CSCI4150U: Data Mining

Anomaly Detection

Syed Naqvi
Student ID: 100590852

November 24, 2024

Abstract

This report analyzes the effectiveness of Anomaly Detection using Parametric Models, Distance-based models and Density-based Models. When the dataset is roughly normally distributed, Mahalanobis distance provides an effective way to identify anomalies. In the case of more abstract distributions, anomalies can be detected using largest distance to k^{th} nearest neighbor, or minimal density.

1 Introduction

1.1 Methodology

First we use Mahalanobis distance to identify outliers on the *G-data* dataset as this dataset follows a roughly normal distribution. We then use a distance based approach to assign points an anomaly score using the maximal distance from their k^{th} nearest neighbors for $k = [1, 2, 5]$. The points with the highest anomaly scores have the highest likelihood of being outliers. Finally, we use the inverse of each of the previous distances and then the inverse of the average of all k nearest distances for each point to assign density scores. The points with the lowest density score have the highest chance of being anomalies.

1.2 Preprocessing

To initialize our dataset for clustering, we perform feature standardization and remove the existing labels column giving the following initial plots:

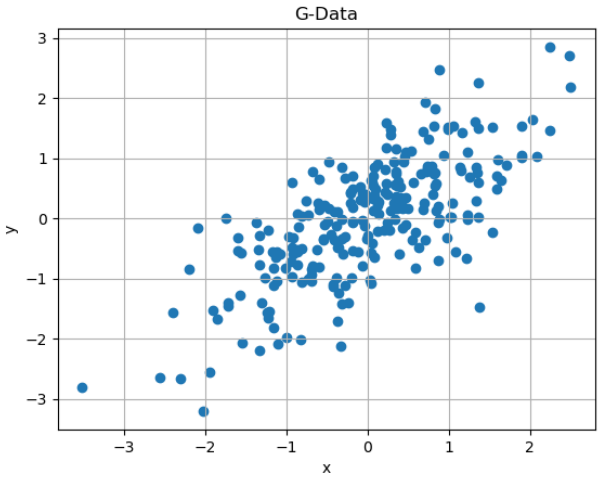

Figure 1: Compound Data

