.81 | 74.45 | 72.59803 | 0.71979 |
| K | 0 | 1.1 | 0.521053 | 0.212315 |
| Ca | 7.08 | 16.19 | 9.073684 | 1.908951 |
| Ba | 0 | 3.15 | 0.050263 | 0.359949 |
| Fe | 0 | 0.35 | 0.079737 | 0.10573 |
| **Type**  Vehicle Windows  Float Processed (3) | **Attribute** | **Min** | **Max** | **Mean** | **SD** |
| RI | 1.5161 | 1.52211 | 1.517964 | 0.001859 |
| Na | 12.16 | 14.32 | 13.43706 | 0.491753 |
| Mg | 3.34 | 3.9 | 3.543529 | 0.157926 |
| Al | 0.58 | 1.76 | 1.201176 | 0.337114 |
| Si | 71.36 | 73.01 | 72.40471 | 0.496981 |
| K | 0 | 0.61 | 0.406471 | 0.223026 |
| Ca | 8.32 | 9.65 | 8.782941 | 0.368762 |
| Ba | 0 | 0.15 | 0.008824 | 0.035294 |
| Fe | 0 | 0.37 | 0.057059 | 0.104643 |
| **Type**  Container (5) | **Attribute** | **Min** | **Max** | **Mean** | **SD** |
| RI | 1.51316 | 1.52369 | 1.518928 | 0.003214 |
| Na | 11.03 | 14.01 | 12.82769 | 0.746553 |
| Mg | 0 | 2.68 | 0.773846 | 0.959948 |
| Al | 1.4 | 3.5 | 2.033846 | 0.666697 |
| Si | 69.89 | 73.88 | 72.36615 | 1.232012 |
| K | 0.13 | 6.21 | 1.47 | 2.054792 |
| Ca | 5.87 | 12.5 | 10.12385 | 2.098118 |
| Ba | 0 | 2.2 | 0.187692 | 0.584389 |
| Fe | 0 | 0.51 | 0.060769 | 0.149484 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type**  Tableware (6) | **Attribute** | **Min** | **Max** | **Mean** | **SD** |
| RI | 1.51115 | 1.51969 | 1.517456 | 0.002938 |
| Na | 13.79 | 17.38 | 14.64667 | 1.022024 |
| Mg | 0 | 2.41 | 1.305556 | 1.034388 |
| Al | 0.34 | 2.09 | 1.366667 | 0.539156 |
| Si | 72.37 | 75.41 | 73.20667 | 1.017732 |
| K | 0 | 0 | 0 | 0 |
| Ca | 6.65 | 11.22 | 9.356667 | 1.367024 |
| Ba | 0 | 0 | 0 | 0 |
| Fe | 0 | 0 | 0 | 0 |
| **Type**  Headlamps (7) | **Attribute** | **Min** | **Max** | **Mean** | **SD** |
| RI | 1.51131 | 1.52365 | 1.517116 | 0.002501 |
| Na | 11.95 | 15.79 | 14.44207 | 0.674421 |
| Mg | 0 | 3.34 | 0.538276 | 1.098243 |
| Al | 1.19 | 2.88 | 2.122759 | 0.435026 |
| Si | 70.26 | 75.18 | 72.96586 | 0.923881 |
| K | 0 | 2.7 | 0.325172 | 0.656866 |
| Ca | 5.43 | 9.76 | 8.491379 | 0.956573 |
| Ba | 0 | 2.88 | 1.04 | 0.653769 |
| Fe | 0 | 0.09 | 0.013448 | 0.029276 |

# Appendix 2







# Appendix 3

## LOF Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | Type | LOF Prediction Margin | LOF Predicted Outlier | Outlier |
| 40 | 1 | -0.78033 | no | yes |
| 64 | 1 | -0.74068 | no | yes |
| 51 | 1 | -0.71899 | no | yes |
| 39 | 1 | -0.70617 | no | yes |
