Student: Víctor Iván López Rodríguez A00817161

Student: Samuel Guadalupe Rodríguez Rodríguez A00817512

Student: Josué Genaro Almaraz Rivera A00821189

CS5051.501 Computational Techniques for Machine Learning

Assignment 5 - Report

Due: November 03, 2021

Description of the datasets

For this assignment, we used 95 databases from the set that was provided by the professor. Among them we can find data about bacterium (e.g., ecoli1, ecoli2), yeast, wine quality, and dermatology.

For testing our algorithms and variations with these datasets, we uploaded them into Google Drive, mounted the unit in Google Colab, and using the *os* Python dependency, we went through the files directory, to work with the respective train and test data files.

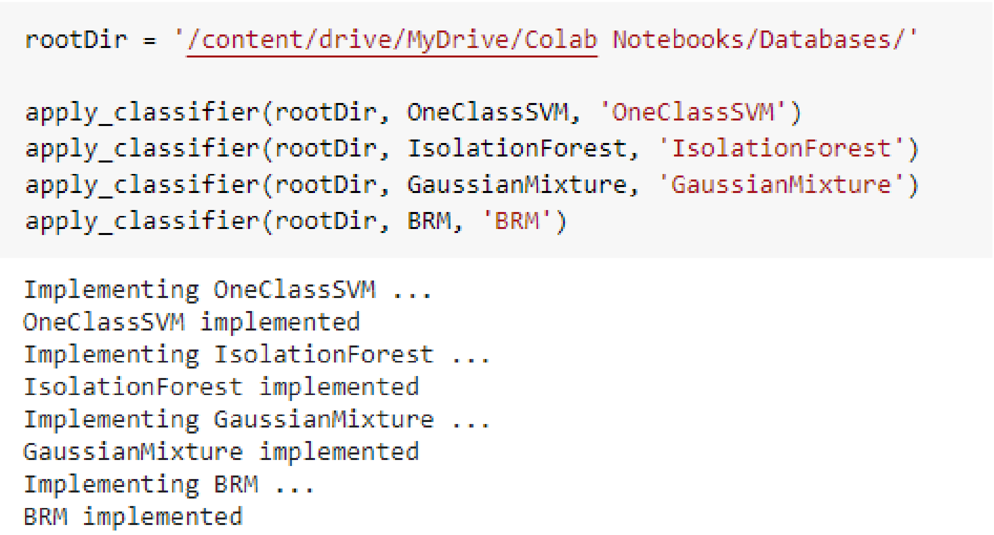
Following is the implementation of the requested homework points in the databases, using one-class classifiers, applying data standardization, and the creation of an own custom Bagging Random Miner model.

Source code can be found in this GitHub repository link:
https://github.com/JG-11/one-class-classifiers-anomaly-detection

1. Algorithms + Databases

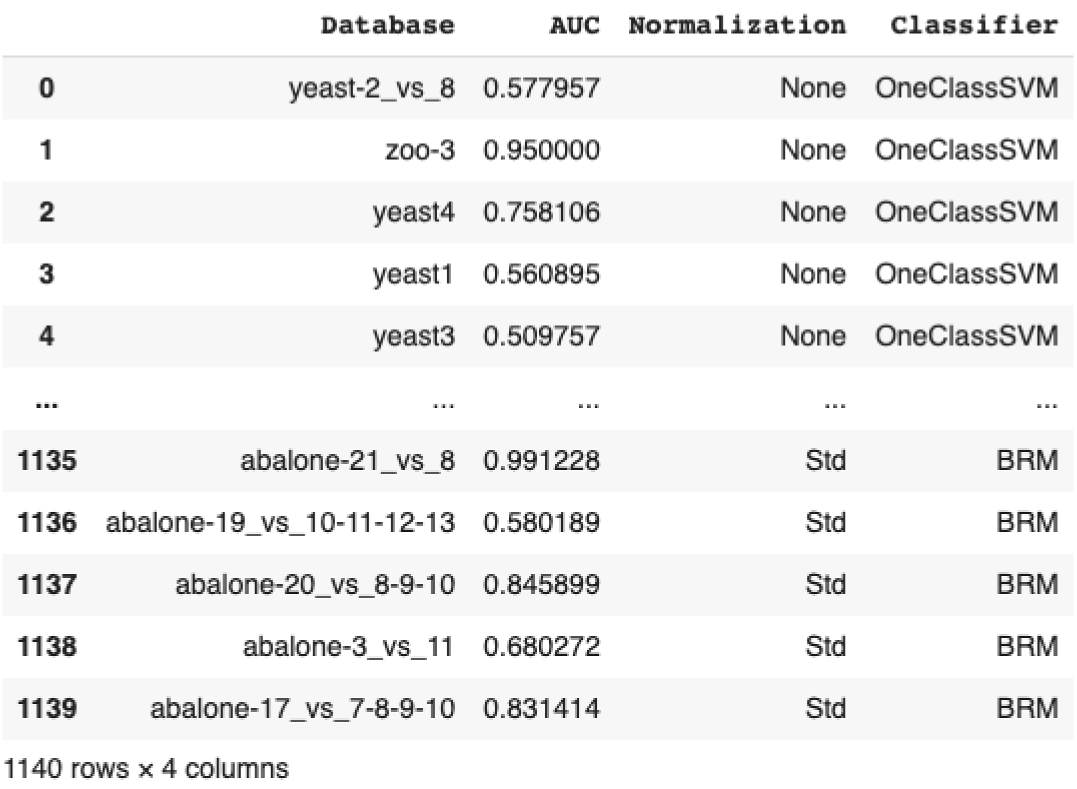
In this section, we evaluate and analyze the results of the algorithms Bagging Random Miner, Gaussian Mixture Model, Isolation Forest, and One Class SVM, according to the Area Under the Curve (AUC), in the 95 databases.

We used a code provided by the professor and modified it to loop through all the databases in Google Drive. We trained and tested the 4 algorithms with all the datasets and obtained the AUC results.



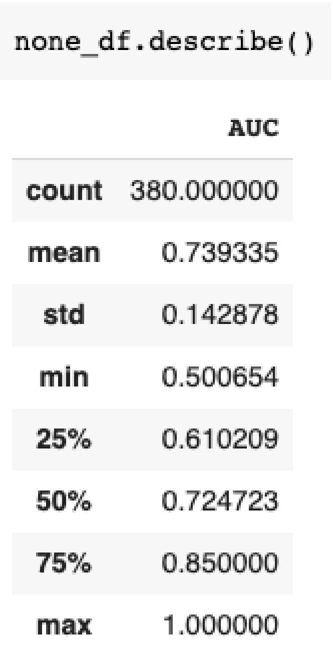
```
rootDir = '/content/drive/MyDrive/Colab Notebooks/Databases/'

apply_classifier(rootDir, OneClassSVM, 'OneClassSVM')
apply_classifier(rootDir, IsolationForest, 'IsolationForest')
apply_classifier(rootDir, GaussianMixture, 'GaussianMixture')
apply_classifier(rootDir, BRM, 'BRM')
```

```
Implementing OneClassSVM ...
OneClassSVM implemented
Implementing IsolationForest ...
IsolationForest implemented
Implementing GaussianMixture ...
GaussianMixture implemented
Implementing BRM ...
BRM implemented
```



