PRML MINOR PROJECT

Nishtha Karki (B21AI051)
Drithi Davuluri (B21AI055)
Uddanti Moksha Akshaya (B21AI041)

PREPROCESSING:

- Loaded the “Healthcare Dataset and analyzed the features.
- Out of the total 27 features, we studied the significance of each feature in finding anomalies in the dataset .
- **Dropped the following columns :**

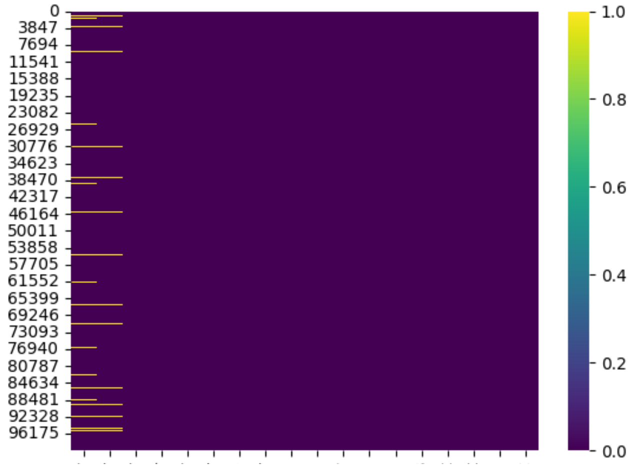
1. **'index', 'National Provider Identifier'** - unique to every datapoint in the dataset and hence, has no significance in finding anomalies.

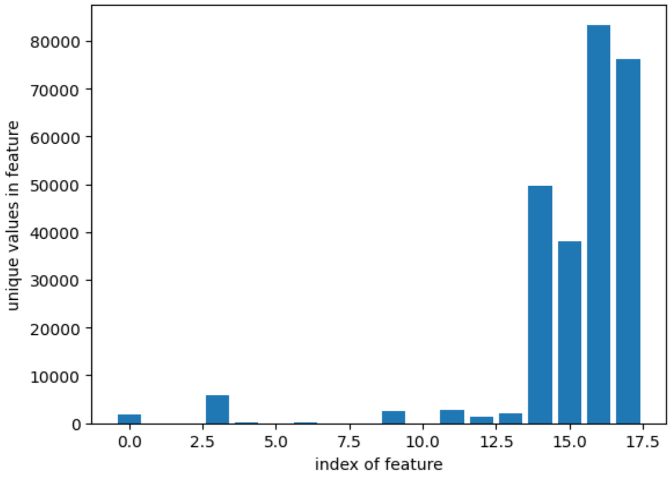
2. 'Last Name/Organization Name of the Provider', 'First Name of the Provider', 'Middle Initial of the Provider'- Name does not contribute to analyzing whether a datapoint is anomalous or not. Furthermore, for organizations, we don't have any middle name or last names which can be tough to encode.

3.**'Street Address 1 of the Provider','Street Address 2 of the Provider', 'City of the Provider'-**Feature too specific to a datapoint to predict anomaly.

4.**HCPCS Code-**Redundant information if we have HCPCS description.

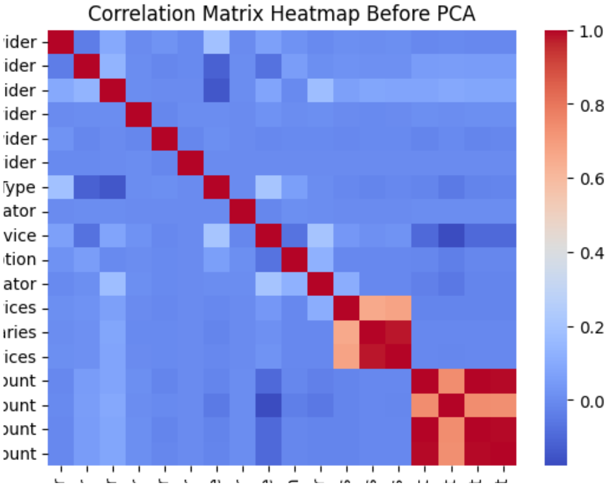
- Found features having Null values and plotted heatmap for it.

