SPATIAL ANALYSIS OF 2023 EARTHQUAKES IN TURKIYE

Presentation by Gozde Yazganoglu
Supervised by Diego Bodas

Master of Data Science | 2023

CUNEF
UNIVERSIDAD

INDEX

INTRODUCTION

- On 6th Of February 2023 a series of Earthquakes have hit the south-east of Turkiye which has affected a very large area.
    - Pazarcik (7.7 RMS in point (37.043, 37.288))
    - Elbistan (7.6 in point (37.239, 38.089))
    - Nurdagi (6.6 in point (36.920, 37.304))
- Resulted in
    - 14 million people (16% of the population) affected
    - 350 000 km2 area (size of whole Germany) affected
    - 59000 people died (Turkiye + Syria)
    - 122000 people are injured
    - $118.8 billion valued economic loss

Source: Copernicus EU

RESEARCH QUESTIONS

- If there is a relationship between damage level and the geophysical fault information?
- How are clustered buildings and different level of damages?
- By using a machine learning algorithm can we find what makes an area more vulnerable to damage?
- Which regions are affected worse compared to other regions?

LITERATURE REVIEV

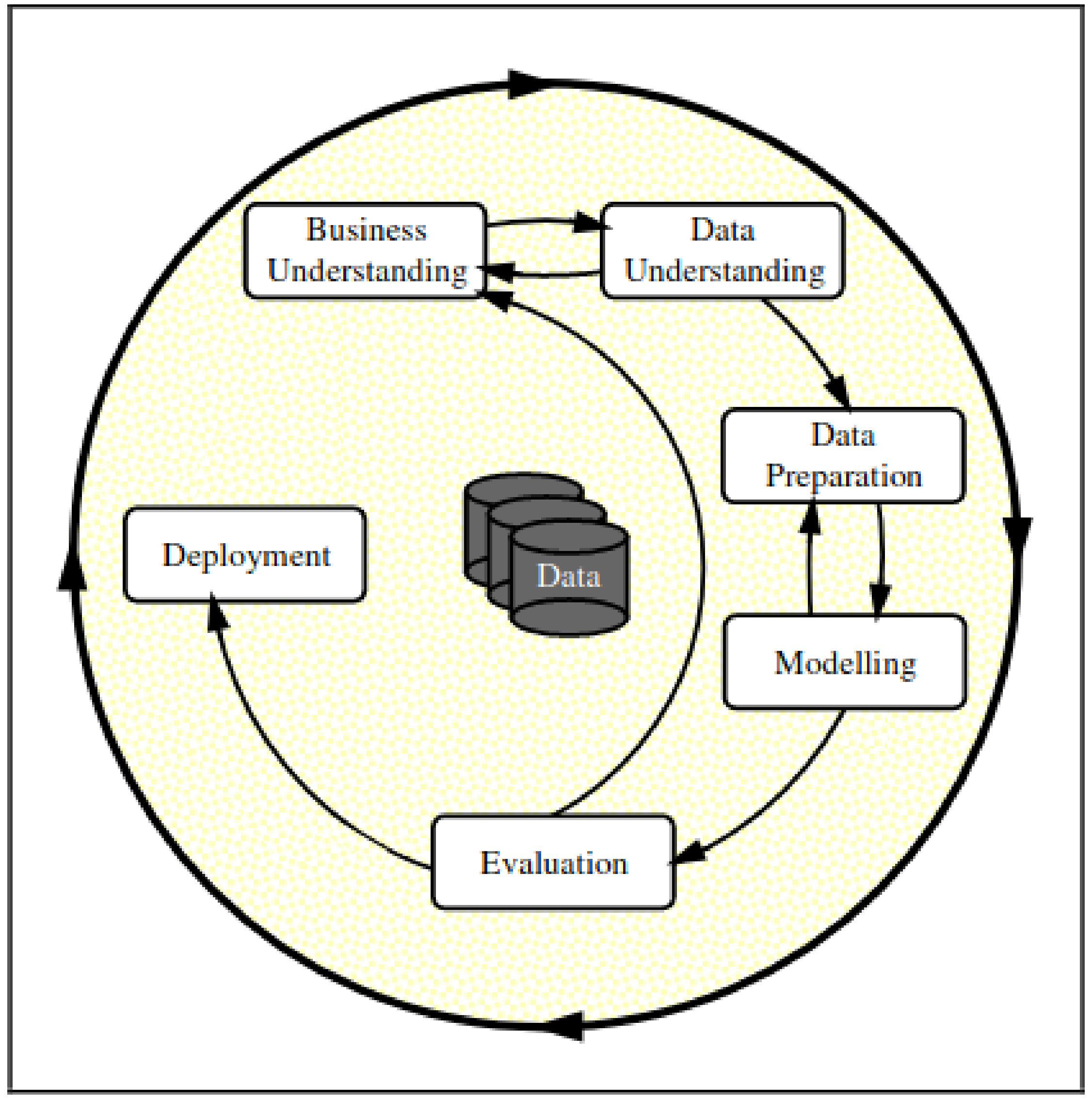
- Spatial Data :
  - Type of information which is related to location, shapes, distances to specific locations, postal code, city...etc that could be mapped.
- Types of Spatial Data:
  - Raster Data (Photos , Numbers, Pixels)
  - Vector Data (Point, line, multi lines, Polygon, Multi polygons)
- GIS (Geographic Information Systems)
  - ArcGIS, QGIS, PostgreGIS...

LITERATURE REVIEV

- Earthquake Engineering Research Institute (EERI) 2020 utilized artificial neural network (ANN) modeling for hazard detection.
- Priambodo et al. (2020) executed a temporal-spatial analysis on Indonesia's seismic data using ANN. They emphasized the need for more geo-specific data for accurate predictions.
- Garg, Masih, and Sharma (2021) focused on bridge damage assessments post-earthquakes and underscored the significance of **distance from the earthquake center**.
- Mangalathu et al. (2020) employed machine learning to assess building damage and achieved a 65% accuracy with their model. They emphasized building age and **fault distance** as key variables.
- Ahmad et al. used actuarial methods to use calculate risk premium for insurance products.

METHODOLOGY - MAIN

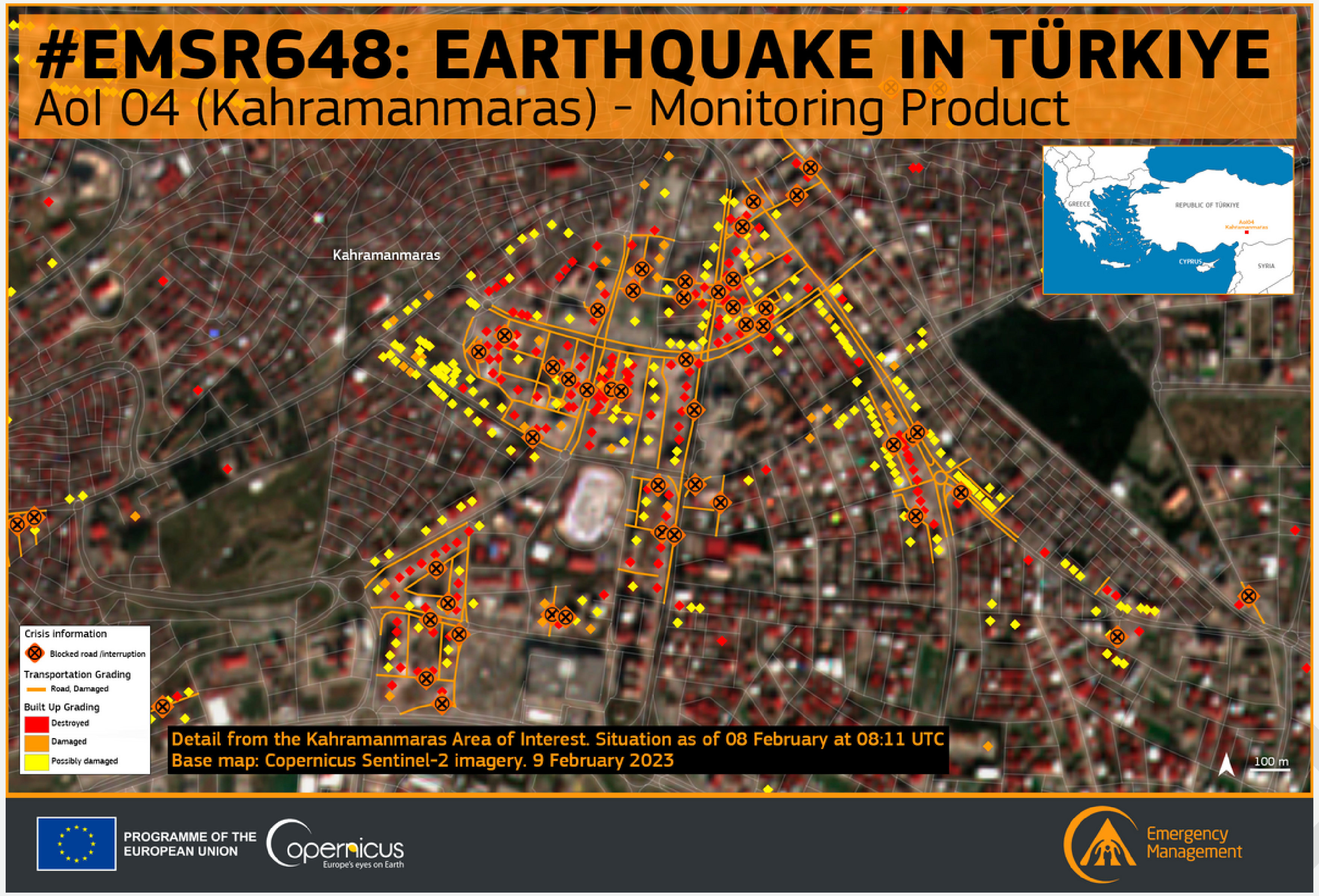
- CRISP-DM Method
  - Business Understanding
  - Data Understandind
  - Data Prepareation
  - Modelling
  - Evaluation
  - Deployment



Source: Wirth and Hipp (2000)

DATA UNDESTNANDING

- Copernicus EU Disaster Data
  - Buildings, Roads, Facilities
  - Campsites
  - Watersources
  - Elevation coutour
  - Area of Interest
- TURKSTAT (TUIK)
- ITU -ATAG - Faults Data
- AFAD - Distaster data