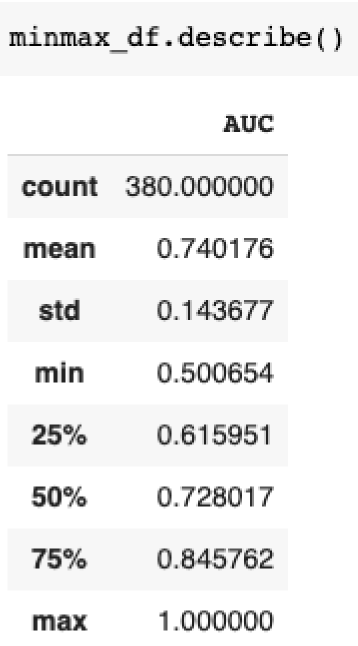
	Database	AUC	Normalization	Classifier
0	yeast-2_vs_8	0.577957	None	OneClassSVM
1	zoo-3	0.950000	None	OneClassSVM
2	yeast4	0.758106	None	OneClassSVM
3	yeast1	0.560895	None	OneClassSVM
4	yeast3	0.509757	None	OneClassSVM
...	...	...	...	...
1135	abalone-21_vs_8	0.991228	Std	BRM
1136	abalone-19_vs_10-11-12-13	0.580189	Std	BRM
1137	abalone-20_vs_8-9-10	0.845899	Std	BRM
1138	abalone-3_vs_11	0.680272	Std	BRM
1139	abalone-17_vs_7-8-9-10	0.831414	Std	BRM

1140 rows × 4 columns



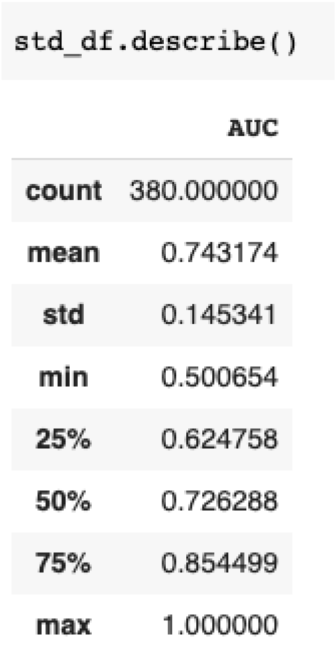
```
none_df.describe()
```

	AUC
count	380.000000
mean	0.739335
std	0.142878
min	0.500654
25%	0.610209
50%	0.724723
75%	0.850000
max	1.000000



```
minmax_df.describe()
```

	AUC
count	380.000000
mean	0.740176
std	0.143677
min	0.500654
25%	0.615951
50%	0.728017
75%	0.845762
max	1.000000



```
std_df.describe()
```

	AUC
count	380.000000
mean	0.743174
std	0.145341
min	0.500654
25%	0.624758
50%	0.726288
75%	0.854499
max	1.000000

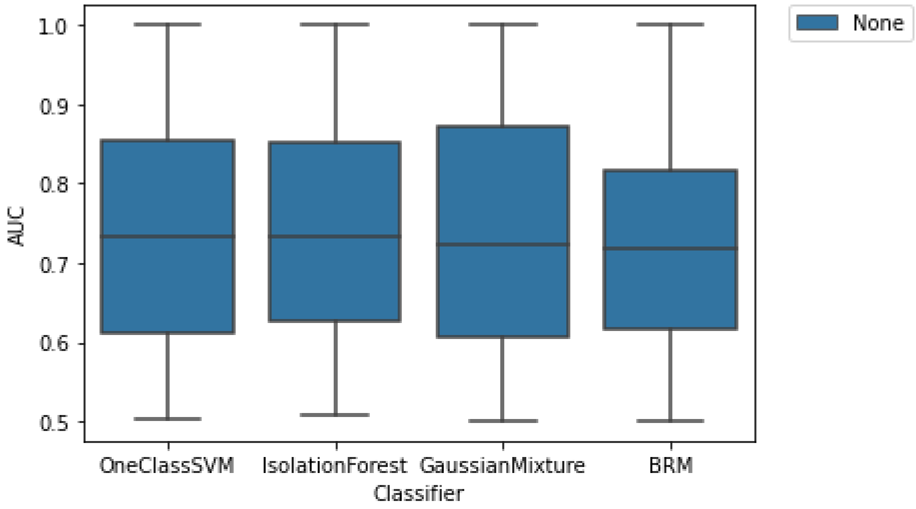
We calculated the AUC for the combination of all databases, classifiers and normalization methods. The results are 95 x 4 x 3 = 1140 experiments. With the last described tables we can see that the mean increased in less than 1% using normalization methods. There is a slight increase in AUC using standard normalization compared to MinMax scaling.

2. Algorithms + Databases + Statistical Tests + Visualizations

Using the AUC results of each database for each algorithm, we generated a box plot to visualize the results, a Friedman test to see if there's a statistical difference, a post hoc test to see how the models differ, and a Critical Difference (CD) diagram with the results of these tests.

a) Without Scaling or Normalizing

- *Boxplot*



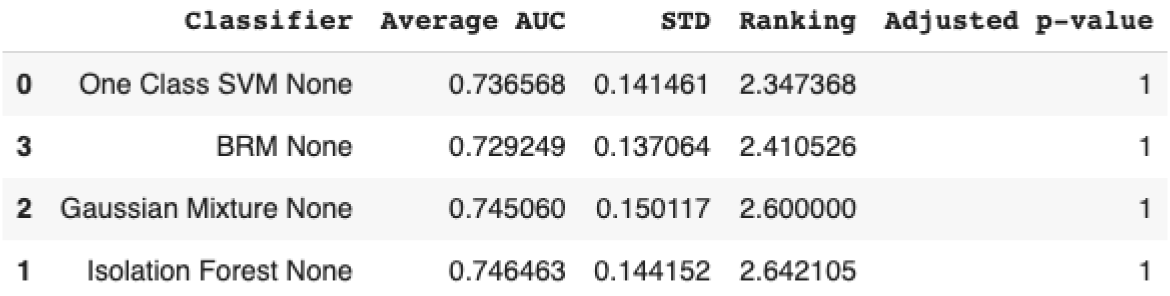
- *Friedman Test and Shaffer Test*

We obtained a statistic = 3.5999, pvalue = 0.3080, and the following ranking:

Algorithm	Ranking
One Class SVM	2.3473684210526318
Isolation Forest	2.642105263157895
Gaussian Mixture	2.60
BRM	2.4105263157894736

With a significance level of alpha = 0.05, we can accept the null hypothesis, so the results obtained by all the tested classifiers, in all the databases, are similar statistically.