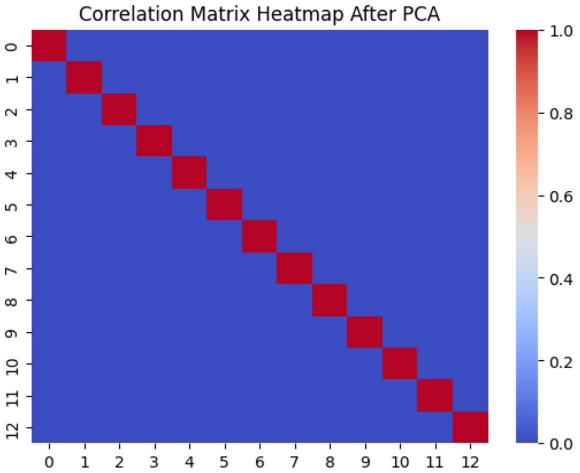




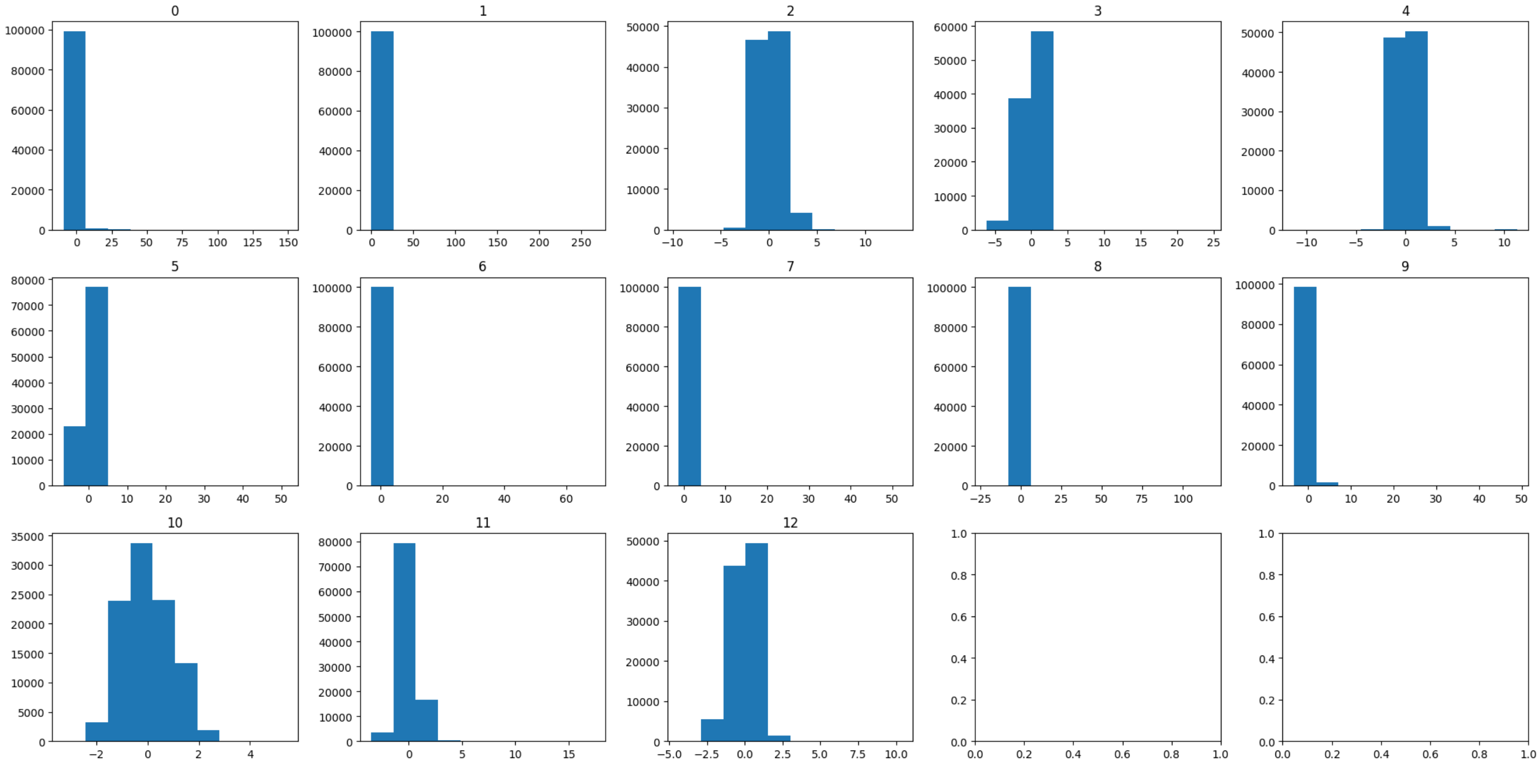
- **Replaced the missing values** with mode because Mode imputation is suitable for categorical variables or numerical variables with a small number of unique values.
- Here , in the first two features , as we can observe from the heatmap that there is a very small yellow part as compared to the purple part indicating to the fact that the percentage of null values is pretty low.