| 184 | 6 | -0.66244 | no | yes |
| 154 | 3 | -0.6601 | yes | no |
| 155 | 3 | -0.65785 | yes | no |
| 148 | 3 | -0.64972 | yes | no |
| 62 | 1 | -0.62053 | no | yes |
| 161 | 3 | -0.61026 | yes | no |
| 55 | 1 | -0.5922 | no | yes |
| 54 | 1 | -0.58256 | no | yes |
| 61 | 1 | -0.56075 | no | yes |
| 48 | 1 | -0.55335 | no | yes |
| 22 | 1 | -0.5516 | no | yes |
| 33 | 1 | -0.54872 | no | yes |
| 104 | 2 | -0.50931 | yes | no |
| 182 | 6 | -0.47556 | no | yes |
| 56 | 1 | -0.44691 | no | yes |
| 183 | 6 | -0.41586 | no | yes |
| 37 | 1 | -0.39065 | no | yes |
| 53 | 1 | -0.38531 | no | yes |
| 1 | 1 | -0.35007 | no | yes |
| 159 | 3 | -0.34761 | yes | no |
| 153 | 3 | -0.30257 | yes | no |
| 20 | 1 | -0.30223 | yes | no |
| 180 | 6 | -0.29814 | no | yes |
| 152 | 3 | -0.27616 | yes | no |
| 149 | 3 | -0.24061 | yes | no |
| 160 | 3 | -0.22725 | yes | no |
| 190 | 7 | -0.21469 | no | yes |
| 208 | 7 | -0.20688 | no | yes |
| 156 | 3 | -0.20608 | yes | no |
| 147 | 3 | -0.19793 | yes | no |
| 157 | 3 | -0.19712 | yes | no |
| 177 | 6 | -0.17821 | no | yes |
| 167 | 5 | -0.17538 | no | yes |
| 71 | 2 | -0.08936 | yes | no |
| 163 | 3 | -0.08221 | yes | no |
| 187 | 7 | -0.05149 | no | yes |
| 172 | 5 | -0.02719 | no | yes |
| 57 | 1 | -0.0199 | no | yes |
| 131 | 2 | 0.001439 | no | no |
| 164 | 5 | 0.011756 | yes | yes |
| 178 | 6 | 0.015921 | yes | yes |
| 151 | 3 | 0.02181 | no | no |
| 109 | 2 | 0.061698 | yes | yes |
| 158 | 3 | 0.072605 | no | no |
| 181 | 6 | 0.076455 | yes | yes |
| 179 | 6 | 0.082001 | yes | yes |
| 173 | 5 | 0.089643 | yes | yes |
| 105 | 2 | 0.115138 | no | no |
| 102 | 2 | 0.122296 | no | no |
| 166 | 5 | 0.132901 | no | no |
| 185 | 6 | 0.134733 | yes | yes |
| 85 | 2 | 0.153577 | yes | yes |
| 146 | 2 | 0.160076 | no | no |
| 45 | 1 | 0.165484 | no | no |
| 130 | 2 | 0.178655 | no | no |
| 132 | 2 | 0.178893 | yes | yes |
| 162 | 3 | 0.180181 | no | no |
| 81 | 2 | 0.186355 | no | no |
| 188 | 7 | 0.190198 | yes | yes |
| 169 | 5 | 0.223721 | no | no |
| 186 | 7 | 0.263752 | yes | yes |
| 72 | 2 | 0.269574 | no | no |
| 11 | 1 | 0.306367 | no | no |
| 113 | 2 | 0.321633 | yes | yes |
| 165 | 5 | 0.325384 | no | no |
| 90 | 2 | 0.35846 | no | no |
| 59 | 1 | 0.365822 | no | no |
| 6 | 1 | 0.37665 | no | no |
| 10 | 1 | 0.383642 | no | no |
| 99 | 2 | 0.398704 | no | no |
| 189 | 7 | 0.40227 | yes | yes |