Even though we accept the null hypothesis, we proceeded to make the Shaffer post hoc test:



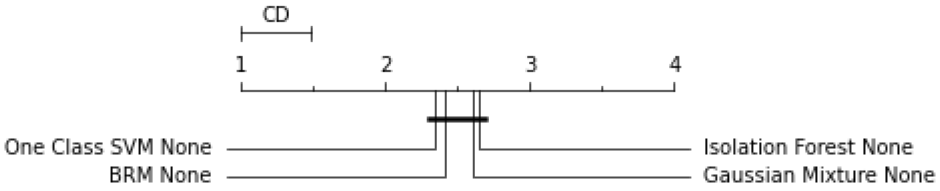
	Classifier	Average AUC	STD	Ranking	Adjusted p-value
0	One Class SVM None	0.736568	0.141461	2.347368	1
3	BRM None	0.729249	0.137064	2.410526	1
2	Gaussian Mixture None	0.745060	0.150117	2.600000	1
1	Isolation Forest None	0.746463	0.144152	2.642105	1

Algorithms	Pvalue
AUC_OneClass_SVM_None vs	0.7681949167459463

AUC_Isolation_Forest_None	
AUC_OneClass_SVM_None vs AUC_Gaussian_Mixture_None	0.800552924922785
AUC_Isolation_Forest_None vs AUC_BRM_None	0.8168650606090959
AUC_Gaussian_Mixture_None vs AUC_BRM_None	0.8497215784715073
AUC_OneClass_SVM_None vs AUC_BRM_None	0.9496407729386862
AUC_Isolation_Forest_None vs AUC_Gaussian_Mixture_None	0.9664147845002375

- *Critical Diagram*

This CD diagram was generated using the rankings produced by the Friedman Test.



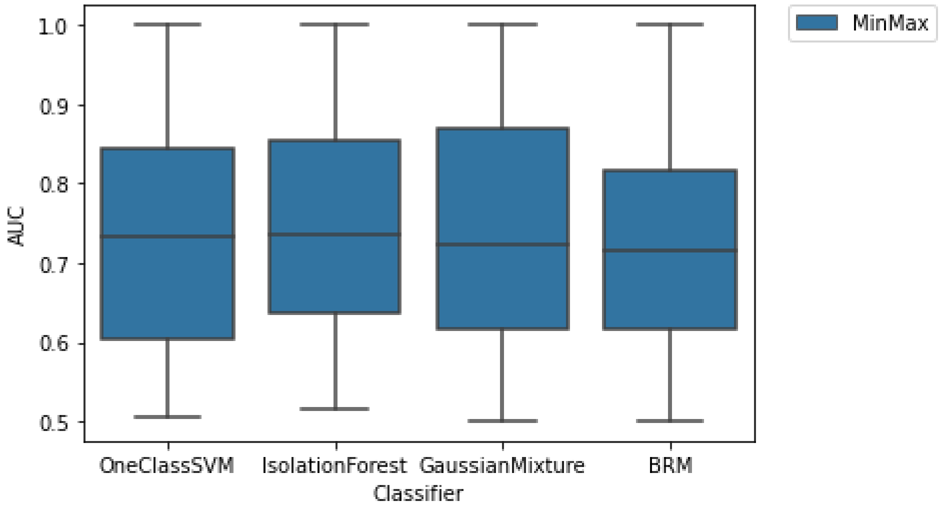
We can clearly see that on average the One Class SVM and BRM have a better ranking over the 95 databases. The thick horizontal line groups all classifiers, this means that they are not significantly different.

3. Algorithms + Databases + Statistical Tests + Visualizations + Data Transformation

In this section, we evaluated the 95 databases using the 4 mentioned algorithms (Bagging Random Miner, Gaussian Mixture Model, Isolation Forest, and One Class SVM). Also, we added to the evaluation without scaling or normalizing the database (as in the previous section), the MinMax scaling and Standard normalizing. We generated a box plot to visualize the results, a Friedman test with a post hoc test, and a CD diagram with the results of the statistical tests.

a) MinMax Scaling

- *Boxplot*



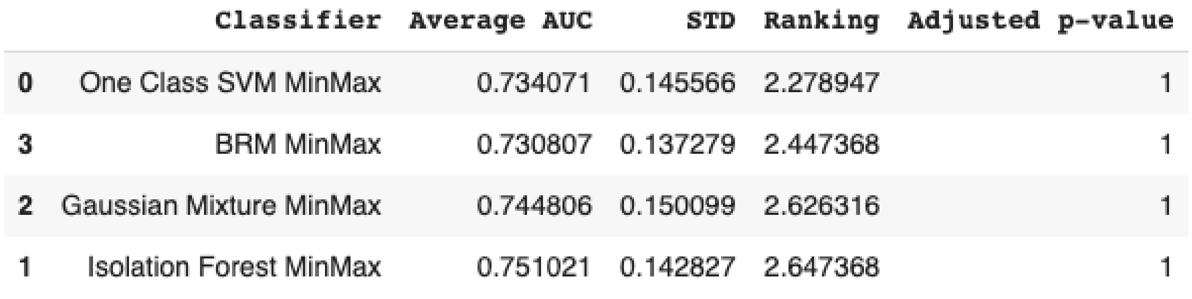
- *Friedman Test and Shaffer Test*

We obtained a statistic = 5.2000, pvalue = 0.1577, and the following ranking:

Algorithm	Ranking
One Class SVM	2.278947368421053
Isolation Forest	2.642105263157895
Gaussian Mixture	2.626315789473684
BRM	2.4473684210526314

With a significance level of alpha = 0.05, we can accept the null hypothesis, so the results obtained by all the tested classifiers, in all the databases, are similar statistically.

Even though we accept the null hypothesis, we proceeded to make the Shaffer post hoc test.

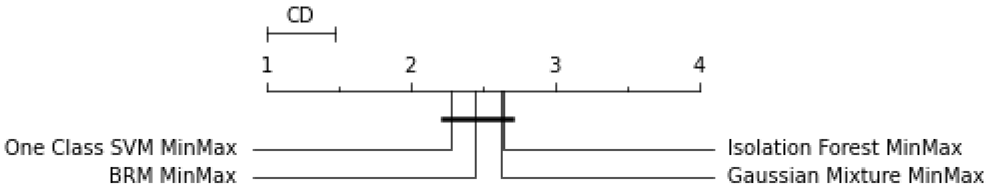


	Classifier	Average AUC	STD	Ranking	Adjusted p-value
0	One Class SVM MinMax	0.734071	0.145566	2.278947	1
3	BRM MinMax	0.730807	0.137279	2.447368	1
2	Gaussian Mixture MinMax	0.744806	0.150099	2.626316	1
1	Isolation Forest MinMax	0.751021	0.142827	2.647368	1

Algorithms	Pvalue
AUC_OneClass_SVM_MinMax vs AUC_Isolation_Forest_MinMax	0.7125593020062799
AUC_OneClass_SVM_MinMax vs AUC_Gaussian_Mixture_MinMax	0.7283145543573779
AUC_Isolation_Forest_MinMax vs AUC_BRM_MinMax	0.8414805811217938
AUC_Gaussian_Mixture_MinMax vs AUC_BRM_MinMax	0.8579790284090667
AUC_OneClass_SVM_MinMax vs AUC_BRM_MinMax	0.8662520470606498
AUC_Isolation_Forest_MinMax vs AUC_Gaussian_Mixture_MinMax	0.9832036710341057