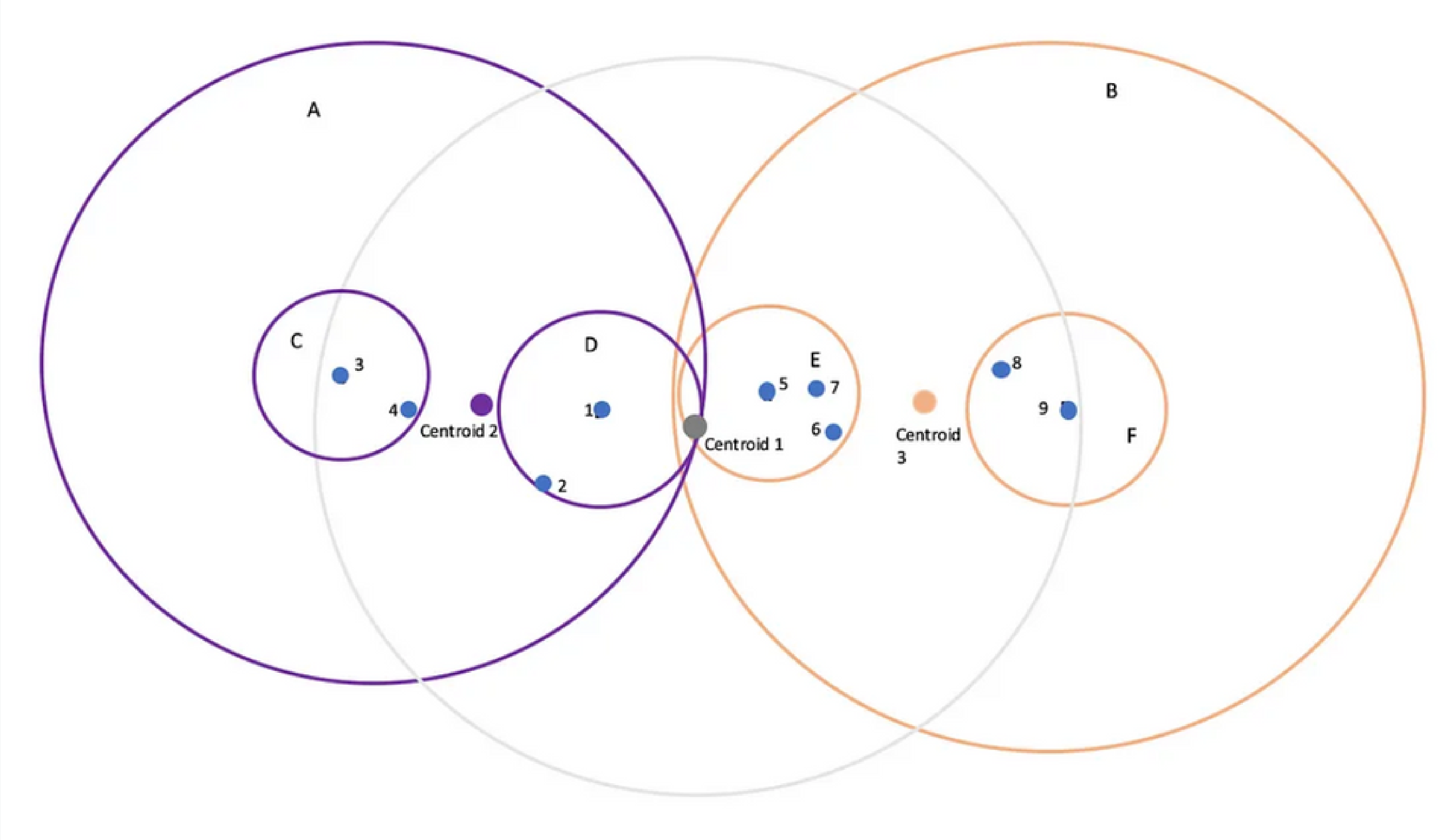
Source: Copernicus EU

DATA PREPARATION

- Spatial Feature Engineering
  - Distances to Important points
  - Percentage Score
  - Spatial Lags/ Standardization

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Moran's I Statistic. Source: Rey, S. J., Arribas-Bel, D., & Wolf, L. J(2020)



Euclidien Distance with ball tree. Source: Hucker, M (2020)

- Data Cleaning (Null variables)

MODELLING

- Supervised Model on Damage level
  - Random Forrest
- SHAP Values to evaluate:

$$SHAP_{feature}(x) = \sum_{set:feature \in set} [|set| \times \binom{F}{|set|}]^{-1} [Predict_{set}(x) - Predict_{set \setminus feature}(x)]$$

Source: Mazzanti, 2020

- Unsupervised Model to explore clusters
  - K-means (4 Clusters)