| 106 | 2 | 0.402461 | yes | yes |
| 150 | 3 | 0.405926 | yes | yes |
| 170 | 5 | 0.443337 | no | no |
| 112 | 2 | 0.445917 | yes | yes |
| 13 | 1 | 0.451084 | no | no |
| 136 | 2 | 0.455151 | no | no |
| 108 | 2 | 0.45734 | yes | yes |
| 110 | 2 | 0.475657 | yes | yes |
| 14 | 1 | 0.48427 | no | no |
| 88 | 2 | 0.488008 | no | no |
| 3 | 1 | 0.492226 | no | no |
| 168 | 5 | 0.49843 | no | no |
| 176 | 5 | 0.503685 | no | no |
| 129 | 2 | 0.513294 | no | no |
| 9 | 1 | 0.523095 | no | no |
| 60 | 1 | 0.523615 | no | no |
| 128 | 2 | 0.546435 | no | no |
| 79 | 2 | 0.564301 | no | no |
| 46 | 1 | 0.564961 | no | no |
| 31 | 1 | 0.578981 | no | no |
| 50 | 1 | 0.593904 | no | no |
| 34 | 1 | 0.596036 | no | no |
| 58 | 1 | 0.628101 | no | no |
| 21 | 1 | 0.638504 | no | no |
| 145 | 2 | 0.641205 | no | no |
| 119 | 2 | 0.644397 | no | no |
| 2 | 1 | 0.655925 | no | no |
| 91 | 2 | 0.67774 | no | no |
| 98 | 2 | 0.683443 | no | no |
| 174 | 5 | 0.687275 | no | no |
| 36 | 1 | 0.689275 | no | no |
| 103 | 2 | 0.698478 | no | no |
| 66 | 1 | 0.702711 | no | no |
| 18 | 1 | 0.705161 | no | no |
| 143 | 2 | 0.707301 | no | no |
| 47 | 1 | 0.719573 | no | no |
| 19 | 1 | 0.736017 | no | no |
| 80 | 2 | 0.738694 | no | no |
| 191 | 7 | 0.739326 | no | no |
| 68 | 1 | 0.746922 | no | no |
| 4 | 1 | 0.749868 | no | no |
| 134 | 2 | 0.750745 | no | no |
| 69 | 1 | 0.751506 | no | no |
| 122 | 2 | 0.758008 | no | no |
| 100 | 2 | 0.763267 | no | no |
| 139 | 2 | 0.773201 | no | no |
| 49 | 1 | 0.773574 | no | no |
| 175 | 5 | 0.7747 | yes | yes |
| 70 | 1 | 0.781442 | no | no |
| 171 | 5 | 0.782093 | no | no |
| 137 | 2 | 0.78324 | no | no |
| 63 | 1 | 0.784602 | no | no |
| 125 | 2 | 0.787044 | no | no |
| 93 | 2 | 0.796435 | no | no |
| 101 | 2 | 0.800218 | no | no |
| 142 | 2 | 0.805988 | no | no |
| 52 | 1 | 0.806643 | no | no |
| 67 | 1 | 0.810828 | no | no |
| 84 | 2 | 0.810895 | no | no |
| 44 | 1 | 0.814245 | no | no |
| 77 | 2 | 0.817712 | no | no |
| 126 | 2 | 0.823448 | no | no |
| 202 | 7 | 0.831719 | yes | yes |
| 8 | 1 | 0.836107 | no | no |
| 97 | 2 | 0.839535 | no | no |
| 138 | 2 | 0.850172 | no | no |
| 140 | 2 | 0.851124 | no | no |
| 114 | 2 | 0.855677 | no | no |
| 96 | 2 | 0.855845 | no | no |
| 118 | 2 | 0.861305 | no | no |
| 123 | 2 | 0.863028 | no | no |
| 65 | 1 | 0.863989 | no | no |
| 111 | 2 | 0.868553 | yes | yes |