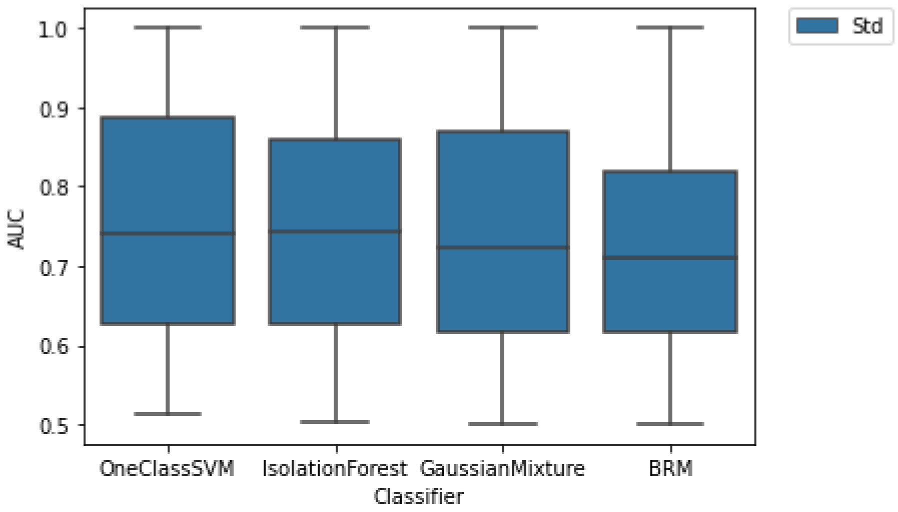
- Critical Diagram

This CD diagram was generated using the rankings produced by the Friedman Test.



b) Standard Normalizing

- *Boxplot*



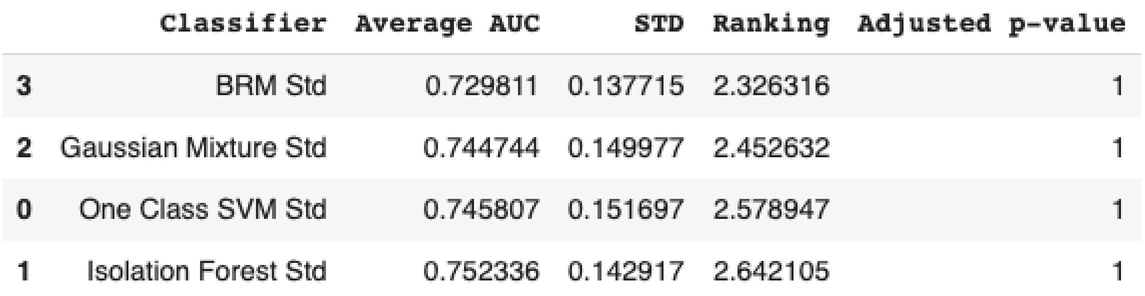
- *Friedman Test and Shaffer Test*

We obtained a statistic = 3.4368, pvalue = 0.3290, and the following ranking:

Algorithm	Ranking
AUC_OneClass_SVM_MinMax	2.5789473684210527
AUC_Isolation_Forest_MinMax	2.642105263157895
AUC_Gaussian_Mixture_MinMax	2.4526315789473685
AUC_BRM_MinMax	2.3263157894736843

With a significance level of alpha = 0.05, we can accept the null hypothesis, so the results obtained by all the tested classifiers, in all the databases, are similar statistically.

Even though we accept the null hypothesis, we proceeded to make the Shaffer post hoc test.

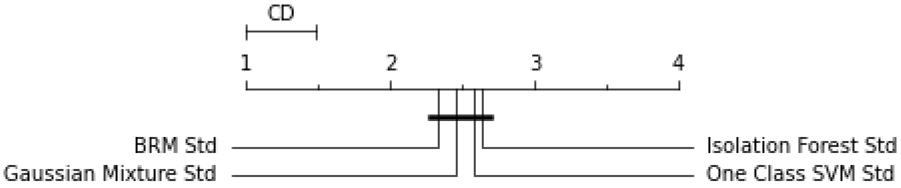


	Classifier	Average AUC	STD	Ranking	Adjusted p-value
3	BRM Std	0.729811	0.137715	2.326316	1
2	Gaussian Mixture Std	0.744744	0.149977	2.452632	1
0	One Class SVM Std	0.745807	0.151697	2.578947	1
1	Isolation Forest Std	0.752336	0.142917	2.642105	1

Algorithms	Pvalue
AUC_Isolation_Forest_Std vs AUC_BRM_Std	0.7521623083780538
AUC_OneClass_SVM_Std vs AUC_BRM_Std	0.800552924922785
AUC_Isolation_Forest_Std vs AUC_Gaussian_Mixture_Std	0.8497215784715073
AUC_OneClass_SVM_Std vs AUC_Gaussian_Mixture_Std	0.899481958176398
AUC_Gaussian_Mixture_Std vs AUC_BRM_Std	0.899481958176398
AUC_OneClass_SVM_Std vs AUC_Isolation_Forest_Std	0.949640772938686

- *Critical Diagram*

This CD diagram was generated using the rankings produced by the Friedman Test.

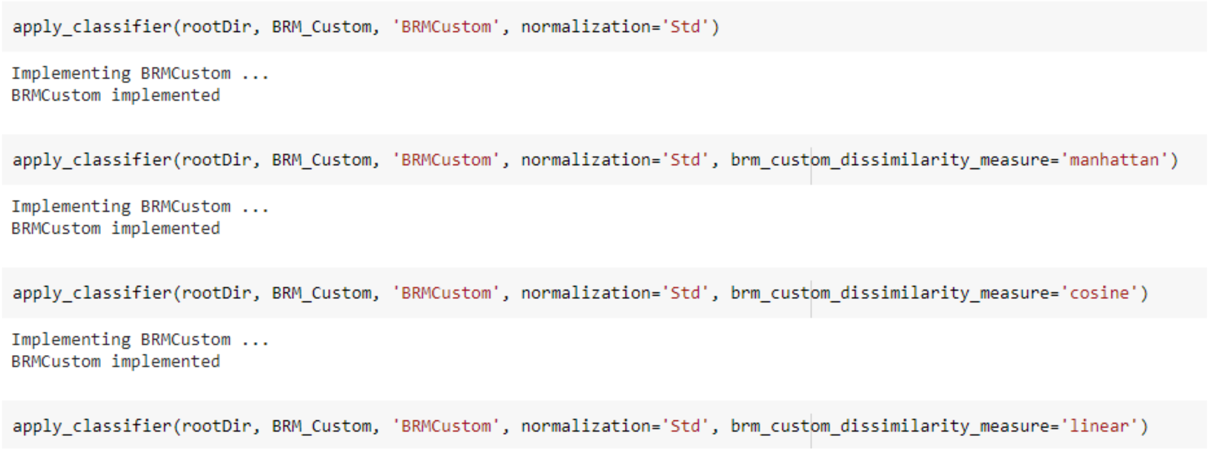


4. **Algorithms + Databases + Statistical Tests + Visualizations + Data Transformation + Dissimilarity measure**

In this section, we evaluated the 95 databases using the 4 mentioned algorithms (Bagging Random Miner, Gaussian Mixture Model, Isolation Forest, and One Class SVM). Also, we added to the evaluation without scaling or normalizing the database (as in the previous section), the MinMax scaling and Standard normalizing. We generated a box plot to visualize the results, a Friedman test with a post hoc test, and a CD diagram with the results of the statistical tests.