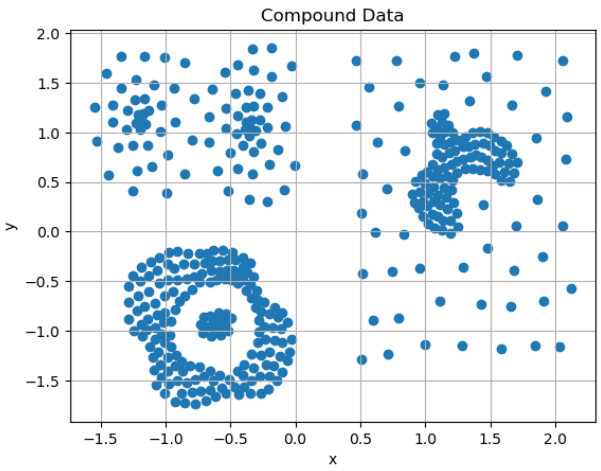

Figure 2: Flame Data

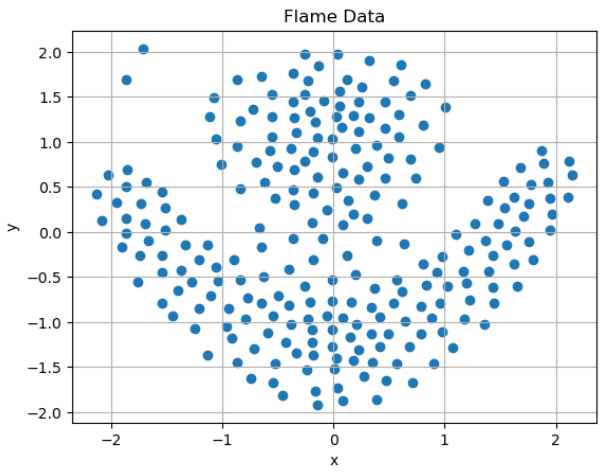

Figure 3: Pathbased Data

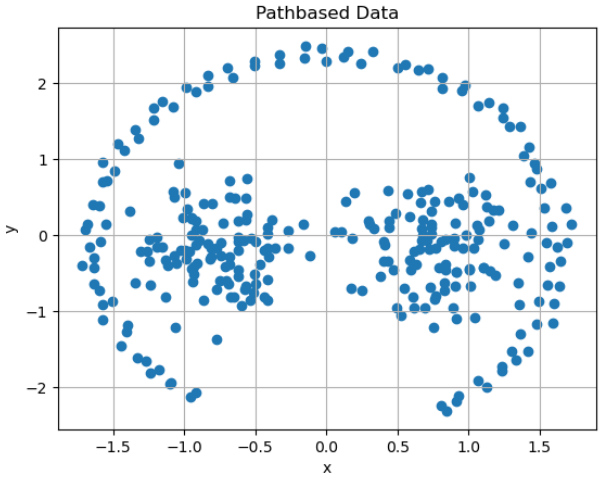

Figure 4: Spiral Data

2 Part I (Using Parametric Models):

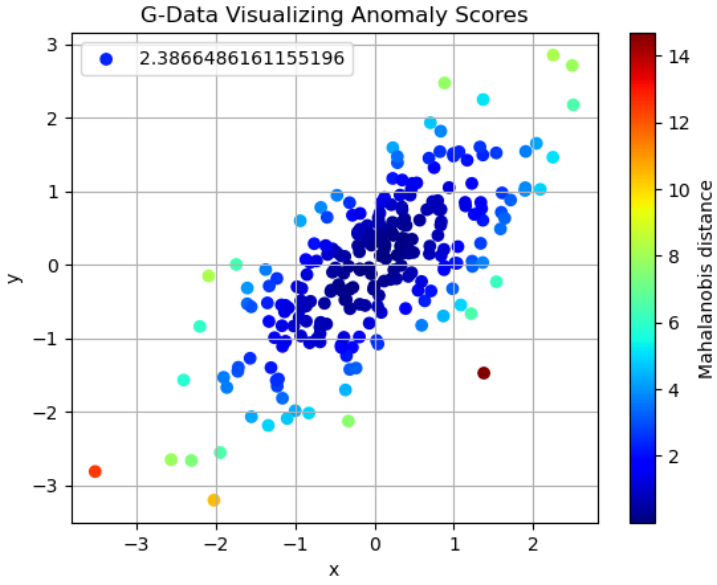

Figure 5: G-Data Anomaly Visualization

	x	y	anomaly_score
41	1.376915	-1.474245	14.678287
8	-3.514537	-2.813186	12.453138
26	-2.019632	-3.202631	10.403718
67	-2.086532	-0.153715	8.193830
204	2.249512	2.849704	8.160705

Figure 6: G-Data top 5 Anomalies

As expected, we see that points along the distribution are penalized less than points off-distribution. This is evident in some points being a darker red color (higher anomaly score) despite being closer to the central cluster than other points in a euclidean sense, but actually being farther off from the main distribution of the dataset.

3 Part II (Using Distance-Based Models):

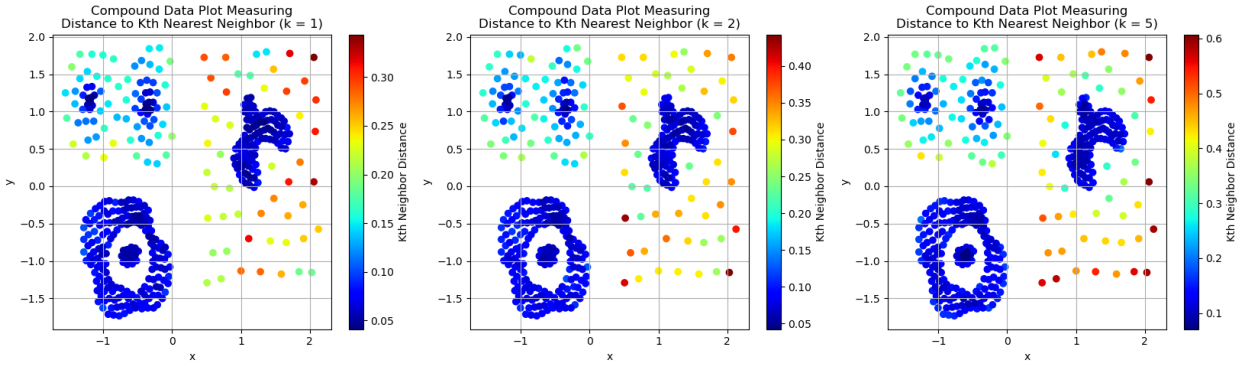

Figure 7: Compound Data Anomaly Visualizations

	x	y	distance_score
42	2.060247	1.726481	0.607148
28	2.060247	0.058959	0.604712
36	2.127088	-0.574277	0.588630
37	2.029397	-1.154744	0.588630
20	0.713140	-1.239175	0.578571

Figure 8: Anomalies with Highest Kth Distance (k = 5)

Using distance from the 5th nearest neighbor as an anomaly score seems to identify the most intuitive points, farthest from the main, most compact clusters of the dataset.