- Also , the first two features are of object data type and are hence, categorical.
- From the following graph , we can observe that the unique values in the first two features are low.
- Removed the ‘,’ from data points . for eg in 1,000.
- Converted the data type of features having numbers in them from object to float.
- Encoded the rest of the features having data type as object
- Applied PCA and preserved 95% of the variance of the dataset reducing the number of features from 18 to 13.





- Correlation plots are used to understand which variables are related to each other and the strength of this relationship. Since the principal components are orthogonal to each other hence, they are uncorrelated i.e. each one captures a unique aspect of the variation in the data.
- Plotted hist plots of each principal component in the dataset.



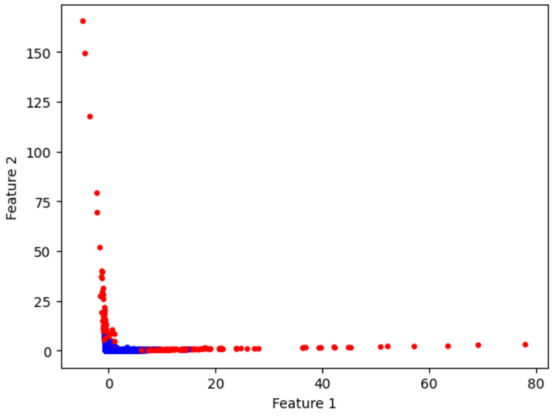
ANOMALY DETECTION:

1) K-NEAREST NEIGHBORS FOR ANOMALY DETECTION :

K-nearest neighbors (KNN) is a popular algorithm used for anomaly detection. The basic idea behind KNN anomaly detection is to use the distance of a data point to its k nearest neighbors to determine whether it is an outlier or not.

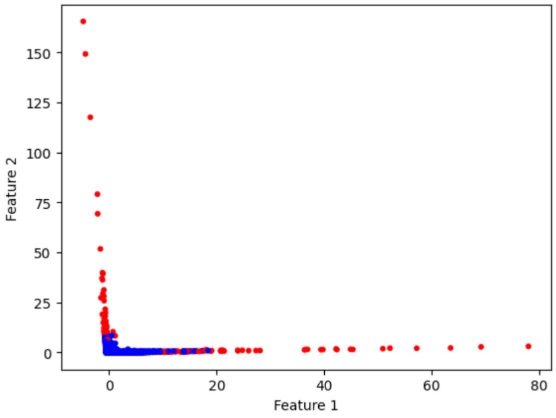
- For choosing the number of nearest neighbors there is a rule of thumb to take k as the square root of the number of samples or the number of features. As the square root of the number of samples is quite large we considered taking number of features. We took k from 5 to 13(i.e. The number of features).
- The distances to the k nearest neighbors are computed for each point in the data. A threshold is set for anomaly detection based on the mean distance to the kth neighbor plus two times the standard deviation of the distances to the kth neighbor. Anomalies are flagged based on whether their distance to the kth neighbor is greater than the threshold set.

k=5



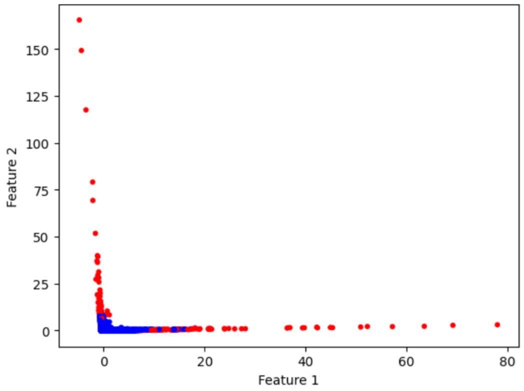
```
Number of anomalies detected: 458
```

k=7

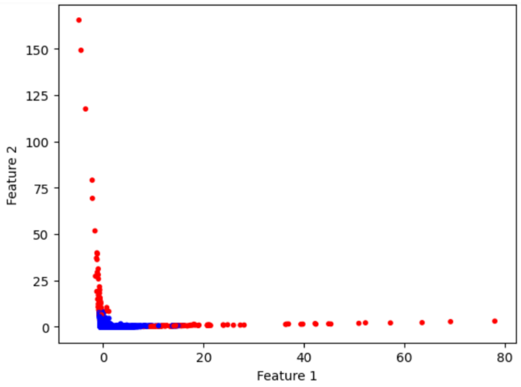


```
Number of anomalies detected: 144
```