We modified the BRM class to accept any dissimilarity measure. We tried with euclidean (default), manhattan, cosine and linear distances.

Example:

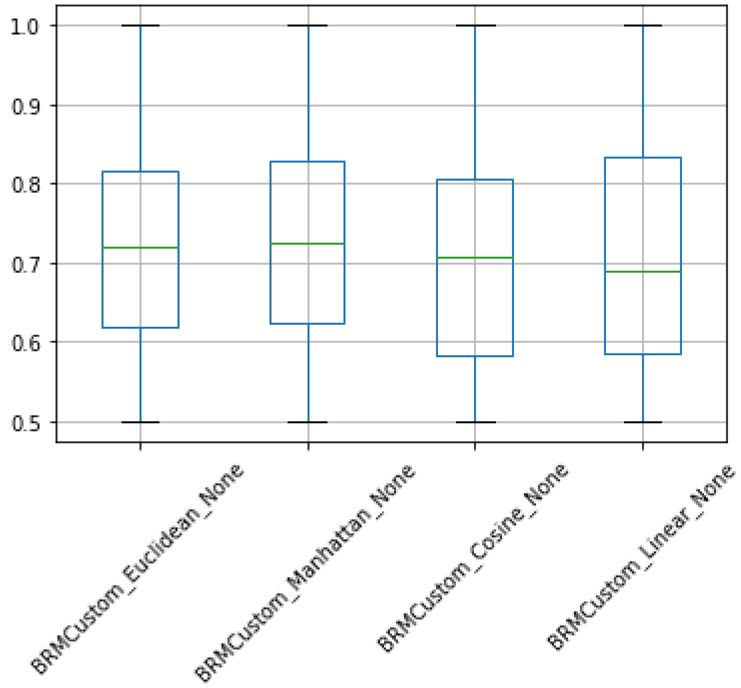


```
apply_classifier(rootDir, BRM_Custom, 'BRMCustom', normalization='Std')

Implementing BRMCustom ...
BRMCustom implemented

apply_classifier(rootDir, BRM_Custom, 'BRMCustom', normalization='Std', brm_custom_dissimilarity_measure='manhattan')

Implementing BRMCustom ...
BRMCustom implemented

apply_classifier(rootDir, BRM_Custom, 'BRMCustom', normalization='Std', brm_custom_dissimilarity_measure='cosine')

Implementing BRMCustom ...
BRMCustom implemented

apply_classifier(rootDir, BRM_Custom, 'BRMCustom', normalization='Std', brm_custom_dissimilarity_measure='linear')
```

We generated a box plot to visualize the results, a Friedman test with a post hoc test, and a CD diagram with the results of the statistical tests.

a) BRM Custom without Scaling and Euclidean, Manhattan, Cosine and Linear distances

- *Boxplot*



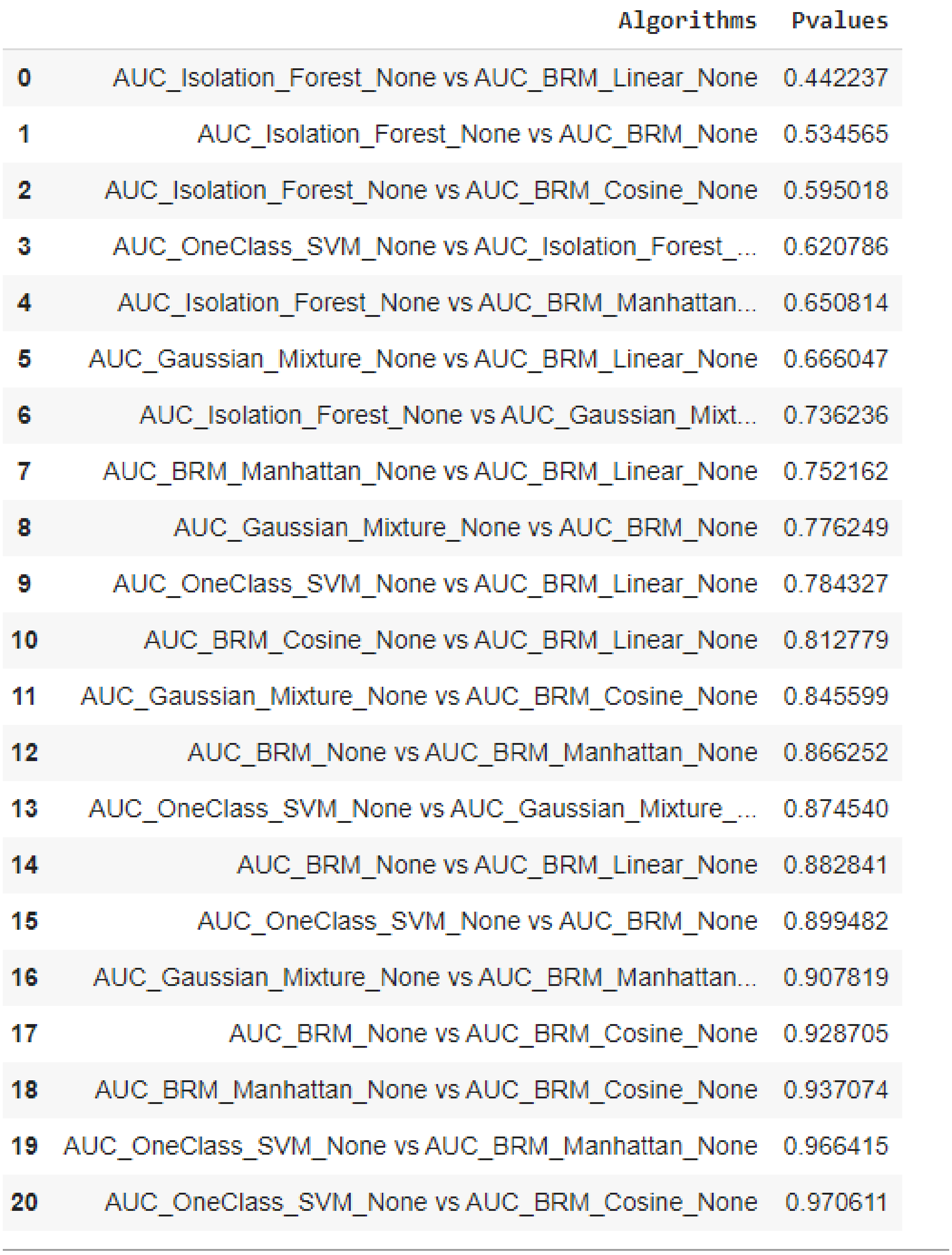
- *Friedman Test and Shaffer Test*

We obtained a statistic = 7.2620, pvalue = 0.2972, and the following ranking:

Algorithm	Ranking
One Class SVM	3.963157894736842
Isolation Forest	4.457894736842105
Gaussian Mixture	4.121052631578947
BRM	3.836842105263158
Custom BRM Manhattan	4.005263157894737
Custom BRM Cosine	3.9263157894736844
Custom BRM Linear	3.6894736842105265

With a significance level of alpha = 0.05, we can accept the null hypothesis, so the results obtained by all the tested classifiers, in all the databases, are similar statistically.