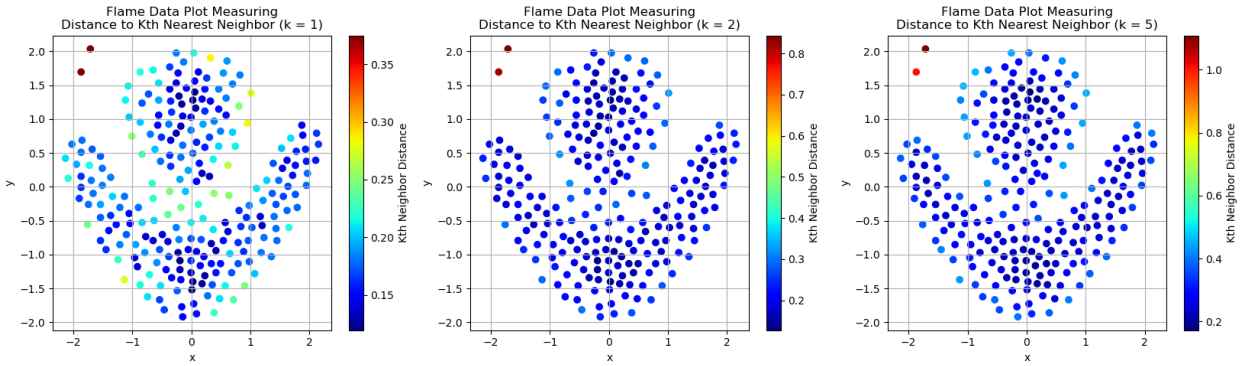

Figure 9: Flame Data Anomaly Visualizations

	x	y	distance_score
0	-1.712779	2.035183	1.108408
1	-1.869233	1.694577	1.007131
197	-0.867928	1.694577	0.460143
196	-1.071318	1.487252	0.459344
204	1.009519	1.383589	0.448655

Figure 10: Flame Data Top 5 Anomalies (k = 5)

Using distance from the 5th nearest neighbor as an anomaly score seems to identify the most intuitive points, farthest from the main, most compact clusters of the dataset.

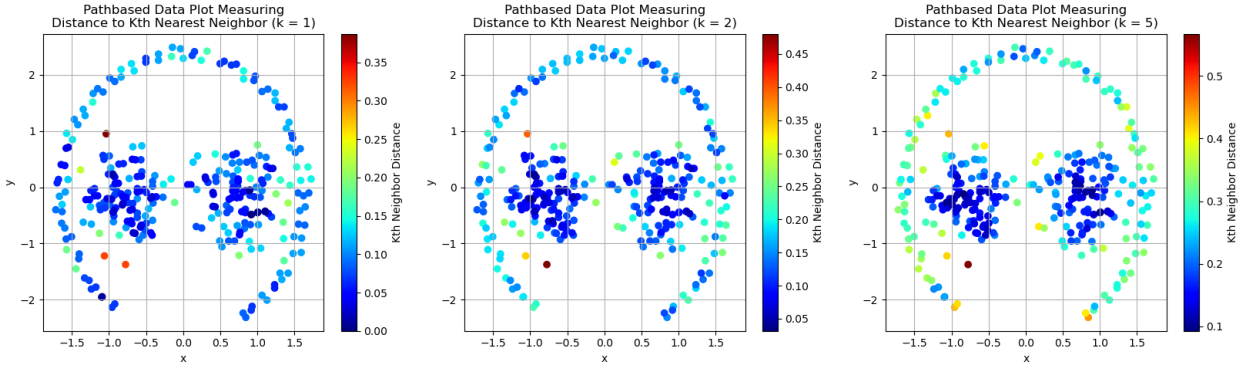

Figure 11: Pathbased Data Anomaly Visualizations

	x	y	distance_score
270	-1.037263	0.951303	0.387203
202	-0.776324	-1.374998	0.324108
203	-1.061537	-1.221052	0.324108
266	-1.383159	0.309860	0.225910
107	1.414351	-0.280268	0.208852

Figure 12: Pathbased Data Top 5 Anomalies (k = 1)

In the case of the pathbased dataset, the distance to the first nearest neighbor seems to do the best job identifying anomalies. This anomaly score was able to remain low for all the points falling within the two central clusters as well as the bordering ring shape while identifying points that deviated most from these two distributions.