TOOLS

python

pandas

NumPy

Shapely

Folium

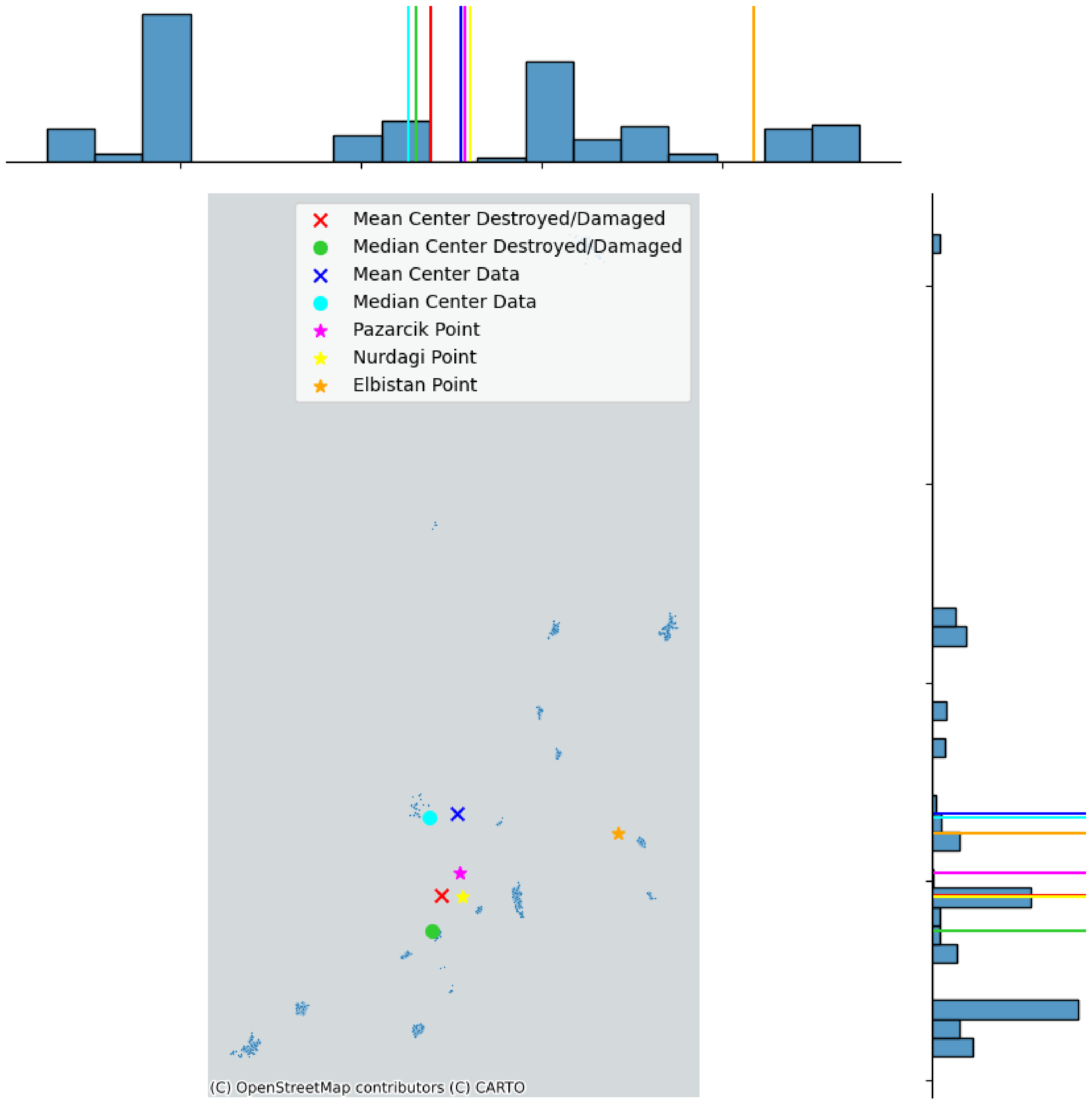
GeoPandas

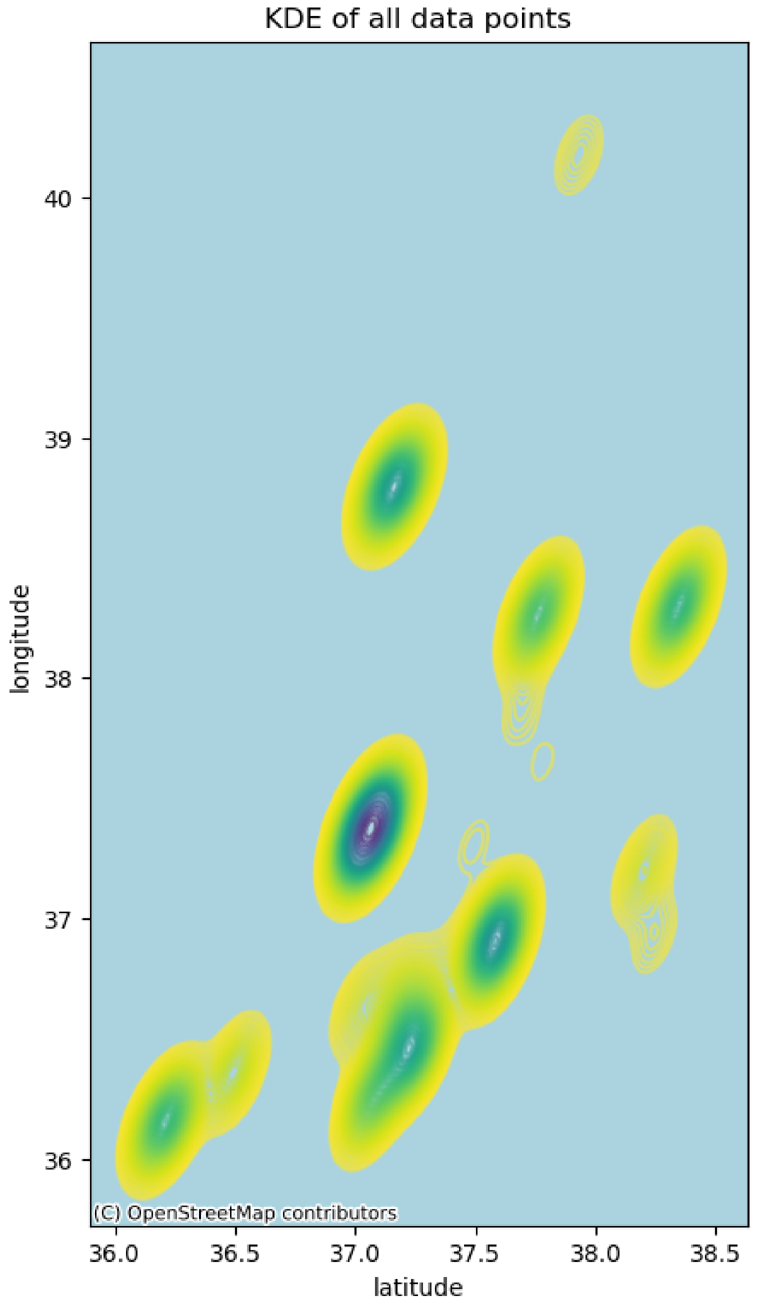
PYCARET

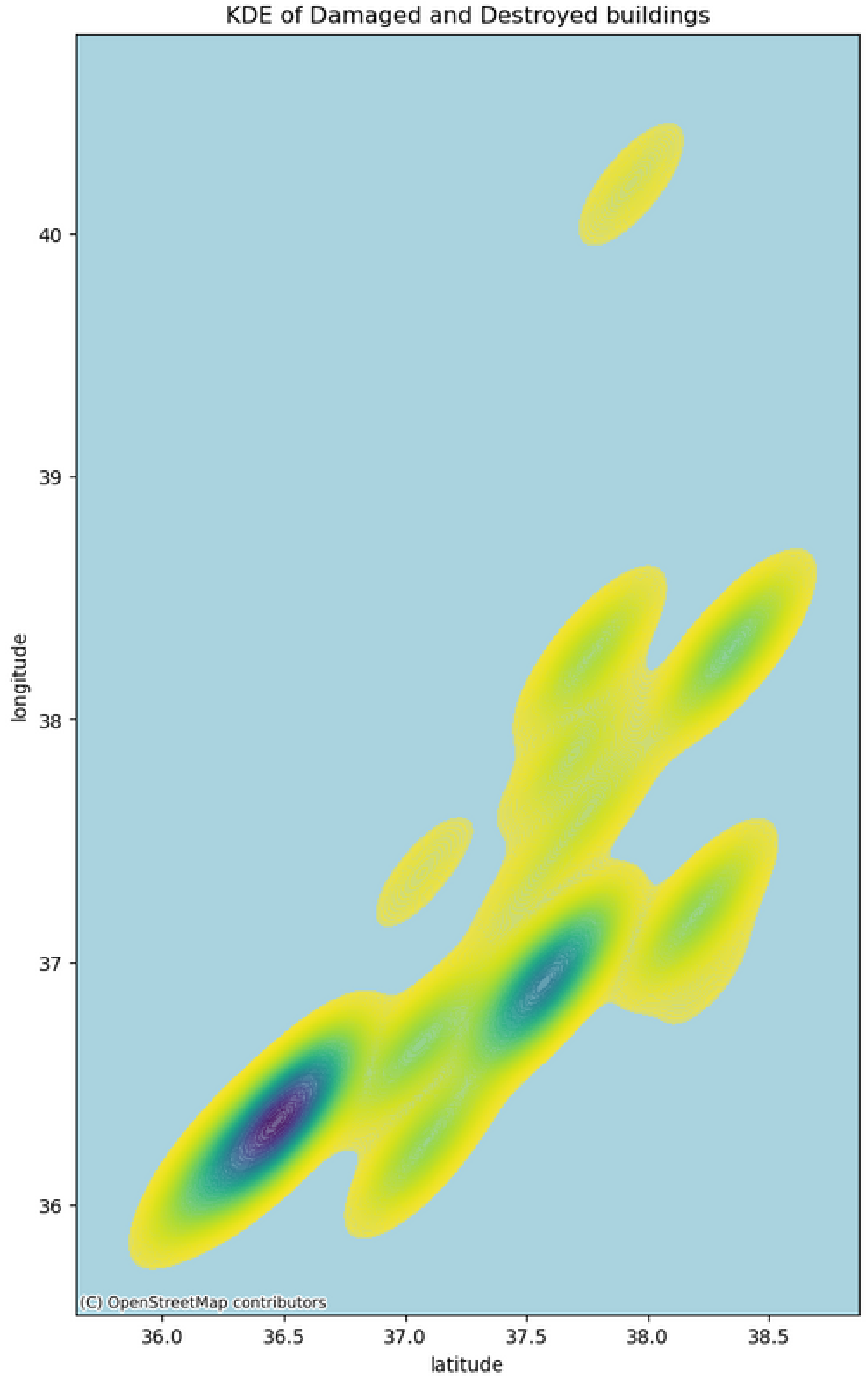
+ableau

SHAP

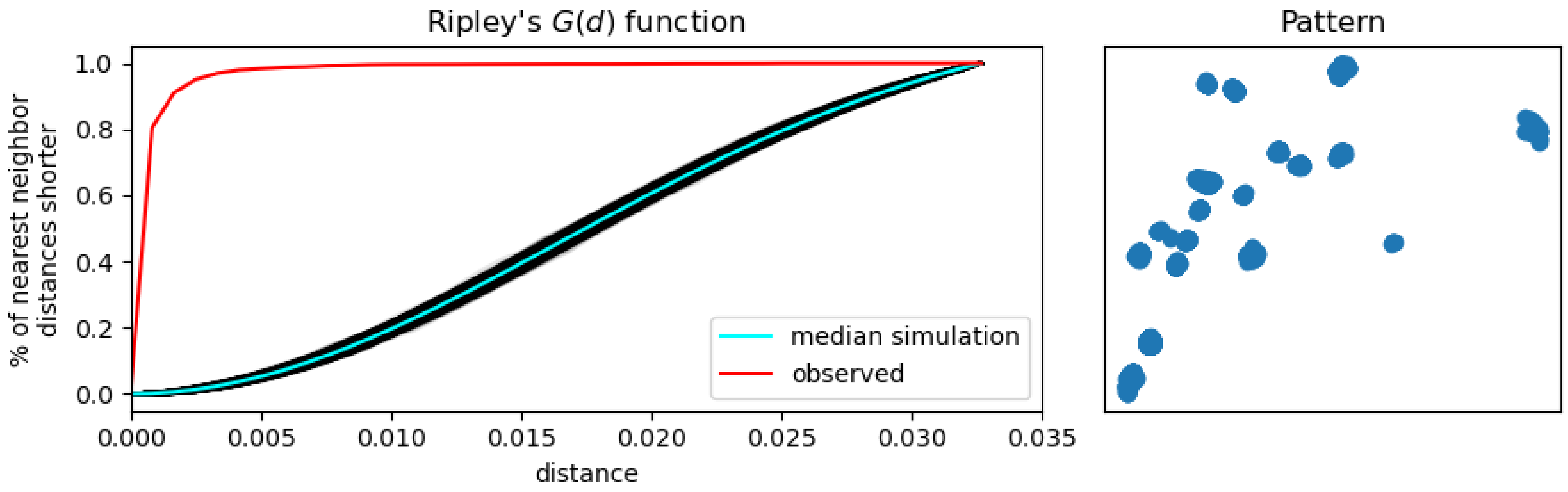
RESULTS -SPATIAL ANALYSIS

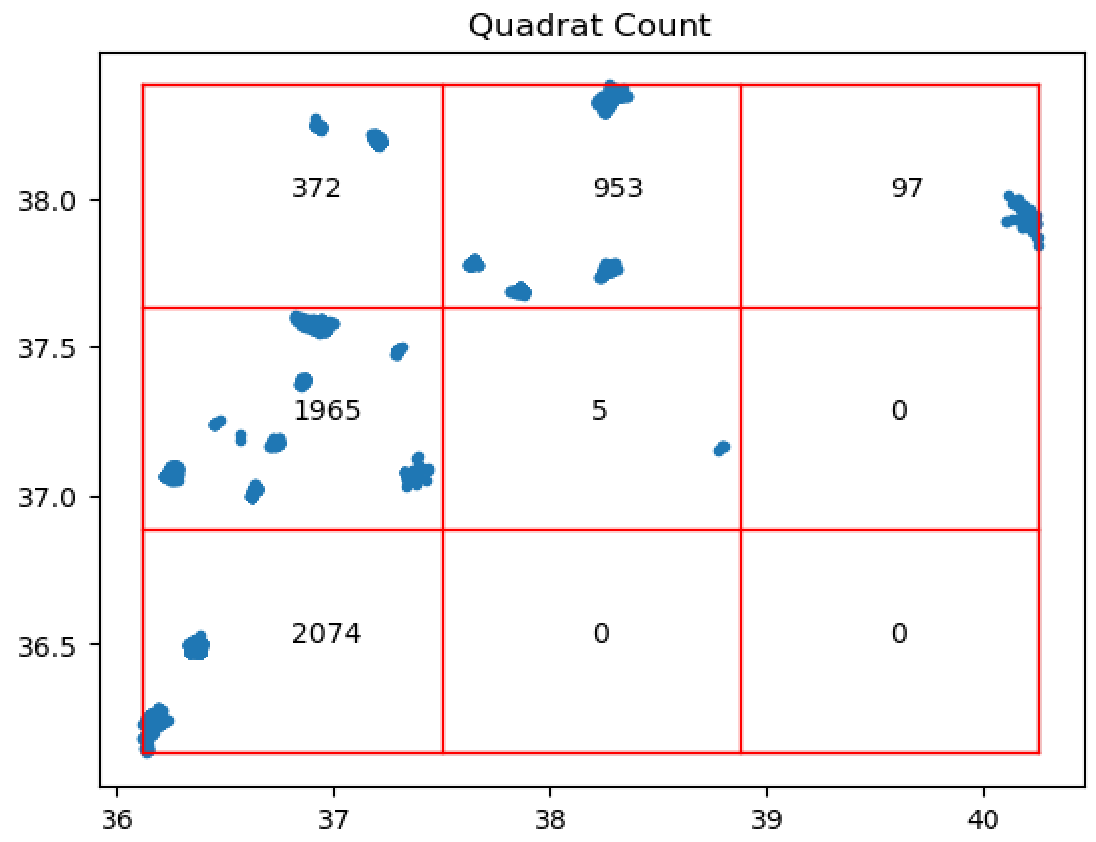


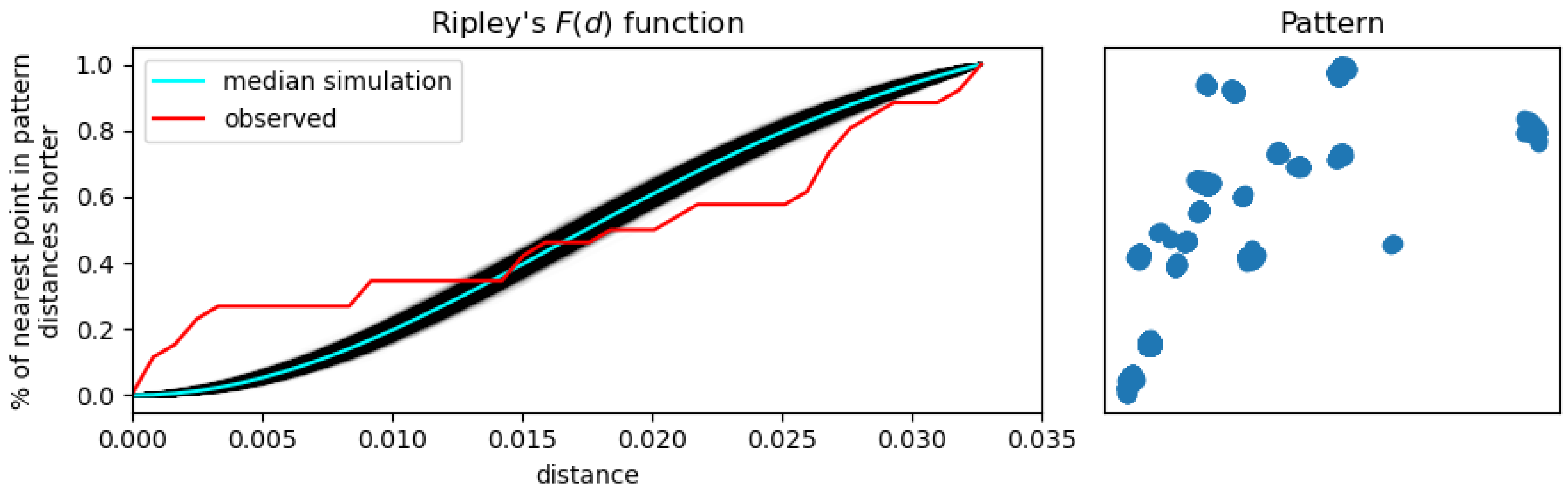




RESULTS -SPATIAL ANALYSIS





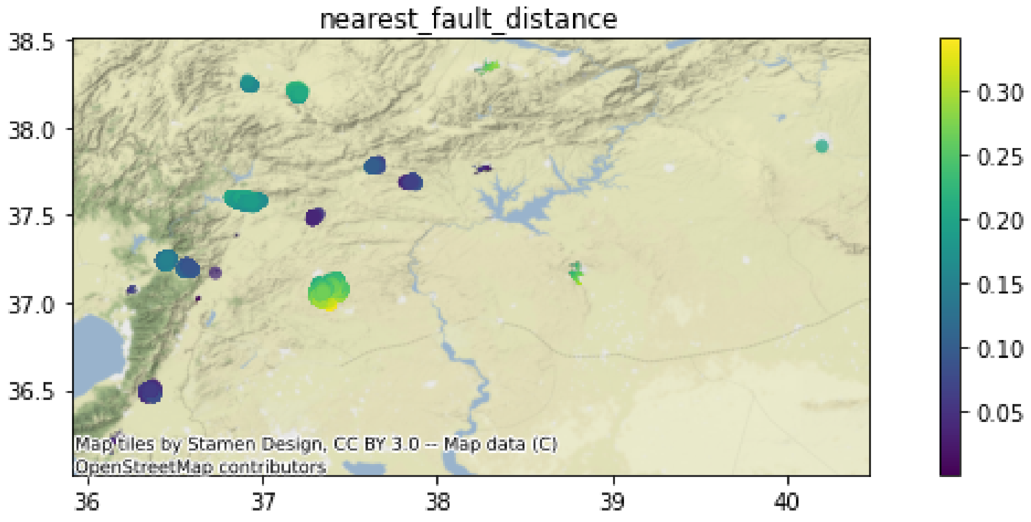


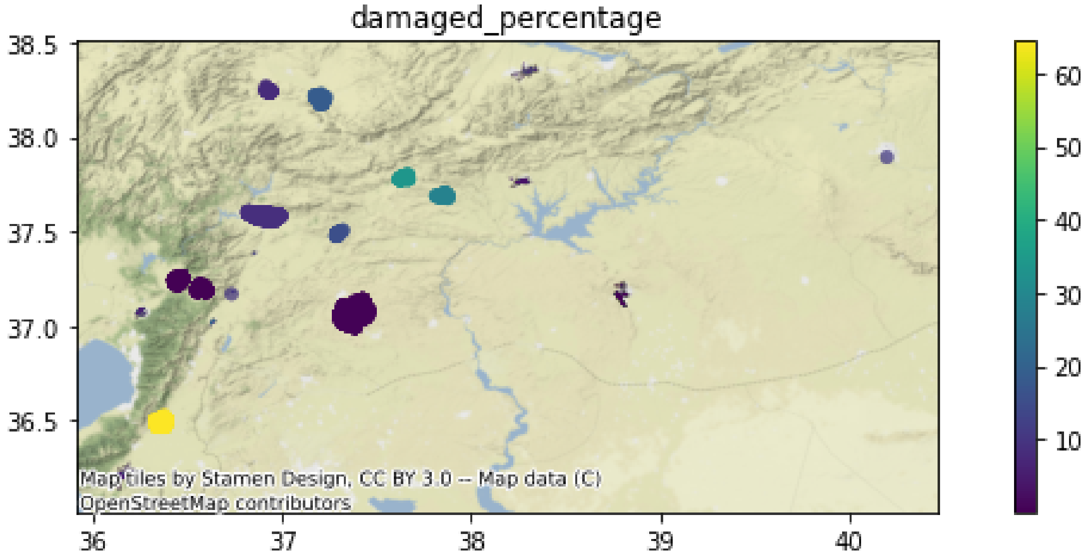

```
#calculating standard spatial distance for buildings

centrography.std_distance(filtered_data[["latitude", "longitude"]])