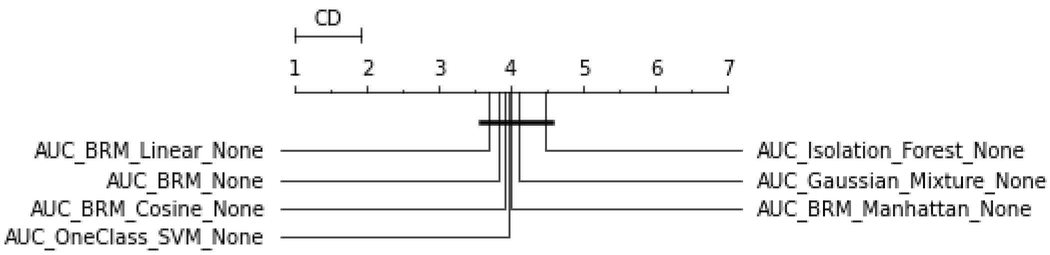
Even though we accept the null hypothesis, we proceeded to make the Shaffer post hoc test.



	Algorithms	Pvalues
0	AUC_Isolation_Forest_None vs AUC_BRM_Linear_None	0.442237
1	AUC_Isolation_Forest_None vs AUC_BRM_None	0.534565
2	AUC_Isolation_Forest_None vs AUC_BRM_Cosine_None	0.595018
3	AUC_OneClass_SVM_None vs AUC_Isolation_Forest_...	0.620786
4	AUC_Isolation_Forest_None vs AUC_BRM_Manhattan...	0.650814
5	AUC_Gaussian_Mixture_None vs AUC_BRM_Linear_None	0.666047
6	AUC_Isolation_Forest_None vs AUC_Gaussian_Mixt...	0.736236
7	AUC_BRM_Manhattan_None vs AUC_BRM_Linear_None	0.752162
8	AUC_Gaussian_Mixture_None vs AUC_BRM_None	0.776249
9	AUC_OneClass_SVM_None vs AUC_BRM_Linear_None	0.784327
10	AUC_BRM_Cosine_None vs AUC_BRM_Linear_None	0.812779
11	AUC_Gaussian_Mixture_None vs AUC_BRM_Cosine_None	0.845599
12	AUC_BRM_None vs AUC_BRM_Manhattan_None	0.866252
13	AUC_OneClass_SVM_None vs AUC_Gaussian_Mixture_...	0.874540
14	AUC_BRM_None vs AUC_BRM_Linear_None	0.882841
15	AUC_OneClass_SVM_None vs AUC_BRM_None	0.899482
16	AUC_Gaussian_Mixture_None vs AUC_BRM_Manhattan...	0.907819
17	AUC_BRM_None vs AUC_BRM_Cosine_None	0.928705
18	AUC_BRM_Manhattan_None vs AUC_BRM_Cosine_None	0.937074
19	AUC_OneClass_SVM_None vs AUC_BRM_Manhattan_None	0.966415
20	AUC_OneClass_SVM_None vs AUC_BRM_Cosine_None	0.970611

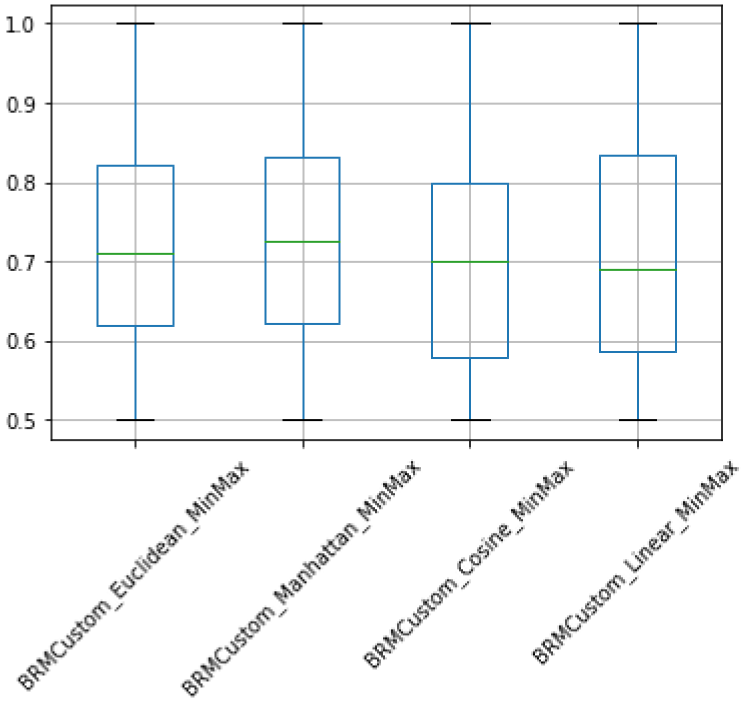
- Critical Diagram

This CD diagram was generated using the rankings produced by the Friedman Test.



b) **BRM Custom with MinMax Scaling and Euclidean, Manhattan, Cosine and Linear distances**