k=9



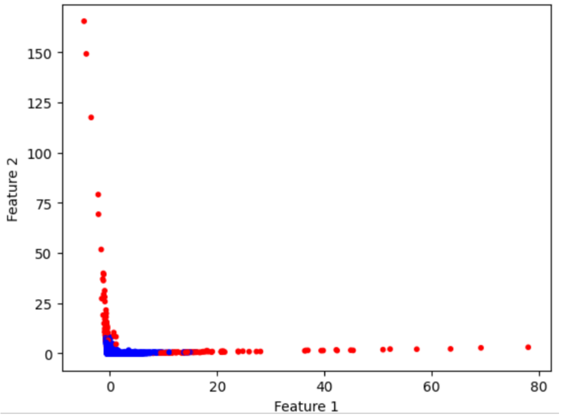
k=11



```
Number of anomalies detected: 166
```

```
Number of anomalies detected: 179
```

k=13



```
Number of anomalies detected: 189
```

- The results showed that k=5 produced the best results. The reason for this could be that a smaller value of k (e.g., k=5) captures more local outliers that may be missed by larger values of k. A larger value of k result in a smoother decision boundary but may also miss some local outliers.
- **Drawbacks of using KNN**
  - KNN is computationally expensive, especially when the size of the dataset is large and the number of features is high. This resulted in slow execution times and was not feasible.
  - The choice of k can have a significant impact on the performance of the algorithm.
  - If the data is highly imbalanced, with very few anomalies compared to normal points, KNN may not be the best choice. This is because the algorithm may classify most of the points as normal, resulting in a high false negative rate.

2) SVM FOR ANOMALY DETECTION:

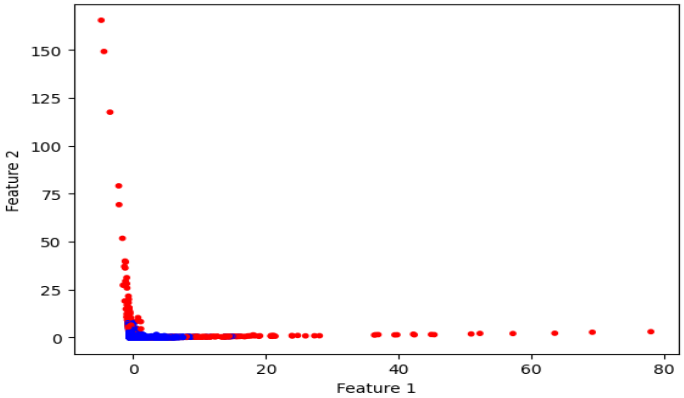
Support Vector Machines (SVMs) can also be used for anomaly detection.The basic idea behind SVM-based anomaly detection is to define a decision boundary that separates the normal data points from the anomalous ones.

- One common approach is the one-class SVM, which involves training an SVM on only the normal data points, without any labeled anomalous data. The SVM learns a boundary that encloses the normal data points in a high-dimensional space, and any data points that fall outside this boundary are flagged as anomalies.

Here we are considering the One-class SVM model for detection of anomalies.

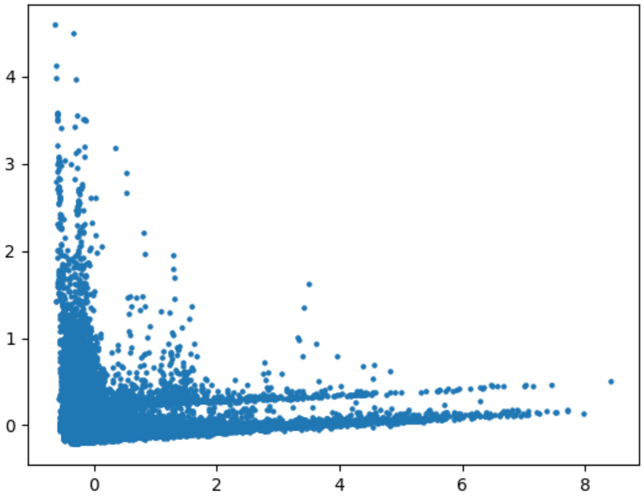
- We are initially considering that dataset to have around 0.5% outliers.
- Anomalies are identified as data points with a predicted label of -1, while normal data points are identified as those with a predicted label of 1.