```

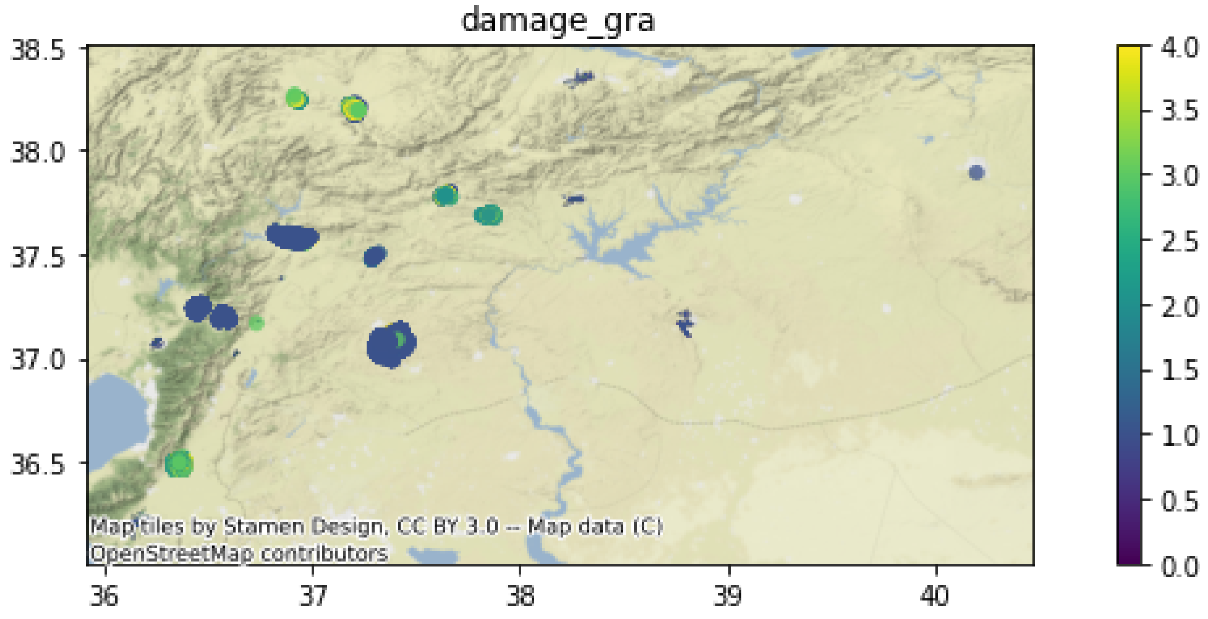
```
1.0462483908354068
```

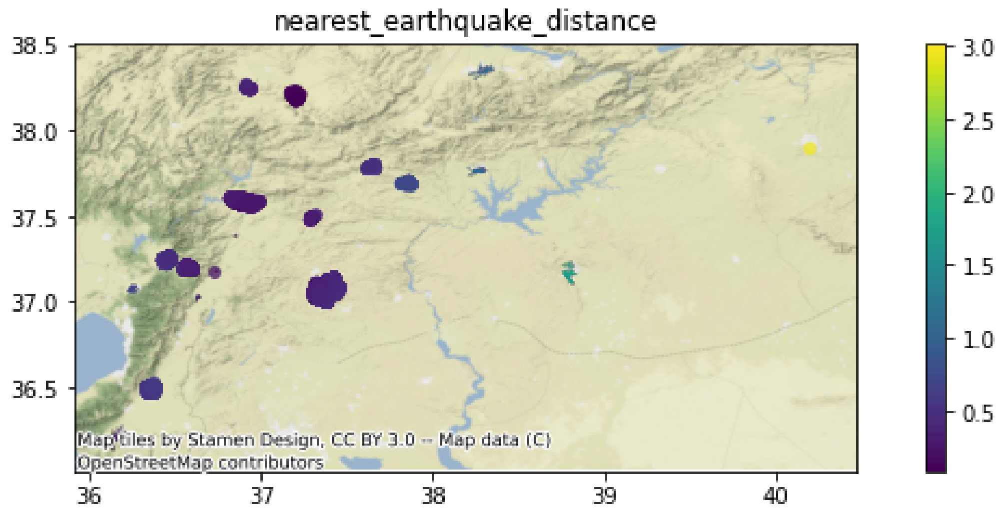
RESULTS -SPATIAL ANALYSIS

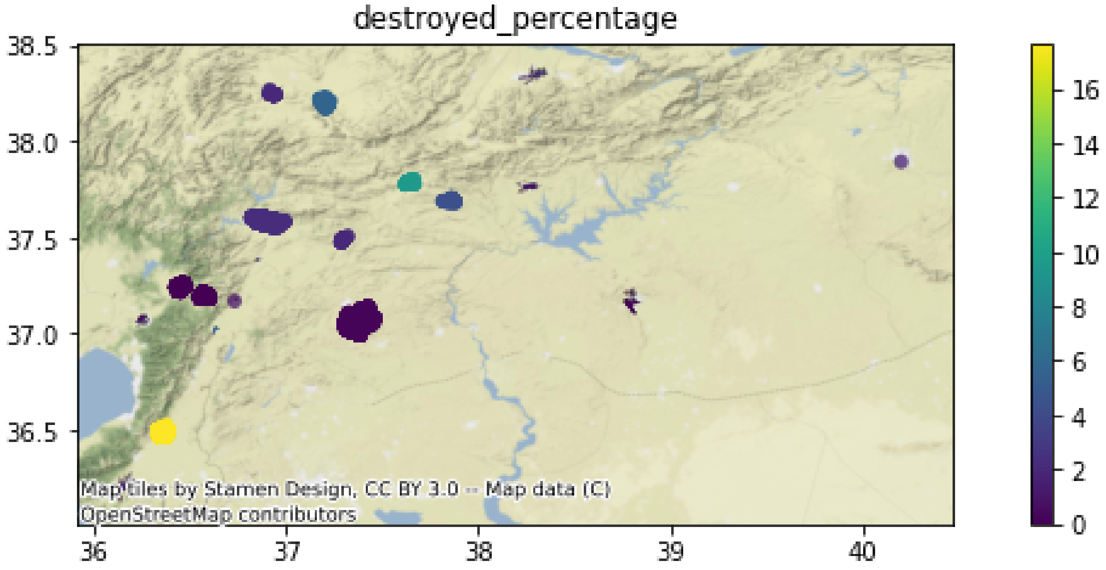
Column	Moran's I	Expected I	p-value
percentage	0.658767	-0.00001	0.001
damaged_percentage	1.000000	-0.00001	0.001
destroyed_percentage	1.000000	-0.00001	0.001
nearest_water_source_distance	0.999846	-0.00001	0.001
nearest_camping_distance	1.000024	-0.00001	0.001
nearest_earthquake_distance	1.000005	-0.00001	0.001
nearest_fault_distance	0.999976	-0.00001	0.001
damage_gra	0.532654	-0.00001	0.001