4 Part III (Using Density-Based Models):

4.1 A. Density as the Inverse of the kth Nearest Neighbor

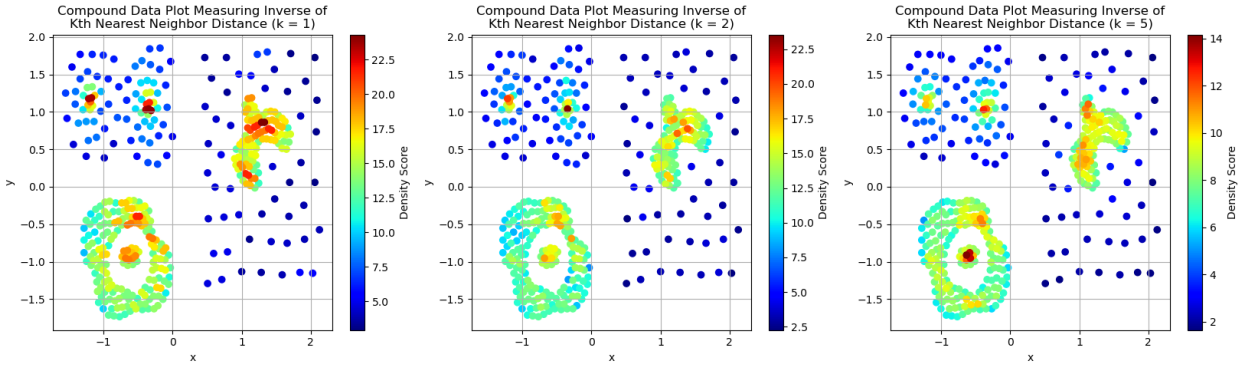

Figure 13: Compound Data Anomaly Visualizations

	x	y	density_score
42	2.060247	1.726481	1.647045
28	2.060247	0.058959	1.653679
37	2.029397	-1.154744	1.698861
36	2.127088	-0.574277	1.698861
20	0.713140	-1.239175	1.728396

Figure 14: Compound Data Top 5 Anomalies (k = 5)

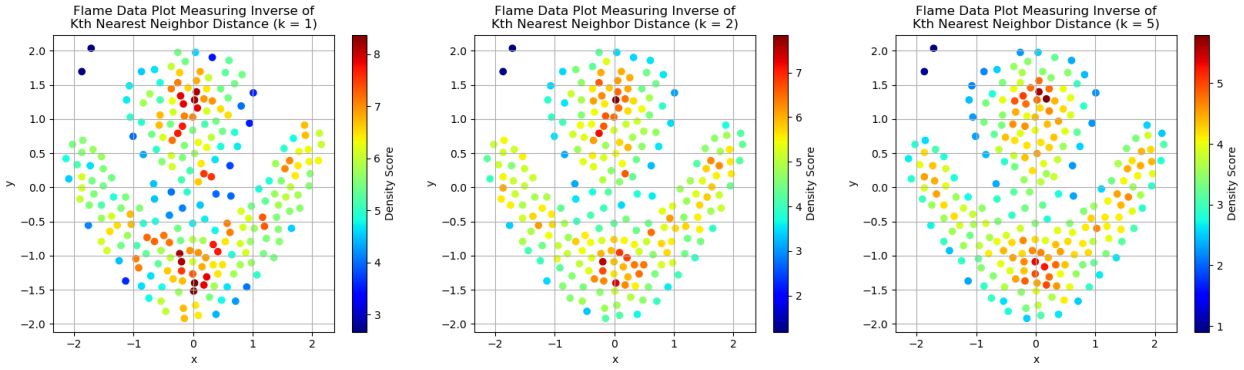

Figure 15: Flame Data Anomaly Score

	x	y	density_score
0	-1.712779	2.035183	0.902195
1	-1.869233	1.694577	0.992920
197	-0.867928	1.694577	2.173238
196	-1.071318	1.487252	2.177016
192	0.946937	0.939320	2.228885

Figure 16: Flame Data Top 5 Anomalies (k = 5)