- OneClassSVM(nu=0.005, kernel='rbf', gamma=0.1)

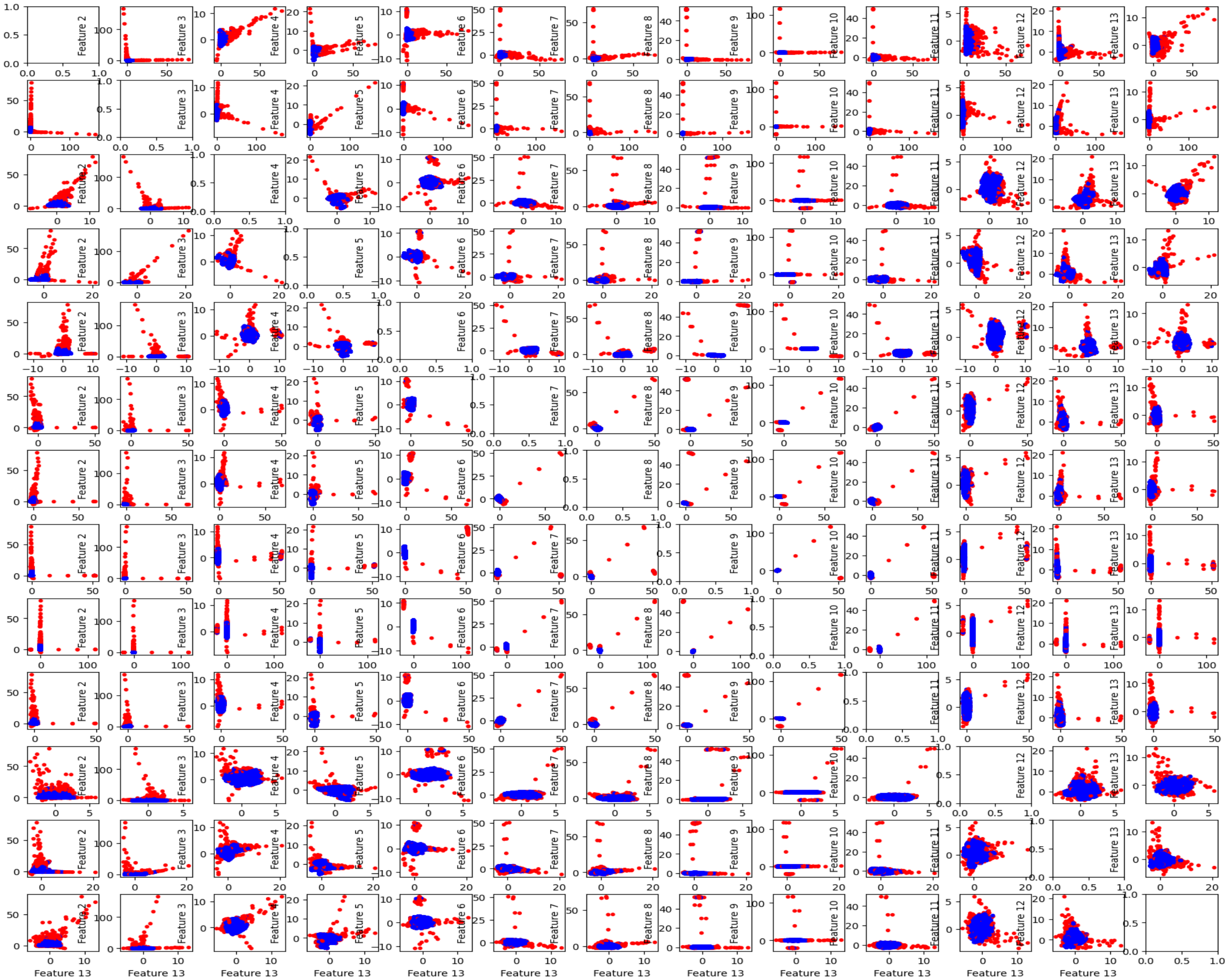


Number of anomalies detected: 508

- After applying one-class SVM we are removing the anomalies datapoints and plotting the remaining points