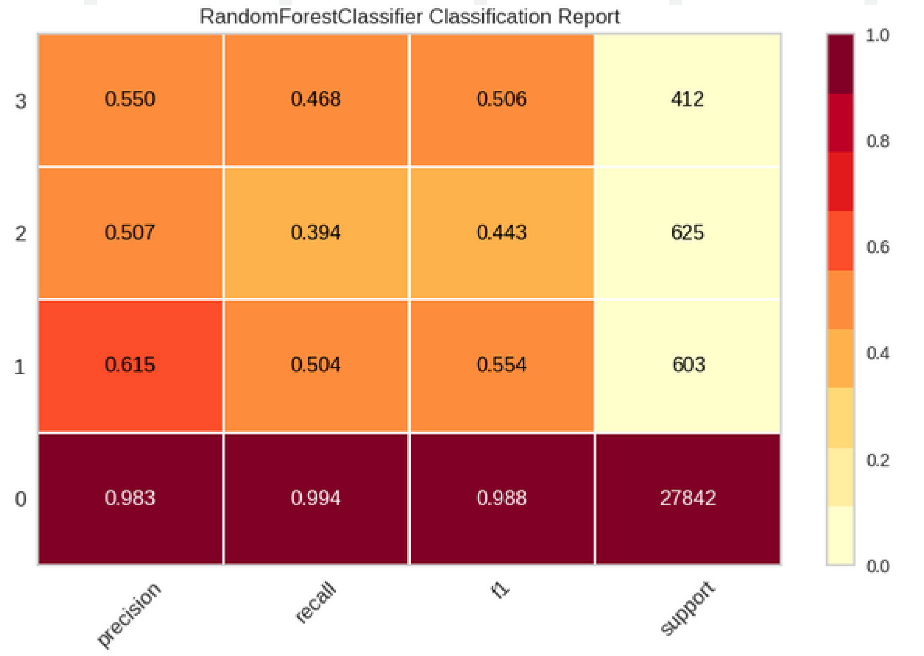


RESULTS - SUPERVISED LEARNING

	Description	Value
0	Session id	123
1	Target	damage_gra
2	Target type	Multiclass
3	Target mapping	1: 0, 2: 1, 3: 2, 4: 3
4	Original data shape	(98272, 49)
5	Transformed data shape	(98272, 10)
6	Transformed train set shape	(68790, 10)
7	Transformed test set shape	(29482, 10)
8	Numeric features	45
9	Categorical features	3
10	Preprocess	True
11	Imputation type	simple
12	Numeric imputation	mean
13	Categorical imputation	mode
14	Maximum one-hot encoding	25
15	Encoding method	None
16	Remove multicollinearity	True
17	Multicollinearity threshold	0.900000
18	Feature selection	True
19	Feature selection method	classic
20	Feature selection estimator	lightgbm
21	Number of features selected	0.200000
22	Fold Generator	StratifiedKFold
23	Fold Number	10
24	CPU Jobs	-1
25	Use GPU	True
26	Log Experiment	False
27	Experiment Name	clf-default-name
28	USI	ecb9

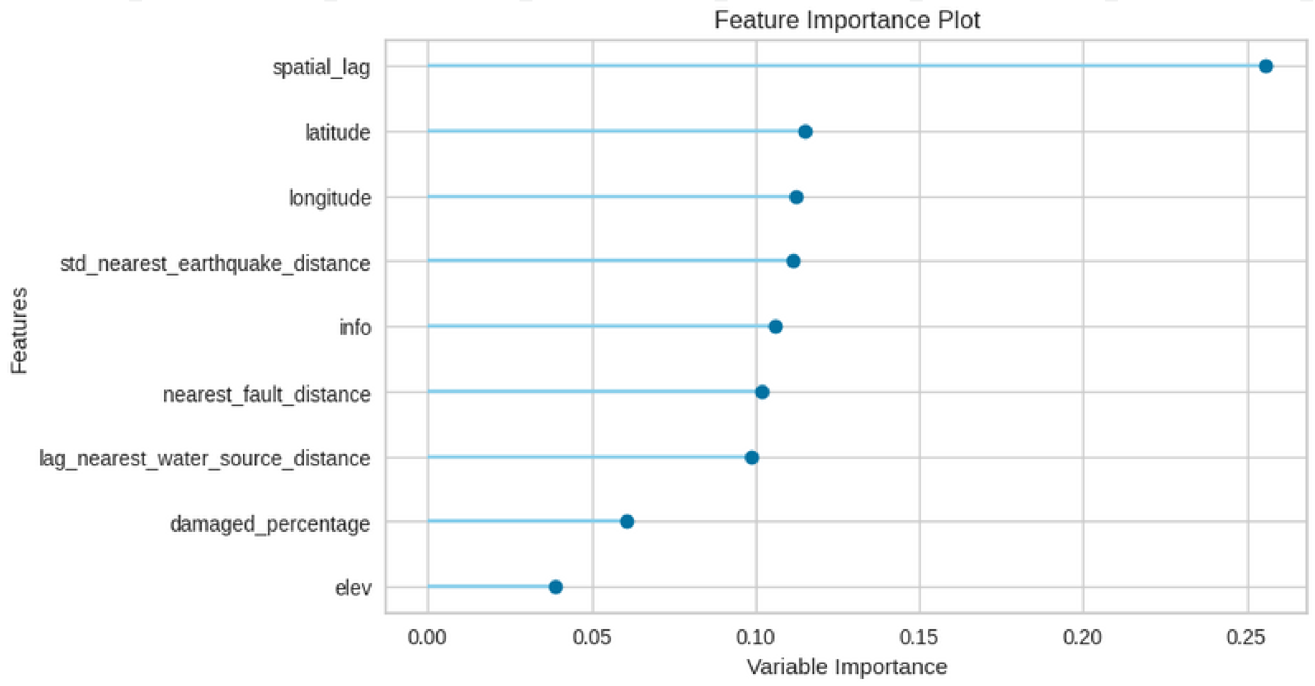
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9636	0.9748	0.9636	0.9587	0.9607	0.6216	0.6263	0.5020
lightgbm	Light Gradient Boosting Machine	0.9621	0.9830	0.9621	0.9553	0.9579	0.5928	0.6009	0.4480
et	Extra Trees Classifier	0.9613	0.9688	0.9613	0.9577	0.9593	0.6116	0.6137	0.5820
gbc	Gradient Boosting Classifier	0.9612	0.9786	0.9612	0.9533	0.9562	0.5736	0.5843	0.4730
ada	Ada Boost Classifier	0.9586	0.9672	0.9586	0.9499	0.9533	0.5463	0.5562	0.3740
knn	K Neighbors Classifier	0.9548	0.9157	0.9548	0.9449	0.9486	0.4929	0.5051	0.5950
dt	Decision Tree Classifier	0.9536	0.8494	0.9536	0.9538	0.9537	0.5671	0.5672	0.3090
lr	Logistic Regression	0.9535	0.9611	0.9535	0.9422	0.9463	0.4645	0.4806	0.4330
ridge	Ridge Classifier	0.9503	0.0000	0.9503	0.9323	0.9345	0.2924	0.3553	0.2690
dummy	Dummy Classifier	0.9444	0.5000	0.9444	0.8919	0.9174	0.0000	0.0000	0.3340
lda	Linear Discriminant Analysis	0.9390	0.9627	0.9390	0.9453	0.9414	0.4791	0.4816	0.3800
qda	Quadratic Discriminant Analysis	0.9210	0.9461	0.9210	0.9457	0.9317	0.4219	0.4343	0.4870
nb	Naive Bayes	0.9182	0.9380	0.9182	0.9436	0.9292	0.4018	0.4139	0.3090
svm	SVM - Linear Kernel	0.8846	0.0000	0.8846	0.9356	0.8903	0.3330	0.3522	0.2420

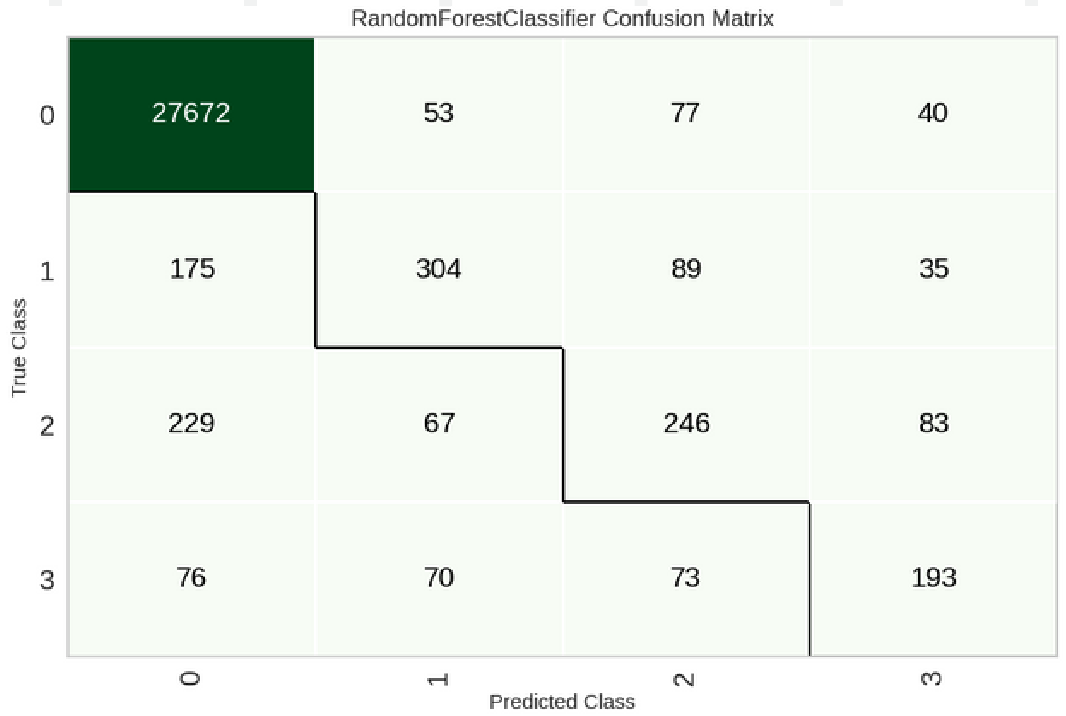
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9594	0.9778	0.9594	0.9505	0.9532	0.5510	0.5623
1	0.9589	0.9727	0.9589	0.9494	0.9527	0.5367	0.5504
2	0.9596	0.9739	0.9596	0.9532	0.9532	0.5499	0.5622
3	0.9602	0.9777	0.9602	0.9537	0.9549	0.5731	0.5806
4	0.9597	0.9722	0.9597	0.9532	0.9535	0.5496	0.5629
5	0.9580	0.9787	0.9580	0.9513	0.9517	0.5363	0.5474
6	0.9597	0.9744	0.9597	0.9511	0.9528	0.5536	0.5657
7	0.9564	0.9714	0.9564	0.9501	0.9499	0.5151	0.5267
8	0.9570	0.9749	0.9570	0.9503	0.9503	0.5172	0.5302
9	0.9581	0.9739	0.9581	0.9521	0.9506	0.5274	0.5419
Mean	0.9587	0.9747	0.9587	0.9515	0.9523	0.5410	0.5530
Std	0.0012	0.0024	0.0012	0.0014	0.0015	0.0170	0.0160