| 43 | 1 | 0.878313 | no | no |
| 212 | 7 | 0.880629 | no | no |
| 117 | 2 | 0.883534 | no | no |
| 194 | 7 | 0.889737 | no | no |
| 210 | 7 | 0.894389 | no | no |
| 5 | 1 | 0.897959 | no | no |
| 205 | 7 | 0.899982 | no | no |
| 198 | 7 | 0.902599 | no | no |
| 124 | 2 | 0.904422 | no | no |
| 7 | 1 | 0.905272 | no | no |
| 133 | 2 | 0.909739 | no | no |
| 135 | 2 | 0.920855 | no | no |
| 213 | 7 | 0.924474 | no | no |
| 201 | 7 | 0.926784 | no | no |
| 115 | 2 | 0.927402 | no | no |
| 195 | 7 | 0.927661 | no | no |
| 199 | 7 | 0.930406 | no | no |
| 204 | 7 | 0.931882 | no | no |
| 209 | 7 | 0.932303 | no | no |
| 141 | 2 | 0.933636 | no | no |
| 42 | 1 | 0.934024 | no | no |
| 214 | 7 | 0.93466 | no | no |
| 193 | 7 | 0.934821 | no | no |
| 144 | 2 | 0.939027 | no | no |
| 203 | 7 | 0.941083 | no | no |
| 207 | 7 | 0.947606 | no | no |
| 83 | 2 | 0.950886 | no | no |
| 211 | 7 | 0.951242 | no | no |
| 27 | 1 | 0.953535 | no | no |
| 127 | 2 | 0.95551 | no | no |
| 206 | 7 | 0.955519 | no | no |
| 15 | 1 | 0.958558 | no | no |
| 197 | 7 | 0.960299 | no | no |
| 196 | 7 | 0.960334 | no | no |
| 200 | 7 | 0.961506 | no | no |
| 192 | 7 | 0.963038 | no | no |
| 95 | 2 | 0.963804 | no | no |
| 41 | 1 | 0.968625 | no | no |
| 73 | 2 | 0.970653 | no | no |
| 25 | 1 | 0.973474 | no | no |
| 120 | 2 | 0.975104 | no | no |
| 76 | 2 | 0.976641 | no | no |
| 121 | 2 | 0.978502 | no | no |
| 12 | 1 | 0.982992 | no | no |
| 17 | 1 | 0.984118 | no | no |
| 74 | 2 | 0.98808 | no | no |
| 116 | 2 | 0.989422 | no | no |
| 75 | 2 | 0.990896 | no | no |
| 87 | 2 | 0.991123 | no | no |
| 78 | 2 | 0.993294 | no | no |
| 94 | 2 | 0.993596 | no | no |
| 86 | 2 | 0.99503 | no | no |
| 82 | 2 | 0.995436 | no | no |
| 89 | 2 | 0.997594 | no | no |
| 32 | 1 | 0.998095 | no | no |
| 29 | 1 | 0.99903 | no | no |
| 26 | 1 | 0.999305 | no | no |
| 30 | 1 | 0.999378 | no | no |
| 28 | 1 | 0.999507 | no | no |
| 16 | 1 | 0.999964 | no | no |
| 107 | 2 | 0.999998 | yes | yes |
| 23 | 1 | 0.999999 | no | no |
| 24 | 1 | 0.999999 | no | no |
| 35 | 1 | 0.999999 | no | no |
| 38 | 1 | 0.999999 | no | no |
| 92 | 2 | 0.999999 | no | no |

## ISF Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type | id | ISF Prediction Margin | ISF Predicted Outlier | Outlier |
| 1 | 54 | -0.21479 | no | yes |
| 1 | 61 | -0.20724 | no | yes |
| 1 | 33 | -0.20146 | no | yes |