These are graphs for all 13 features using svm-



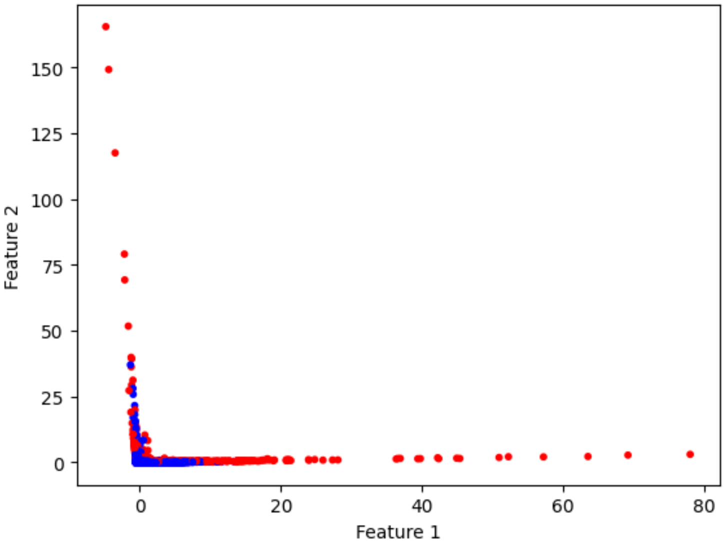
From the above graphs we can see that one-class SVM works good for detection of anomalies when compared to k-Nearest Neighbors

- SVMs are particularly well-suited for anomaly detection because they are effective in handling high-dimensional feature spaces, which is often the case in real-world datasets.
- SVMs also have the ability to handle non-linear data distributions, thanks to the use of kernel functions. In addition, SVMs are robust to overfitting, meaning that they are less likely to incorrectly classify novel or unseen data points.
- In the context of anomaly detection, SVMs are trained on a dataset consisting of both normal and anomalous data points. The SVM then identifies a decision boundary that separates the normal data points from the anomalous ones

3) ISOLATION FOREST FOR ANOMALY DETECTION

The IsolationForest algorithm is a popular machine learning technique for detecting anomalies and outliers in datasets. It works by constructing a number of decision trees and isolating the outliers that are more likely to be isolated in the trees. The algorithm returns a score for each data point, with lower scores indicating a higher likelihood of being an outlier.

- We are initially considering that dataset to have around 0.5% outliers.
- We are considering 100 decision trees for detection of outliers
- The parameter ‘contamination’ is set to 0.005 considering the results that we received in knn and svm where the percentage of anomalous points came out to be 0.005%.
- Also, we varied the value and for 0.005.which we obtained the best result.
-



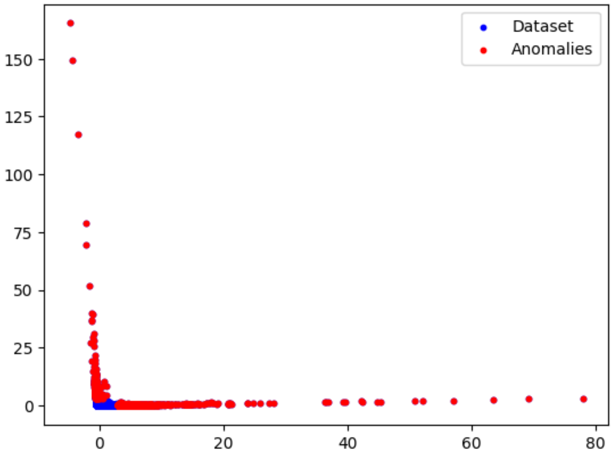
- **Drawbacks of using Isolation forest**

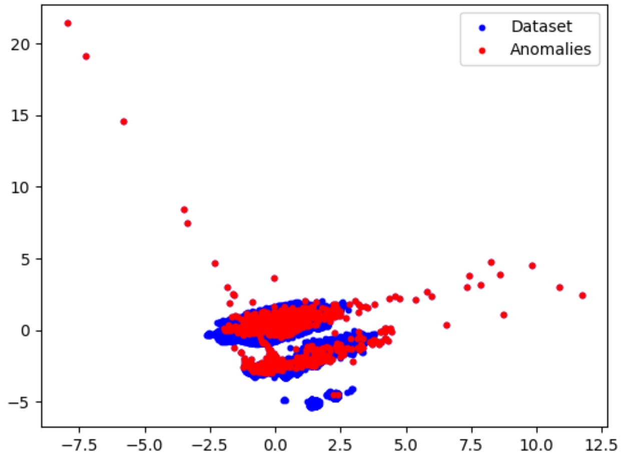
- Difficulty with high-dimensional data: Isolation Forest can struggle with datasets that have a large number of features or dimensions. As the number of dimensions increases, the algorithm may become less effective at isolating anomalies.
- Since here we are considering 13 features which has large dimension. So it is not able to perform good on this dataset when compared to One-Class SVM
- Bias towards outliers of smaller size: The algorithm may have a bias towards outliers of smaller size due to the nature of the isolation process. Outliers that are more isolated in the decision trees are more likely to be detected as anomalies, but this may not always be desirable.