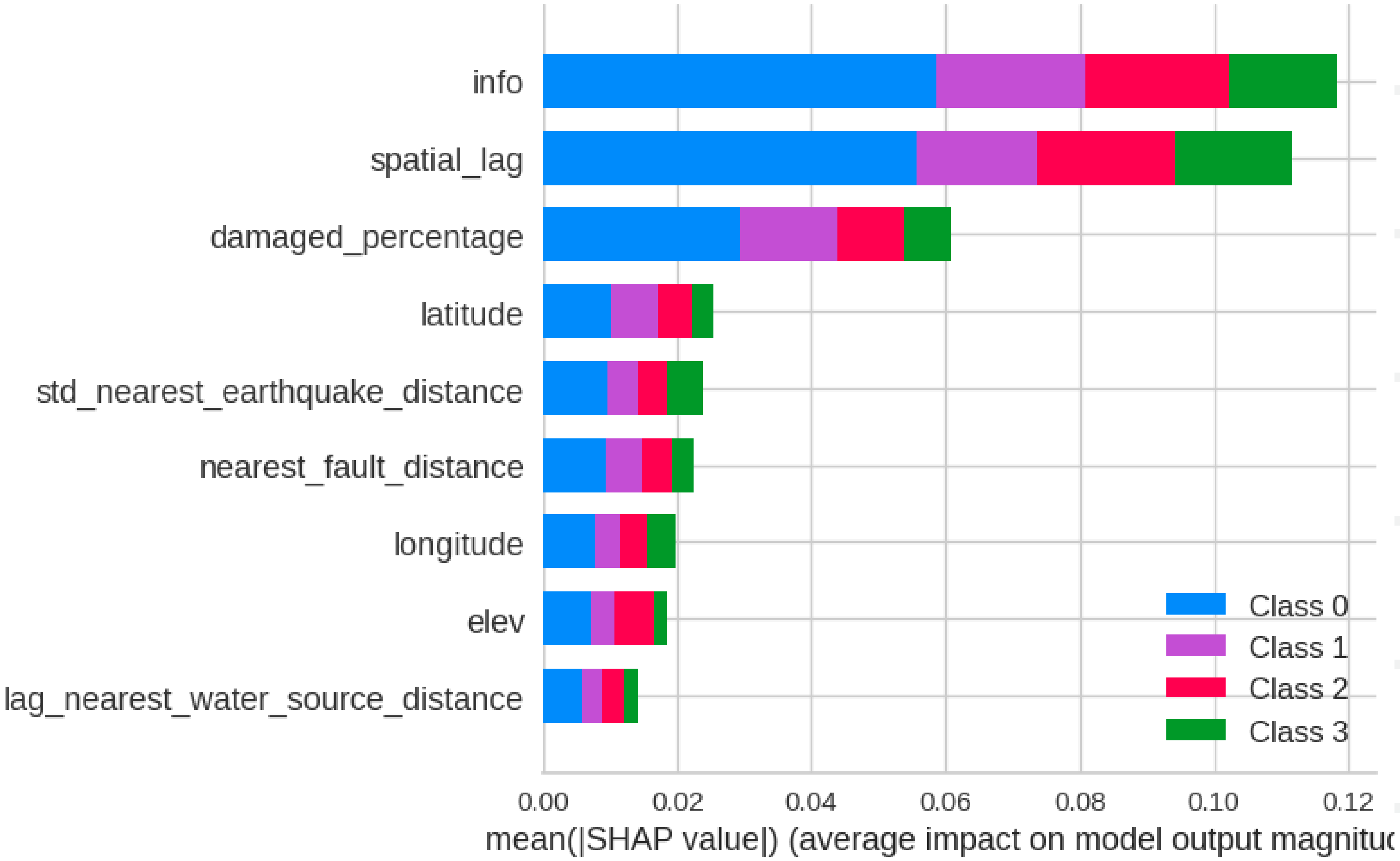


Source: Yazganoglu, G

RESULTS - SUPERVISED LEARNING

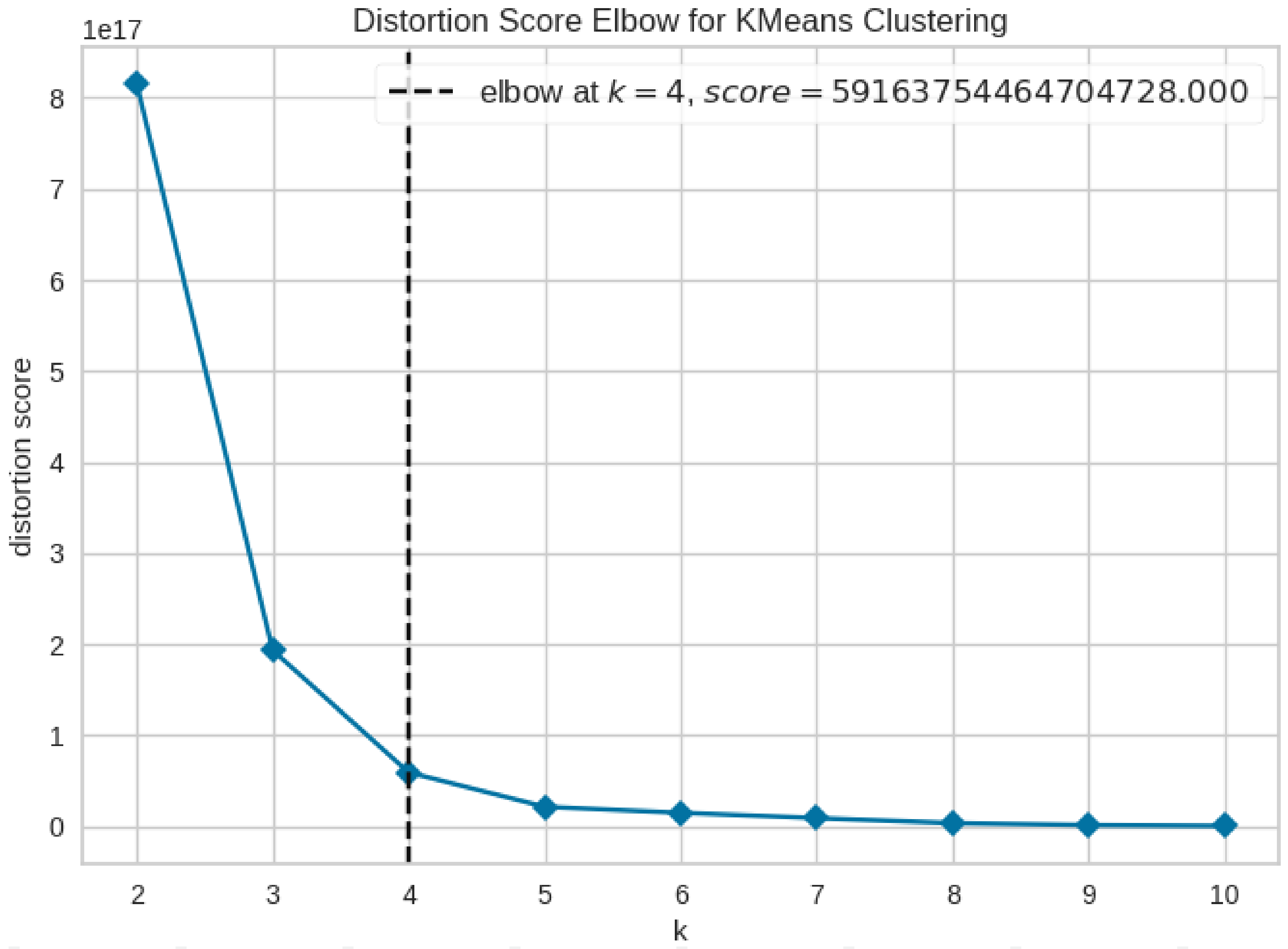


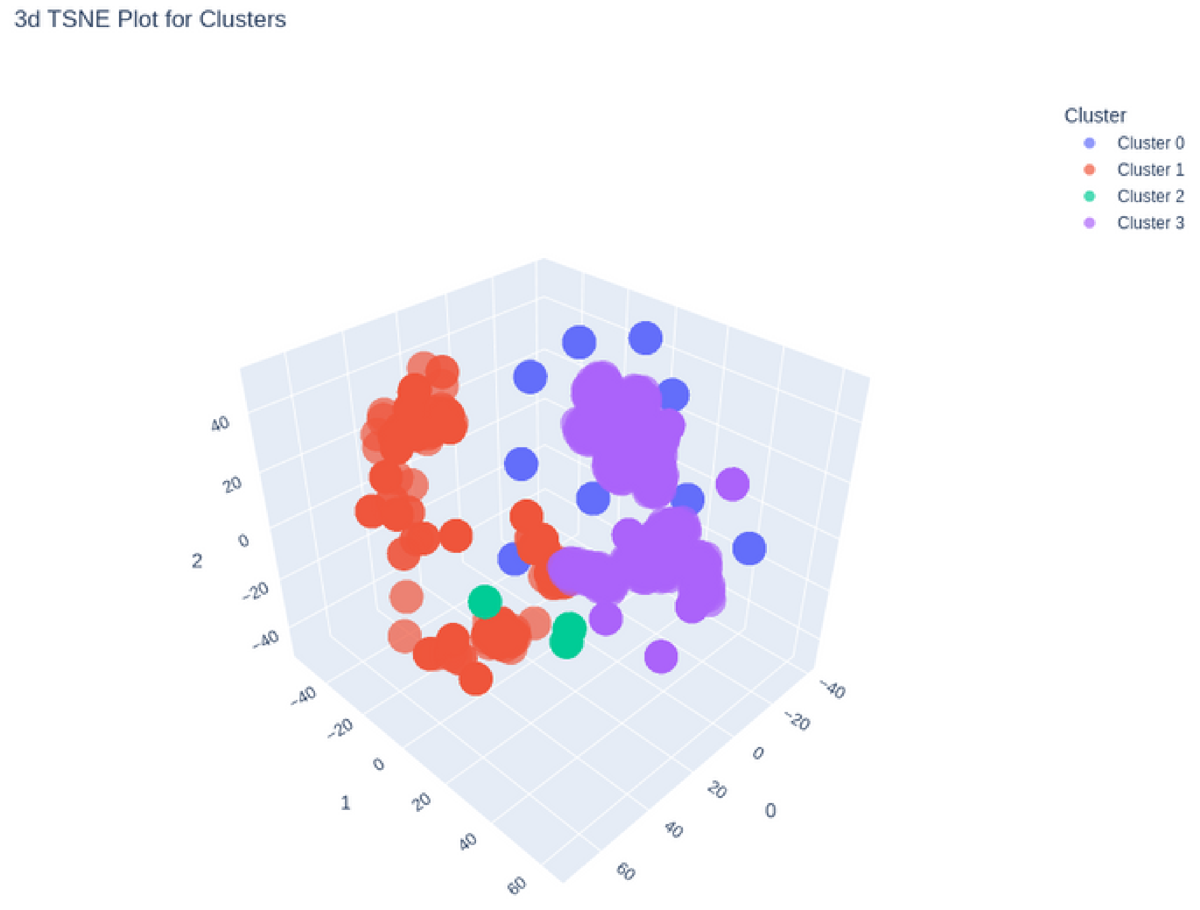




Source: Yazganoglu, G

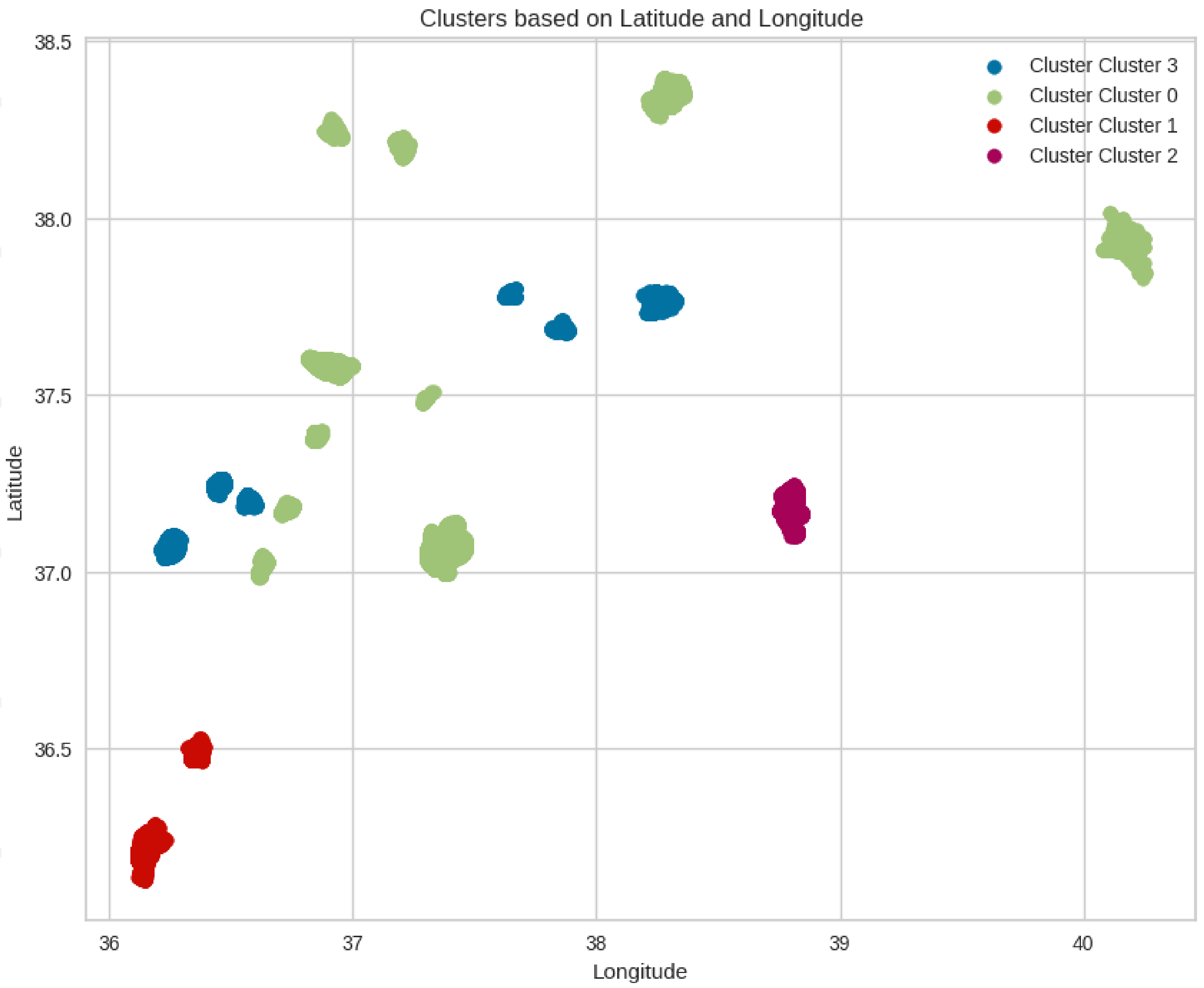
RESULTS - UNSUPERVISED LEARNING





Source: Yazganoglu, G

RESULTS - UNSUPERVISED LEARNING



Source: Yazganoglu, G

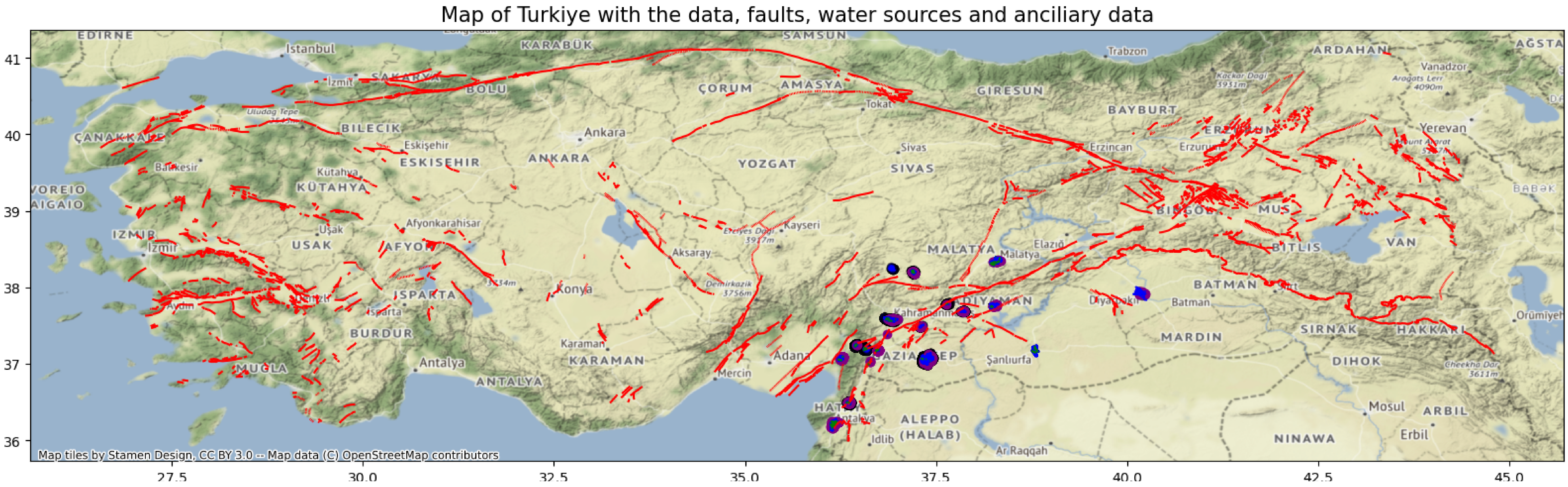
RESEARCH QUESTIONS

- **If there is a relationship between damage and the geophysical fault information?**
  - *Yes, specifically fault distance, earthquake distance and damaged percantage of locality is important variables*
- **How are clustered buildings and different level of damages?**
  - *Buildings that are in the most damaged localities likely to end up damaged or destroyed.*
- **By using a machine learning algorithm can we find what makes an area more vulnerable?**
  - *We have detected more vulnerable clusters but we need more detailed regions.*
- **Which regions are affected worse compared to other regions?**
  - *Cluster 0 and Cluster 1 was affected worse compared to other clusters*

FUTURE WORK

- Multidisiplinary approach is necessary for collection of data.
  - Specifically more info about is needed (year, material, floors...etc)
  - Coloboration of University, Business and Municipality
- Geocoding or smaller local units.
- Temporal movements (Latitude, Longitude, Impact, Depth)
- More actuarial analysis
- Same analysis with sythetic data to determine risky areas for buildings in different regions

Map of Turkiye



BUSINESS THREADS/ OPPORTUNITIES

- Buildings Collapse -> New Durable Buildings are necessary.
- Buildings are Vulnerable - > Existing buildings should be fixed.
- Housing Loans can bankrupt -> Buildings should be insured as part of loan.
- Roads may collapse -> Roads should be fixed, alternatives should be planned.
- Tenants are living in dangerous housing -> Eartquake Certificate
- Supplies can be hard to deliver -> Alternative scenarios should be considered.

