Introduction

This research was conducted to find a better algorithm than the Extended Isolation Forest (EIF) for anomaly detection. EIF performed on four different data sets showed high accuracy in detecting the intentionally introduced anomalies. However, the best results on each data set were acquired using a different hyperparameter (threshold). The purpose of this research is to find a more generalizable algorithm whose hyperparameter that results in the highest accuracy does not change depending on the provided data set.

Methods

In the comparison of different algorithms, silhouette scores of the models were considered and used. The 3D plots were also observed since they are a good visual indicator of the results. The silhouette coefficient is a metric that is used to assess the performance of a clustering method. It gives a brief picture of how well each data point has been categorized. The silhouette value gauges an object's cohesiveness with its own cluster in comparison to other clusters. The silhouette coefficient has a range of 1 to +1 and, a high number on the silhouette implies that the object is well matched to its own cluster and poorly matched to nearby clusters.

For this study, K-Means Clustering, Local Outlier Factor (LOF), and DBSCAN algorithms were considered in the beginning. However, silhouette scores and 3D plots of the data and found anomalies show that K-Means and LOF were not working well in finding anomalies. The silhouette scores were around 0.30 for K-Means and LOF models which is very low. They also failed to detect most of the known anomalies. For this reason, DBSCAN and Extended Isolation Forest (EIF) were compared.

DBSCAN (Density-based spatial clustering of applications with noise), is a commonly used unsupervised clustering algorithm. It works by grouping the data points that are close to each other based on a distance metric (usually Euclidean) together. It also detects outliers that do not belong in any of the clusters, which is useful in our case. DBSCAN has two important hyperparameters; “minPoints” and “eps”. minPoints is the minimum number of points required to form a cluster. “eps” determines the required distance between points in order to belong to a cluster. There are some conventions used for determining these two hyperparameters. For “minPoints”, if the dimension (D) of the data is small, D+1 should be chosen, if the dimension is large, the value should be 2*D. For the optimum “eps” value, a k-distance graph is created to see the plot of sorted distances from each point to its closest neighbor. The distance value where the angle of the plot changes sharply (elbow point) should be chosen as “eps”.

Implementation

Since the clustering is done using a distance metric on the data points, and features with different ranges can affect the results, standardization was applied to the data using Standard Scaler before training the model. According to the dimensions of the data, 16 is chosen as “minPoints” and according to the elbow method, 0.5 is chosen as “eps” for the implementation of DBSCAN. DBSCAN automatically labels anomalies as -1. To simplify the results, only the first detected anomaly in one minute is kept as the final anomaly.

Results

The silhouette score for different data sets is 0.74 on average which means the clustering was done successfully. The percentage of correctly detected anomalies ranges between %70 and %100 for the different data sets.

Data set	**Accuracy**	**Silhouette Score**
Anomaly 1	80%	0,76
Anomaly 2	70%	0,68
Anomaly 3	89%	0,75
Anomaly 4	100%	0,75

Table1. Accuracy and sillhouette scores on data sets (DBSCAN)

The scores after using EIF are shown in Table 2. The sihouette scores in Table 2 are obtained by using a threshold of 0.67.

Data set	Accuracy	Silhouette Score
Anomaly 1	63%	0,62
Anomaly 2	61%	0,59
Anomaly 3	51%	0,60
Anomaly 4	52%	0,54

Table2. Accuracy and sillhouette scores on data sets (EIF)

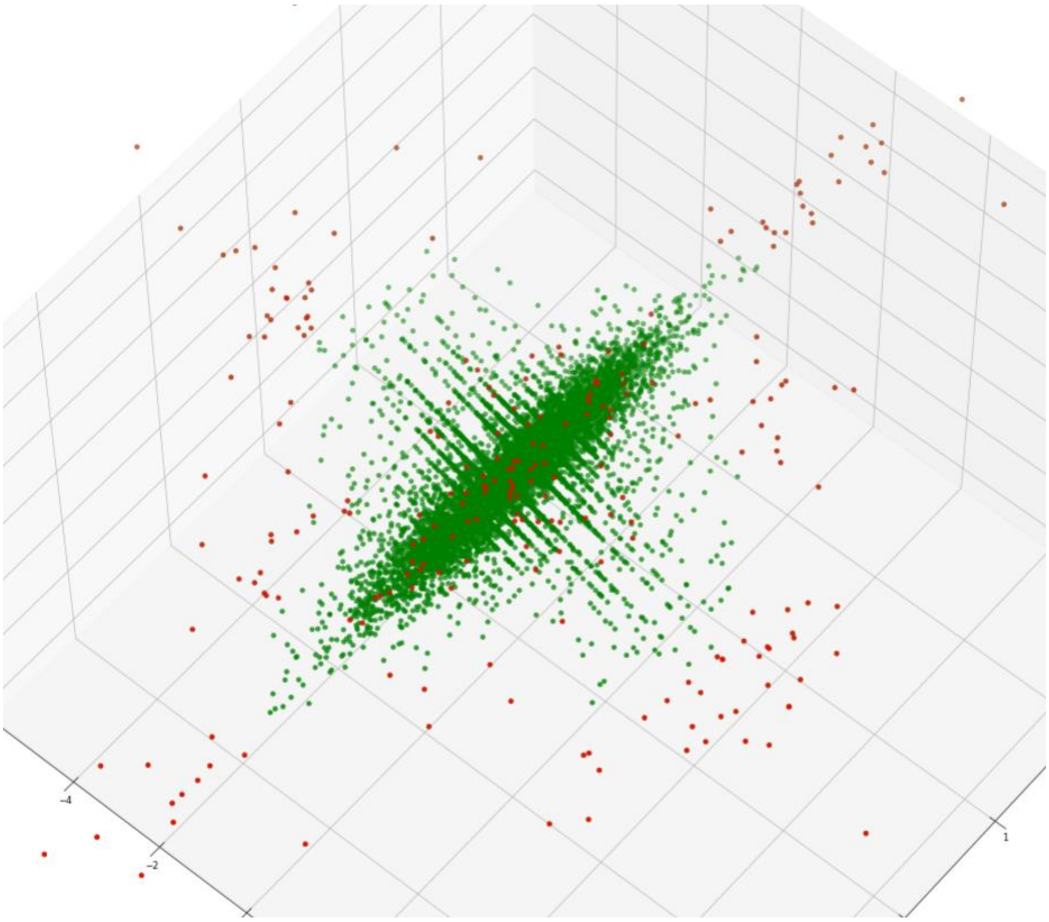

Figure 1. 3D Plot of Anomaly 3 data set. Red points show the found anomalies by DBSCAN.

Discussion & Conclusion

In Table 1 and Table 2, it can be seen that DBSCAN was more successful in detecting the known anomalies. However, the total number of found anomalies is also higher than EIF which results in a high number of false positives. “eps” value can be increased to decrease the total number of anomalies. However, that results in a decrease in accuracy. For example, for the Anomaly 1 data set, when the eps value is 0.5, 56 anomalies are found and, when the eps value is 0.75, the number of anomalies found is 20 but the accuracy drops from %80 to %60. This situation requires a decision to be made in terms of a trade-off between the number of false positives and accuracy.