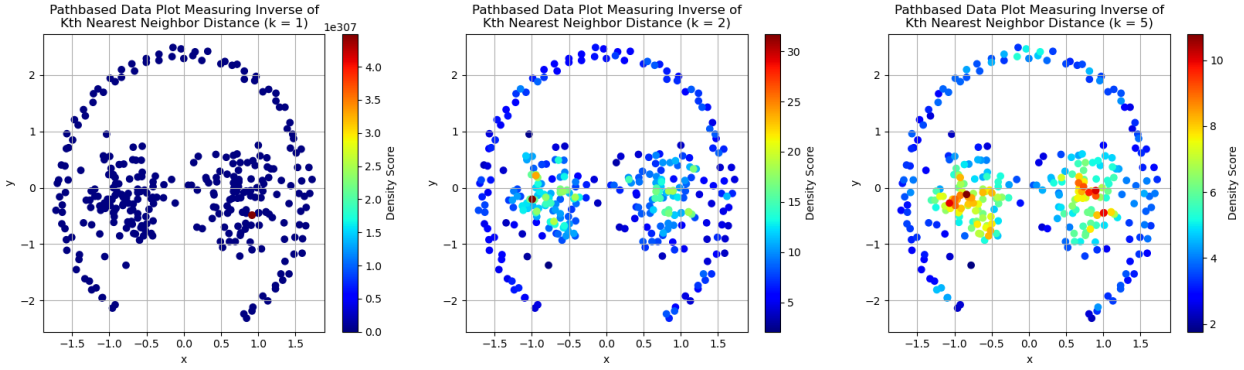

Figure 17: Pathbased Data Anomaly Score

	x	y	density_score
202	-0.776324	-1.374998	1.762296
101	0.843926	-2.315782	2.222204
270	-1.037263	0.951303	2.312673
1	-0.958375	-2.136178	2.334149
203	-1.061537	-1.221052	2.388068

Figure 18: Pathbased Data Top 5 Anomalies (k = 5)

For all above datasets, using the inverse of the distance to the 5th nearest neighbor as the density measure works best. The points with the shortest distance to the 5th nearest neighbor end up with a higher density making them less likely to be anomalies. This results in the anomalies becoming easily identifiable as the coldest points, farthest away from the warm, densely packed regions.

4.2 B. Density as the Inverse of the Average of all k Nearest Neighbor Distances

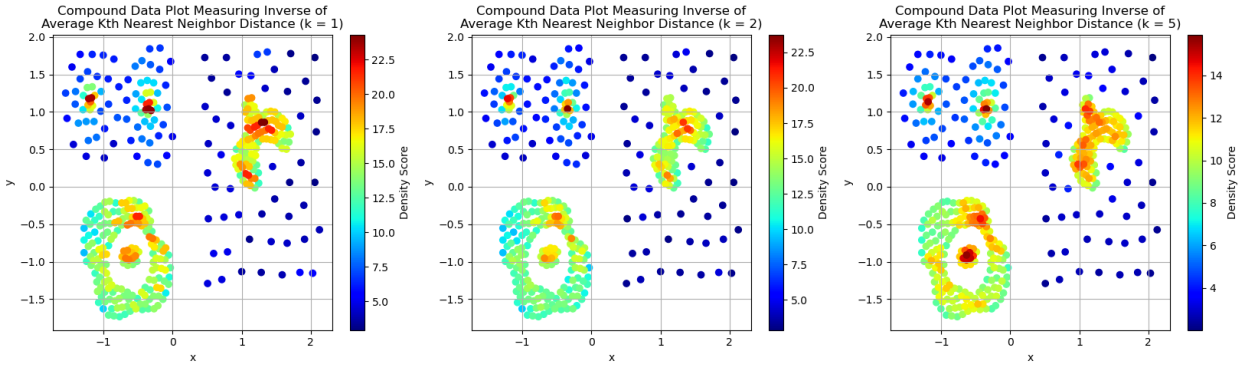

Figure 19: Compound Data Anomaly Score

	x	y	density_score
42	2.060247	1.726481	2.018238
28	2.060247	0.058959	2.236964
37	2.029397	-1.154744	2.240668
36	2.127088	-0.574277	2.257285
19	0.507475	-1.291945	2.258527

Figure 20: Compound Data Top 5 Anomalies (k = 5)