| 1 | 55 | -0.18857 | no | yes |
| 1 | 1 | -0.17914 | no | yes |
| 1 | 37 | -0.16289 | no | yes |
| 1 | 53 | -0.16228 | no | yes |
| 1 | 39 | -0.12953 | no | yes |
| 1 | 40 | -0.11126 | no | yes |
| 1 | 57 | -0.10244 | no | yes |
| 1 | 22 | -0.093 | no | yes |
| 1 | 62 | -0.08492 | no | yes |
| 3 | 163 | -0.08181 | yes | no |
| 2 | 85 | -0.07969 | no | yes |
| 1 | 56 | -0.0792 | no | yes |
| 1 | 64 | -0.07841 | no | yes |
| 5 | 176 | -0.07235 | yes | no |
| 6 | 178 | -0.05798 | no | yes |
| 6 | 180 | -0.05754 | no | yes |
| 1 | 48 | -0.05748 | no | yes |
| 3 | 152 | -0.05646 | yes | no |
| 3 | 150 | -0.0492 | no | yes |
| 6 | 177 | -0.04768 | no | yes |
| 5 | 171 | -0.03771 | yes | no |
| 1 | 51 | -0.03437 | no | yes |
| 7 | 188 | -0.02542 | no | yes |
| 5 | 166 | -0.0247 | yes | no |
| 3 | 162 | -0.01291 | yes | no |
| 6 | 183 | -0.01109 | no | yes |
| 5 | 169 | -0.00861 | yes | no |
| 5 | 174 | -0.00556 | yes | no |
| 2 | 104 | -0.00536 | yes | no |
| 5 | 168 | -0.00531 | yes | no |
| 2 | 109 | -0.00461 | no | yes |
| 3 | 153 | -0.0035 | yes | no |
| 5 | 170 | -0.00179 | yes | no |
| 7 | 194 | 0.004577 | no | no |
| 6 | 179 | 0.011781 | yes | yes |
| 7 | 195 | 0.028832 | no | no |
| 3 | 158 | 0.032028 | no | no |
| 2 | 110 | 0.032048 | yes | yes |
| 6 | 182 | 0.033468 | yes | yes |
| 7 | 210 | 0.038068 | no | no |
| 7 | 196 | 0.039836 | no | no |
| 7 | 192 | 0.040038 | no | no |
| 7 | 212 | 0.040721 | no | no |
| 5 | 165 | 0.050642 | no | no |
| 7 | 193 | 0.055394 | no | no |
| 2 | 105 | 0.057286 | no | no |
| 2 | 71 | 0.058006 | no | no |
| 1 | 11 | 0.058155 | no | no |
| 3 | 151 | 0.061569 | no | no |
| 2 | 113 | 0.06255 | yes | yes |
| 7 | 203 | 0.062562 | no | no |
| 2 | 132 | 0.069359 | yes | yes |
| 1 | 18 | 0.070685 | no | no |
| 2 | 72 | 0.07159 | no | no |
| 3 | 161 | 0.072362 | no | no |
| 7 | 201 | 0.07478 | no | no |
| 3 | 160 | 0.075304 | no | no |
| 7 | 191 | 0.076335 | no | no |
| 7 | 209 | 0.076825 | no | no |
| 7 | 207 | 0.077796 | no | no |
| 7 | 211 | 0.082245 | no | no |
| 2 | 106 | 0.083895 | yes | yes |
| 2 | 130 | 0.085563 | no | no |
| 7 | 206 | 0.089066 | no | no |
| 2 | 131 | 0.089651 | no | no |
| 7 | 197 | 0.090617 | no | no |
| 3 | 148 | 0.092068 | no | no |
| 3 | 154 | 0.09335 | no | no |
| 3 | 149 | 0.093415 | no | no |
| 2 | 111 | 0.093636 | yes | yes |
| 3 | 155 | 0.094358 | no | no |
| 7 | 214 | 0.094708 | no | no |