- *Boxplot*



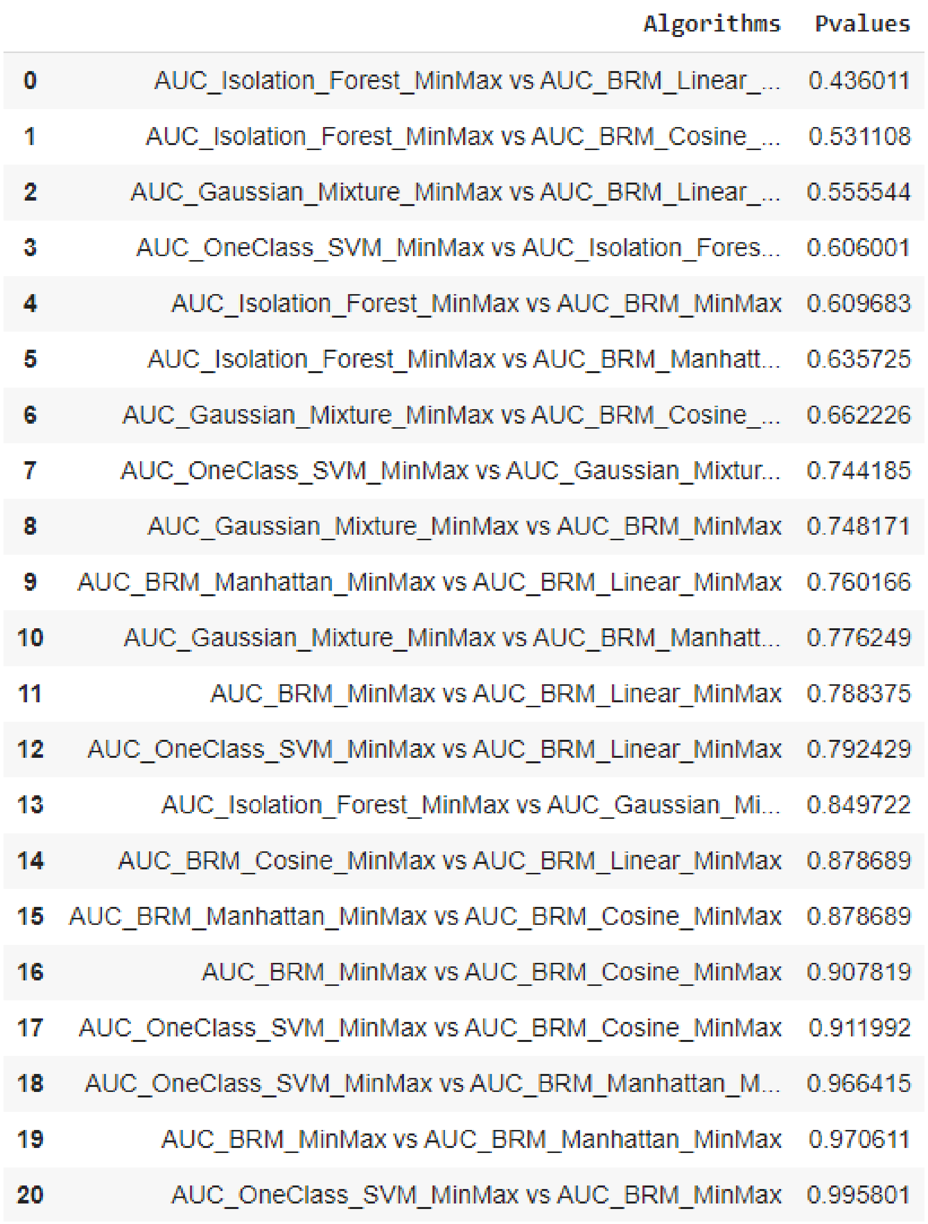
- *Friedman Test and Shaffer Test*

We obtained a statistic = 8.569, pvalue = 0.199, and the following ranking:

Algorithm	Ranking
One Class SVM	3.9263157894736844
Isolation Forest	4.442105263157894
Gaussian Mixture	4.252631578947368
BRM	3.931578947368421
Custom BRM Manhattan	3.968421052631579
Custom BRM Cosine	3.8157894736842106
Custom BRM Linear	3.663157894736842

With a significance level of alpha = 0.05, we can accept the null hypothesis, so the results obtained by all the tested classifiers, in all the databases, are similar statistically.

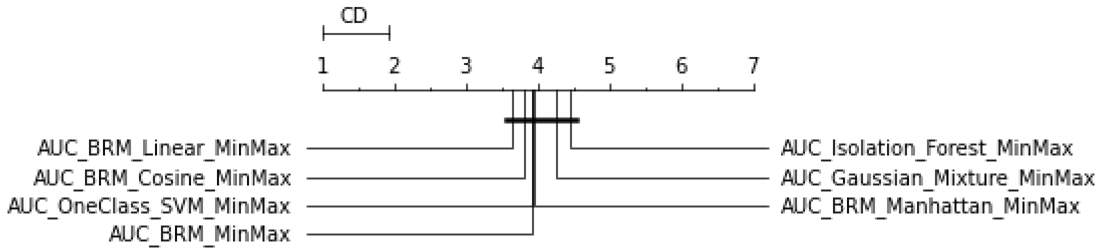
Even though we accept the null hypothesis, we proceeded to make the Shaffer post hoc test.



	Algorithms	Pvalues
0	AUC_Isolation_Forest_MinMax vs AUC_BRM_Linear_...	0.436011
1	AUC_Isolation_Forest_MinMax vs AUC_BRM_Cosine_...	0.531108
2	AUC_Gaussian_Mixture_MinMax vs AUC_BRM_Linear_...	0.555544
3	AUC_OneClass_SVM_MinMax vs AUC_Isolation_Fores...	0.606001
4	AUC_Isolation_Forest_MinMax vs AUC_BRM_MinMax	0.609683
5	AUC_Isolation_Forest_MinMax vs AUC_BRM_Manhatt...	0.635725
6	AUC_Gaussian_Mixture_MinMax vs AUC_BRM_Cosine_...	0.662226
7	AUC_OneClass_SVM_MinMax vs AUC_Gaussian_Mixtur...	0.744185
8	AUC_Gaussian_Mixture_MinMax vs AUC_BRM_MinMax	0.748171
9	AUC_BRM_Manhattan_MinMax vs AUC_BRM_Linear_MinMax	0.760166
10	AUC_Gaussian_Mixture_MinMax vs AUC_BRM_Manhatt...	0.776249
11	AUC_BRM_MinMax vs AUC_BRM_Linear_MinMax	0.788375
12	AUC_OneClass_SVM_MinMax vs AUC_BRM_Linear_MinMax	0.792429
13	AUC_Isolation_Forest_MinMax vs AUC_Gaussian_Mi...	0.849722
14	AUC_BRM_Cosine_MinMax vs AUC_BRM_Linear_MinMax	0.878689
15	AUC_BRM_Manhattan_MinMax vs AUC_BRM_Cosine_MinMax	0.878689
16	AUC_BRM_MinMax vs AUC_BRM_Cosine_MinMax	0.907819
17	AUC_OneClass_SVM_MinMax vs AUC_BRM_Cosine_MinMax	0.911992
18	AUC_OneClass_SVM_MinMax vs AUC_BRM_Manhattan_M...	0.966415
19	AUC_BRM_MinMax vs AUC_BRM_Manhattan_MinMax	0.970611
20	AUC_OneClass_SVM_MinMax vs AUC_BRM_MinMax	0.995801

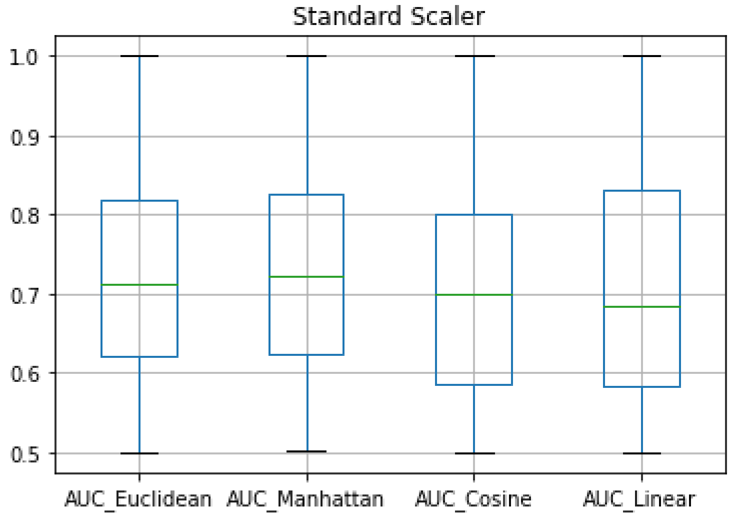
- Critical Diagram

This CD diagram was generated using the rankings produced by the Friedman Test.



c) BRM Custom with Standard Normalization and Euclidean, Manhattan, Cosine and Linear distances