THANK YOU

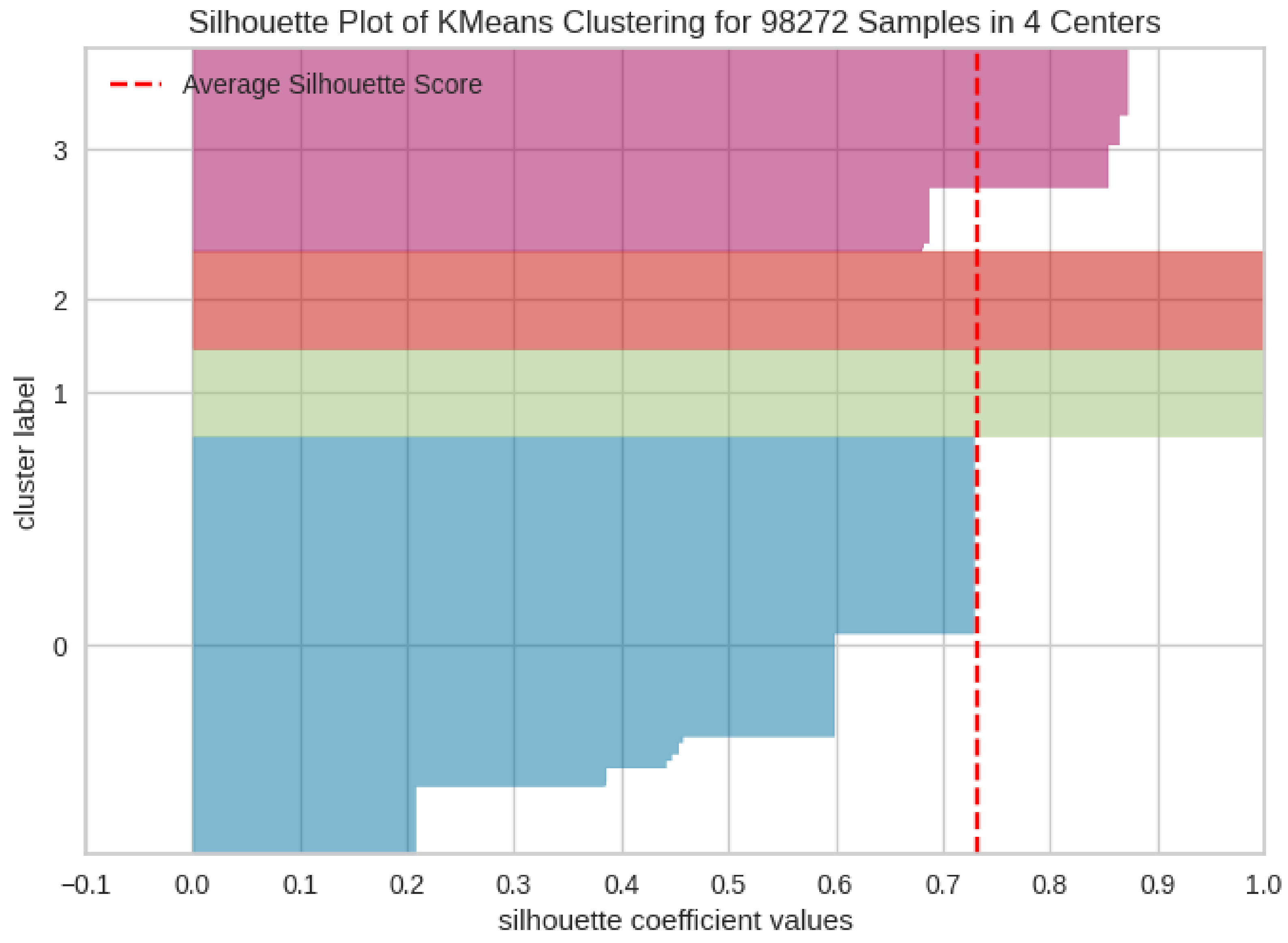
Presentation by Gozde Yazganoglu

CUNEF UNIVERSIDAD

Data Dictionary

0	obj_type	object	0	0.000000	10	NaN	NaN	NaN	NaN
1	name	object	10	0.009703	5082	NaN	NaN	NaN	NaN
2	info	object	0	0.000000	52	NaN	NaN	NaN	NaN
3	damage_gra	object	0	0.000000	5	NaN	NaN	NaN	NaN
4	det_method	object	0	0.000000	2	NaN	NaN	NaN	NaN
5	notation	object	0	0.000000	3	NaN	NaN	NaN	NaN
6	or_src_id	int64	0	0.000000	6	1.000000e+00	9.970000e+02	6.042655e+02	4.848318e+02
7	dmg_src_id	int64	0	0.000000	4	2.000000e+00	9.970000e+02	1.317369e+01	1.019061e+02
8	cd_value	object	0	0.000000	1	NaN	NaN	NaN	NaN
9	real	object	78802	76.460029	2	NaN	NaN	NaN	NaN
10	index_right	int64	0	0.000000	1	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
11	emsr_id	object	0	0.000000	1	NaN	NaN	NaN	NaN
12	glide_no	object	0	0.000000	2	NaN	NaN	NaN	NaN
13	area_id	object	0	0.000000	19	NaN	NaN	NaN	NaN
14	locality	object	0	0.000000	19	NaN	NaN	NaN	NaN
15	map_type	object	0	0.000000	2	NaN	NaN	NaN	NaN
16	population	int64	0	0.000000	17	2.290400e+04	2.170110e+06	1.030682e+06	8.724357e+05
17	income	int64	0	0.000000	7	3.012000e+03	7.819000e+03	6.185708e+03	1.627163e+03
18	total_sales	int64	0	0.000000	19	1.588000e+03	2.481210e+05	1.083212e+05	9.221601e+04
19	second_sales	int64	0	0.000000	19	5.360000e+02	1.414340e+05	6.071037e+04	5.280406e+04
20	water_access	float64	0	0.000000	4	9.500000e-01	1.000000e+00	9.746440e-01	2.280118e-02
21	elec_cons	int64	0	0.000000	7	1.631000e+03	7.413000e+03	3.972904e+03	1.827892e+03
22	building_perm	int64	0	0.000000	8	5.830000e+02	2.959000e+03	1.660737e+03	8.814879e+02
23	land_permited	int64	0	0.000000	8	6.957180e+05	3.019546e+06	1.866989e+06	8.794762e+05
24	labour_fource	float64	0	0.000000	4	4.060000e+01	5.000000e+01	4.788620e+01	2.885233e+00
25	unemployment	float64	0	0.000000	4	1.010000e+01	1.710000e+01	1.368046e+01	3.302016e+00
26	agricultural	int64	0	0.000000	8	1.233061e+06	3.535085e+08	4.008357e+07	1.066765e+08
27	life_time	float64	0	0.000000	8	7.690000e+01	7.970000e+01	7.811588e+01	9.676588e-01
28	hb_per100000	int64	0	0.000000	8	1.930000e+02	3.690000e+02	2.669623e+02	4.670803e+01
29	fertility	float64	0	0.000000	8	1.630000e+00	3.810000e+00	2.376413e+00	5.862124e-01
30	hh_size	float64	0	0.000000	9	3.400000e+00	5.120000e+00	3.902629e+00	4.935363e-01
31	longitude	float64	0	0.000000	98600	3.611427e+01	4.025491e+01	3.735722e+01	8.896378e-01
32	latitude	float64	0	0.000000	98567	3.612831e+01	3.839672e+01	3.729042e+01	5.234444e-01
33	nearest_water_source_distance	float64	0	0.000000	98670	2.067003e-06	3.714001e-01	2.354838e-02	3.965855e-02
34	nearest_camping_distance	float64	0	0.000000	98670	5.749022e-07	8.413472e-01	1.000509e-01	2.467072e-01
35	nearest_earthquake_distance	float64	0	0.000000	98670	8.891392e-02	3.025931e+00	6.889907e-01	5.302256e-01
36	nearest_fault_distance	float64	0	0.000000	98670	1.675678e-04	3.405915e-01	1.703268e-01	9.564513e-02
37	elev	float64	0	0.000000	123	7.000000e+01	1.270000e+03	6.264720e+02	2.805116e+02
38	geometry	geometry	0	0.000000	98671	NaN	NaN	NaN	NaN

Silhouette Plot for K-Means

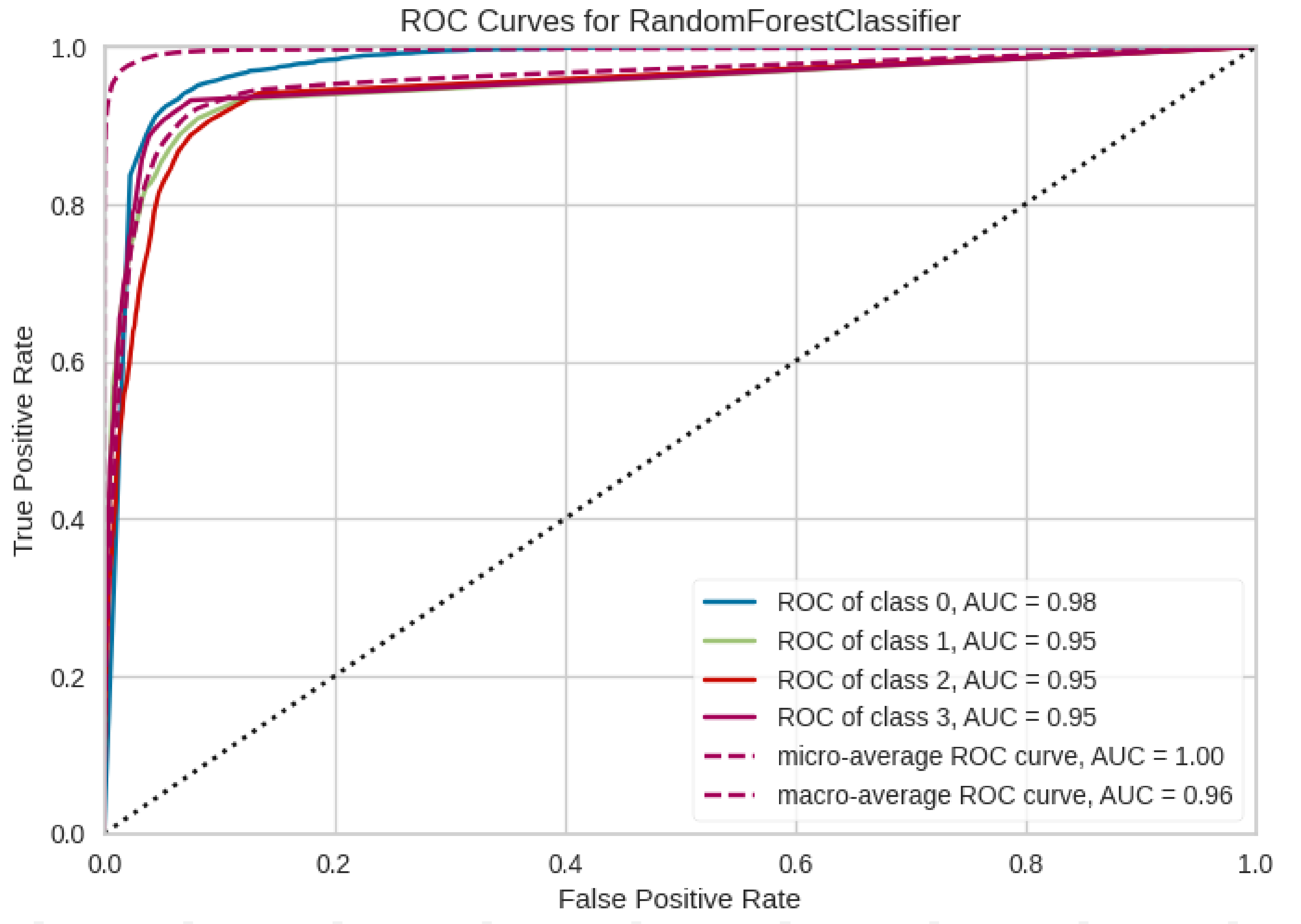


	Description	Value
0	Session id	6993
1	Original data shape	(98272, 49)
2	Transformed data shape	(98272, 120)
3	Numeric features	46
4	Categorical features	3
5	Preprocess	True
6	Imputation type	simple
7	Numeric imputation	mean
8	Categorical imputation	mode
9	Maximum one-hot encoding	-1
10	Encoding method	None
11	CPU Jobs	-1
12	Use GPU	False
13	Log Experiment	False
14	Experiment Name	cluster-default-name
15	USI	8f7c

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.7315	645515079.2386	0.2921	0	0	0

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.4529	82.9497	1.4016	0	0	0

RESULT



ROC Curves for RandomForestClassifier
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.98
ROC of class 1, AUC = 0.95
ROC of class 2, AUC = 0.95
ROC of class 3, AUC = 0.95
micro-average ROC curve, AUC = 1.00
macro-average ROC curve, AUC = 0.96