| 2 | 129 | 0.095502 | no | no |
| 2 | 146 | 0.09661 | no | no |
| 7 | 199 | 0.100308 | no | no |
| 7 | 204 | 0.100331 | no | no |
| 5 | 175 | 0.10143 | yes | yes |
| 7 | 213 | 0.104442 | no | no |
| 7 | 198 | 0.107018 | no | no |
| 7 | 205 | 0.109627 | no | no |
| 3 | 147 | 0.112103 | no | no |
| 7 | 202 | 0.11394 | yes | yes |
| 6 | 184 | 0.114472 | yes | yes |
| 7 | 200 | 0.115101 | no | no |
| 7 | 190 | 0.115869 | yes | yes |
| 5 | 167 | 0.118762 | yes | yes |
| 3 | 159 | 0.120241 | no | no |
| 2 | 103 | 0.120853 | no | no |
| 7 | 187 | 0.121951 | yes | yes |
| 3 | 156 | 0.122409 | no | no |
| 1 | 45 | 0.122476 | no | no |
| 1 | 13 | 0.122827 | no | no |
| 1 | 63 | 0.122917 | no | no |
| 2 | 112 | 0.12673 | yes | yes |
| 1 | 49 | 0.127191 | no | no |
| 7 | 189 | 0.127509 | yes | yes |
| 1 | 70 | 0.12836 | no | no |
| 1 | 69 | 0.132709 | no | no |
| 1 | 6 | 0.134452 | no | no |
| 2 | 136 | 0.136577 | no | no |
| 2 | 98 | 0.143803 | no | no |
| 2 | 101 | 0.144382 | no | no |
| 7 | 208 | 0.147227 | yes | yes |
| 1 | 67 | 0.14872 | no | no |
| 6 | 181 | 0.14967 | yes | yes |
| 1 | 44 | 0.152715 | no | no |
| 1 | 68 | 0.153523 | no | no |
| 7 | 186 | 0.155004 | yes | yes |
| 1 | 2 | 0.158322 | no | no |
| 1 | 66 | 0.15886 | no | no |
| 2 | 79 | 0.160591 | no | no |
| 1 | 65 | 0.163249 | no | no |
| 2 | 145 | 0.163617 | no | no |
| 2 | 93 | 0.165802 | no | no |
| 2 | 108 | 0.165906 | yes | yes |
| 3 | 157 | 0.166183 | no | no |
| 2 | 102 | 0.170908 | no | no |
| 2 | 143 | 0.174084 | no | no |
| 1 | 20 | 0.176248 | no | no |
| 2 | 128 | 0.177113 | no | no |
| 2 | 137 | 0.178899 | no | no |
| 2 | 134 | 0.181594 | no | no |
| 1 | 47 | 0.182604 | no | no |
| 2 | 142 | 0.184353 | no | no |
| 1 | 19 | 0.185355 | no | no |
| 2 | 122 | 0.186715 | no | no |
| 2 | 100 | 0.190899 | no | no |
| 1 | 3 | 0.191681 | no | no |
| 2 | 90 | 0.194013 | no | no |
| 2 | 97 | 0.194527 | no | no |
| 2 | 81 | 0.195088 | no | no |
| 1 | 31 | 0.195196 | no | no |
| 2 | 118 | 0.195267 | no | no |
| 2 | 91 | 0.196765 | no | no |
| 1 | 50 | 0.197334 | no | no |
| 2 | 99 | 0.19919 | no | no |
| 1 | 9 | 0.200194 | no | no |
| 1 | 10 | 0.201038 | no | no |
| 2 | 125 | 0.206458 | no | no |
| 1 | 8 | 0.206723 | no | no |
| 1 | 34 | 0.211256 | no | no |
| 2 | 80 | 0.211258 | no | no |
| 2 | 126 | 0.212348 | no | no |
| 1 | 14 | 0.212876 | no | no |