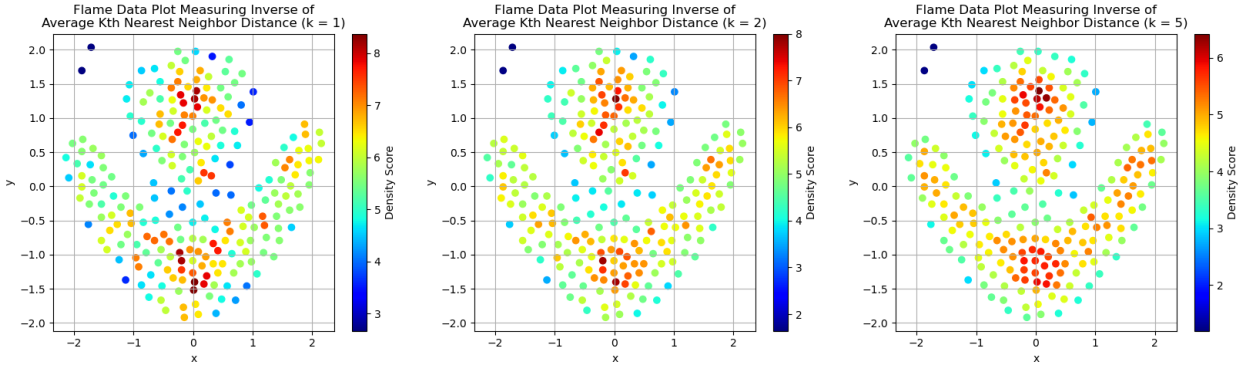

Figure 21: Flame Data Anomaly Score

	x	y	density_score
0	-1.712779	2.035183	1.190771
1	-1.869233	1.694577	1.229857
204	1.009519	1.383589	2.741640
192	0.946937	0.939320	2.783558
158	0.634029	-0.126925	2.865236

Figure 22: Flame Data Top 5 Anomalies (k = 5)

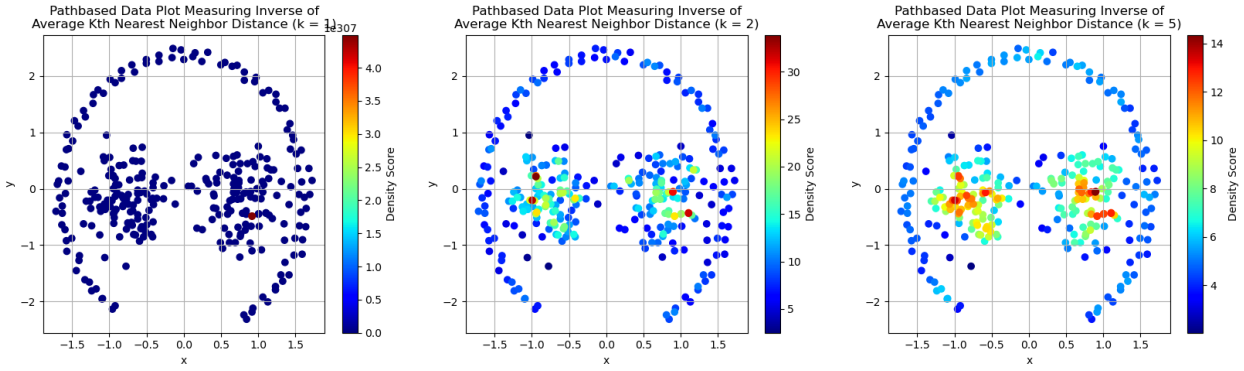

Figure 23: Pathbased Data Anomaly Score

	x	y	density_score
202	-0.776324	-1.374998	2.026356
270	-1.037263	0.951303	2.431010
203	-1.061537	-1.221052	2.715091
198	0.194612	0.557885	3.275485
197	0.133929	0.446701	3.338471

Figure 24: Pathbased Data Top 5 Anomalies (k = 1)

When using the inverse of the average of the distances to all k neighbors, k=5 results in the best anomaly detection. Here, the densely packed regions and the densest points are easily differentiable from the low density (high anomaly points). This is especially visible in the Compound and Flame datasets.

4.3 Conclusion

The effectiveness of anomaly detection models depends heavily on the data distribution. Parametric models like Mahalanobis distance are well-suited for roughly normal distributions, effectively identifying off-distribution anomalies in datasets like G-Data. For more complex distributions, distance-based models (e.g., distance to the k^{th} nearest neighbor) and density-based models (inverse distances) offer robust alternatives. Specifically, using the distance to the k^{th} nearest neighbor when (k=5) highlights anomalies farthest from compact clusters, while density-based models, particularly the inverse of the average distance to k neighbors, excel at distinguishing densely packed regions from sparse anomalies. Across varied datasets like Compound, Flame, and Pathbased, density-based scoring with k=5 proved to be the most effective and flexible approach.