4) FINDING ANOMALIES USING Z-SCORE-

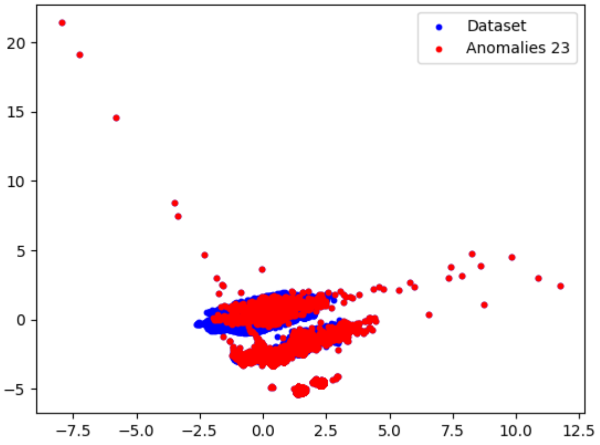
In statistics, the empirical rule states that **99.7%** of data occurs within three standard deviations of the mean within a normal distribution.

- Here we normalize the dataset. The thought process here is to label the data points that are not between -3 and +3 as anomalous.
- Firstly , we took the first two features, found the indexes of data points that were anomalous according to each of the features and took their union to get the indexes of the anomalous dataset.

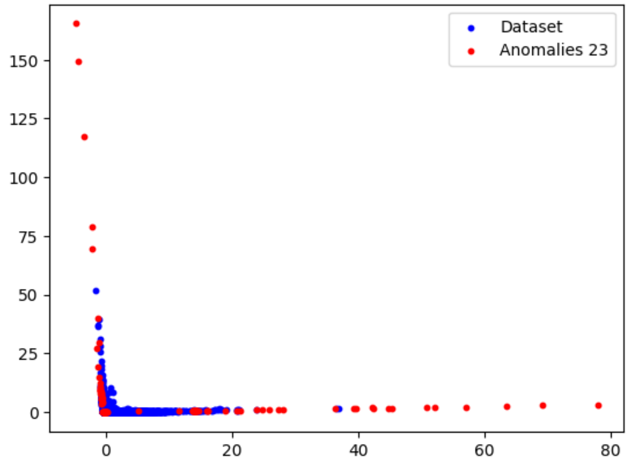


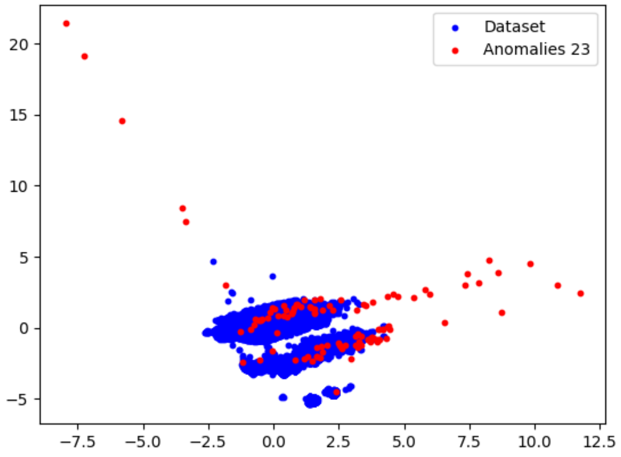


- We observed here that this method is giving us good results in plot of first two features but very inaccurate results for third and fourth feature. Hence this method is rejected.
- Then we proceeded by finding the array of index of anonymous datapoints for each feature and took union of all of them to get 4000 anomalous datapoints.