- *Boxplot*



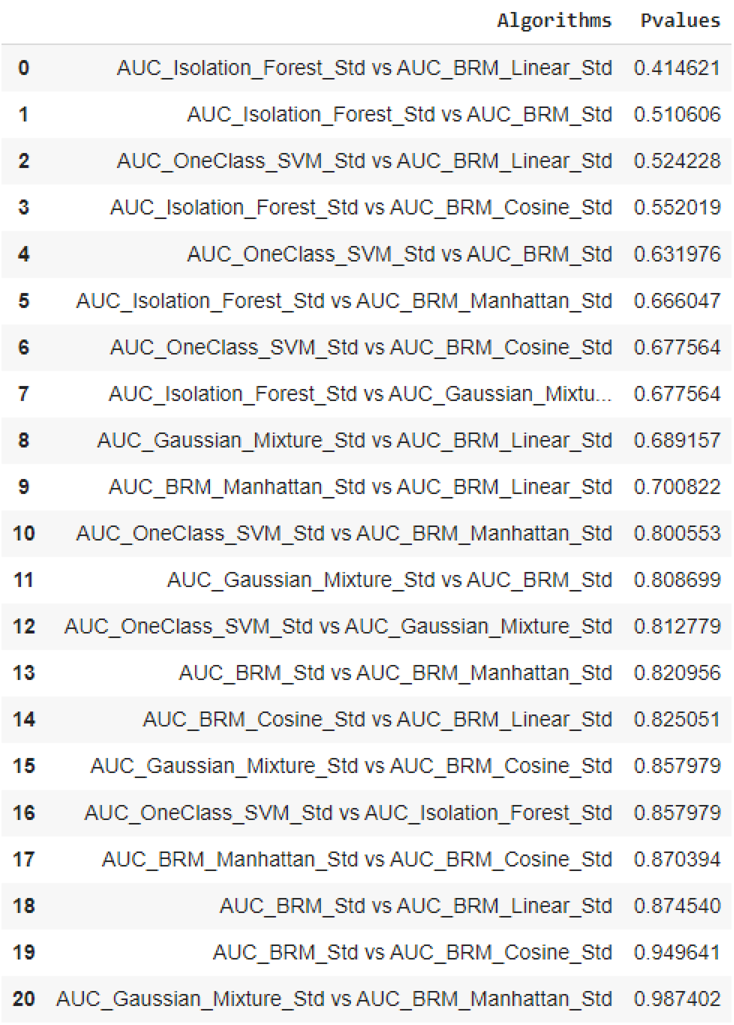
- *Friedman Test and Shaffer Test*

We obtained a statistic = 9.772, pvalue = 0.1345, and the following ranking:

Algorithm	Ranking
One Class SVM	4.2631578947368425
Isolation Forest	4.442105263157894
Gaussian Mixture	4.026315789473684
BRM	3.7842105263157895
Custom BRM Manhattan	4.010526315789473
Custom BRM Cosine	3.8473684210526318
Custom BRM Linear	3.626315789473684

With a significance level of alpha = 0.05, we can accept the null hypothesis, so the results obtained by all the tested classifiers, in all the databases, are similar statistically.

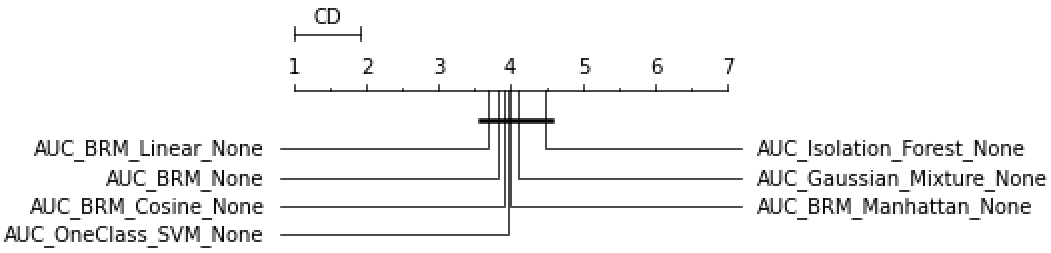
Even though we accept the null hypothesis, we proceeded to make the Shaffer post hoc test.



	Algorithms	Pvalues
0	AUC_Isolation_Forest_Std vs AUC_BRM_Linear_Std	0.414621
1	AUC_Isolation_Forest_Std vs AUC_BRM_Std	0.510606
2	AUC_OneClass_SVM_Std vs AUC_BRM_Linear_Std	0.524228
3	AUC_Isolation_Forest_Std vs AUC_BRM_Cosine_Std	0.552019
4	AUC_OneClass_SVM_Std vs AUC_BRM_Std	0.631976
5	AUC_Isolation_Forest_Std vs AUC_BRM_Manhattan_Std	0.666047
6	AUC_OneClass_SVM_Std vs AUC_BRM_Cosine_Std	0.677564
7	AUC_Isolation_Forest_Std vs AUC_Gaussian_Mixtu...	0.677564
8	AUC_Gaussian_Mixture_Std vs AUC_BRM_Linear_Std	0.689157
9	AUC_BRM_Manhattan_Std vs AUC_BRM_Linear_Std	0.700822
10	AUC_OneClass_SVM_Std vs AUC_BRM_Manhattan_Std	0.800553
11	AUC_Gaussian_Mixture_Std vs AUC_BRM_Std	0.808699
12	AUC_OneClass_SVM_Std vs AUC_Gaussian_Mixture_Std	0.812779
13	AUC_BRM_Std vs AUC_BRM_Manhattan_Std	0.820956
14	AUC_BRM_Cosine_Std vs AUC_BRM_Linear_Std	0.825051
15	AUC_Gaussian_Mixture_Std vs AUC_BRM_Cosine_Std	0.857979
16	AUC_OneClass_SVM_Std vs AUC_Isolation_Forest_Std	0.857979
17	AUC_BRM_Manhattan_Std vs AUC_BRM_Cosine_Std	0.870394
18	AUC_BRM_Std vs AUC_BRM_Linear_Std	0.874540
19	AUC_BRM_Std vs AUC_BRM_Cosine_Std	0.949641
20	AUC_Gaussian_Mixture_Std vs AUC_BRM_Manhattan_Std	0.987402

- Critical Diagram

This CD diagram was generated using the rankings produced by the Friedman Test.



References

1. (Villa-Pérez et al. 2021) M. E. Villa-Pérez, M. A. Álvarez-Carmona, O. Loyola-González, M. A. Medina-Pérez, J. C. Velazco-Rossell, K.-K. Raymond Choo, "Semi-supervised anomaly detection algorithms: A comparative summary and future research directions," Knowledge-Based Systems, vol. 218, pp. 106878, 2021.

2. (Benito Camiña et al. 2019) J. Benito Camiña, M. A. Medina-Pérez, R. Monroy, O. Loyola-González, L. A. Pereyra-Villanueva, L. C. González-Gurrola, "Bagging-RandomMiner: A one-class classifier for file access-based masquerade detection," *Machine Vision and Applications*, vol. 30, no. 5, pp. 959-974, 2019.

3. Derrac, J., García, S., Molina, D., & Herrera, F. (2011). “A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms”. Swarm and Evolutionary Computation, 1(1), 3–18. doi:10.1016/j.swevo.2011.02.002.