| 2 | 119 | 0.213322 | no | no |
| 1 | 17 | 0.215538 | no | no |
| 1 | 46 | 0.216882 | no | no |
| 1 | 4 | 0.217068 | no | no |
| 1 | 52 | 0.218223 | no | no |
| 1 | 21 | 0.218796 | no | no |
| 1 | 5 | 0.218966 | no | no |
| 2 | 114 | 0.221394 | no | no |
| 1 | 59 | 0.222114 | no | no |
| 2 | 141 | 0.222251 | no | no |
| 1 | 36 | 0.222472 | no | no |
| 2 | 139 | 0.222825 | no | no |
| 2 | 144 | 0.223739 | no | no |
| 2 | 87 | 0.225663 | no | no |
| 2 | 95 | 0.228037 | no | no |
| 2 | 115 | 0.228321 | no | no |
| 2 | 73 | 0.230706 | no | no |
| 1 | 60 | 0.230886 | no | no |
| 1 | 42 | 0.232724 | no | no |
| 2 | 94 | 0.232856 | no | no |
| 2 | 84 | 0.233521 | no | no |
| 2 | 133 | 0.23391 | no | no |
| 2 | 75 | 0.235344 | no | no |
| 2 | 92 | 0.23652 | no | no |
| 1 | 7 | 0.236854 | no | no |
| 1 | 12 | 0.237505 | no | no |
| 2 | 117 | 0.239799 | no | no |
| 2 | 82 | 0.240579 | no | no |
| 2 | 116 | 0.241117 | no | no |
| 2 | 127 | 0.242241 | no | no |
| 1 | 26 | 0.242245 | no | no |
| 2 | 76 | 0.242848 | no | no |
| 1 | 43 | 0.243032 | no | no |
| 1 | 27 | 0.244657 | no | no |
| 6 | 185 | 0.245279 | yes | yes |
| 2 | 88 | 0.246379 | no | no |
| 2 | 77 | 0.24818 | no | no |
| 1 | 28 | 0.248397 | no | no |
| 1 | 58 | 0.248528 | no | no |
| 2 | 135 | 0.249761 | no | no |
| 1 | 15 | 0.250048 | no | no |
| 1 | 29 | 0.250347 | no | no |
| 1 | 32 | 0.250507 | no | no |
| 2 | 96 | 0.252758 | no | no |
| 2 | 74 | 0.253229 | no | no |
| 1 | 25 | 0.253701 | no | no |
| 2 | 124 | 0.257708 | no | no |
| 1 | 38 | 0.258075 | no | no |
| 1 | 35 | 0.259644 | no | no |
| 2 | 140 | 0.260086 | no | no |
| 1 | 23 | 0.260386 | no | no |
| 1 | 41 | 0.260502 | no | no |
| 2 | 83 | 0.261873 | no | no |
| 2 | 138 | 0.262924 | no | no |
| 2 | 121 | 0.264123 | no | no |
| 1 | 16 | 0.26593 | no | no |
| 1 | 30 | 0.26714 | no | no |
| 2 | 123 | 0.270073 | no | no |
| 2 | 86 | 0.27586 | no | no |
| 5 | 164 | 0.277251 | yes | yes |
| 2 | 78 | 0.279947 | no | no |
| 2 | 120 | 0.279952 | no | no |
| 2 | 89 | 0.28141 | no | no |
| 1 | 24 | 0.281872 | no | no |
| 5 | 173 | 0.306929 | yes | yes |
| 5 | 172 | 0.323058 | yes | yes |
| 2 | 107 | 0.388315 | yes | yes |

# Appendix 4

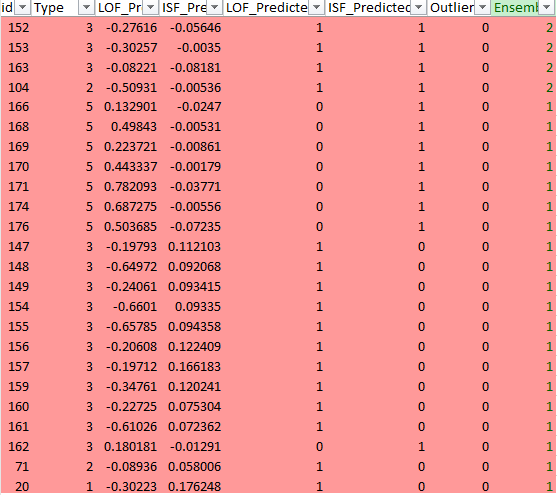


Figure 22 Ensemble - Incorrectly Classified by Either LOF or ISF

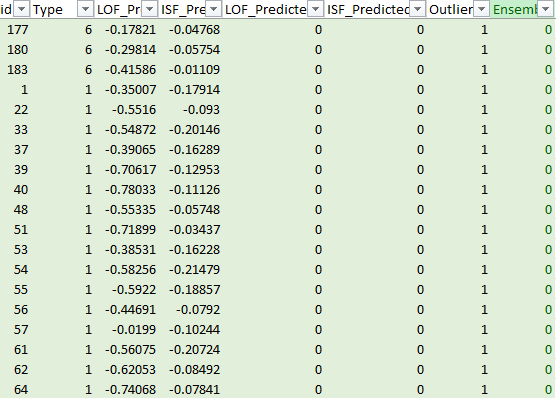


Figure 23 Ensemble - Outliers Not Identified by LOF or ISF