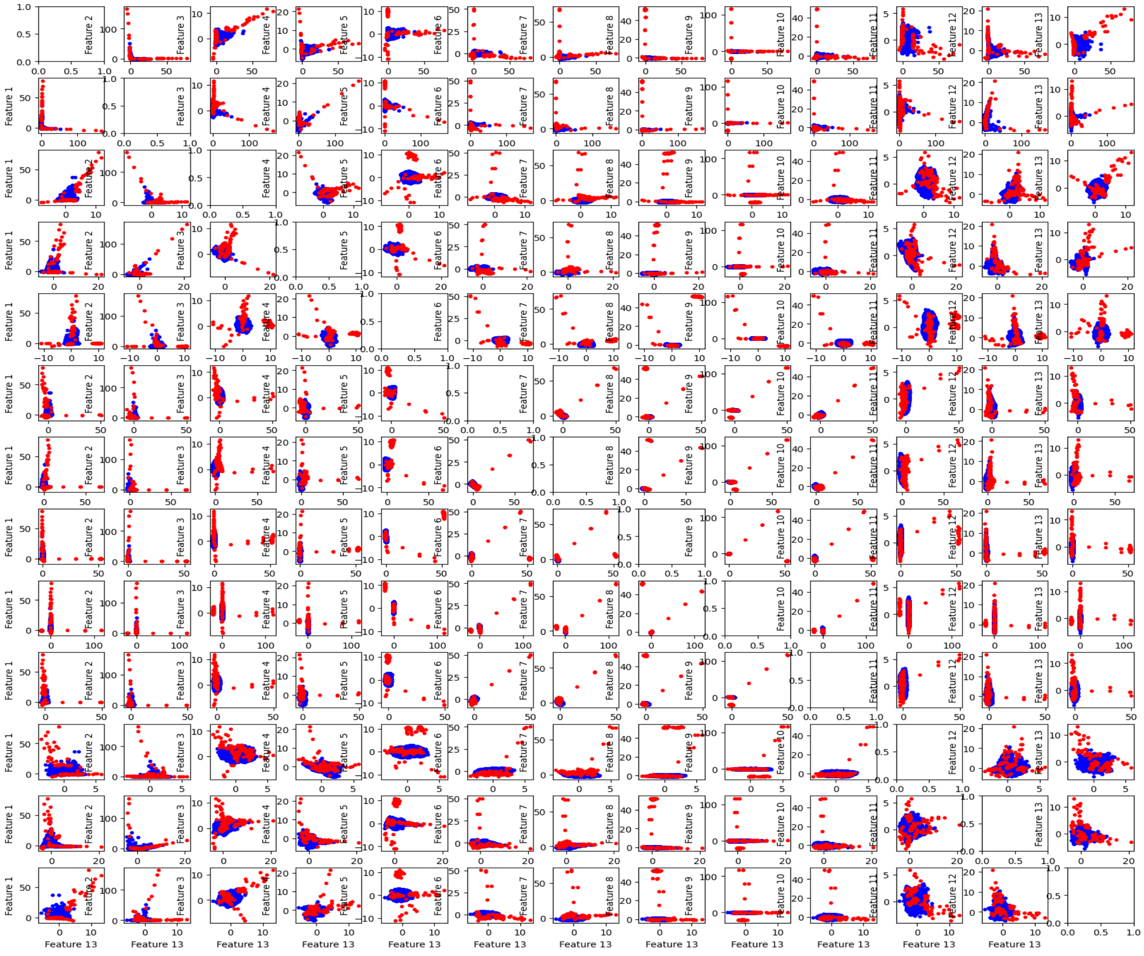


- Here, as we can see that we are getting completely chaotic results, hence this approach is also rejected.
- We concat all the anomalous points arrays of each feature and then we counted the frequencies of all points and got 109 points as anomalous .
- Then the points which were occurring more then twice were classified as anomalous points and again a new plot is plotted on the basis of that.





- Plot for all 13 features with every other feature.



- Here, we can observe that we are getting very inaccurate results and hence we reject this method altogether.

Summary

- We have trained the data on 4 different models - KNN , ISOLATION FOREST , SVM , DETECTION BY Z SCORE
- We cannot decide that a particular model is best for all datasets, but we can find the best performing models according to our dataset. So for our dataset comparing all the models we found that SVM is performing better than other models.
- SVMs are known to have better generalization performance than KNN and Isolation Forest. This means that SVMs are able to perform well on new, unseen data, while KNN and Isolation Forest may not perform as well.
- SVMs are known to be more robust to outliers than KNN and Isolation Forest. This is because SVMs try to find the best decision boundary that separates the data, while KNN and Isolation Forest rely on the nearest neighbors to make decisions.