Analyzing Radiation Intensity in Communication Sites

Author: Eyal Kadosh (208005280)

This project explores the dataset from the Israeli Government Data Portal (data.gov.il) which contains data about radiation intensity on different site types. The first part compares the performances of five classification models while the second part clusters the data using unsupervised learning algorithms.

Supervised Pipeline

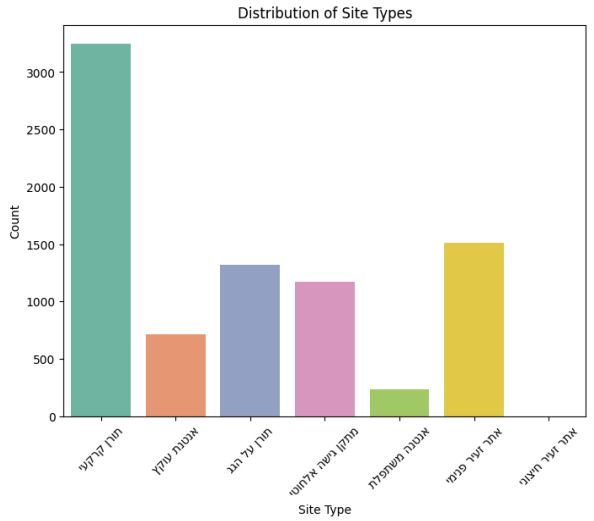
Data Preprocessing

The first step was to explore that data and prepare it for training. Because the data is in Hebrew I translated the columns to English for simplicity and looked at the columns. I choose a few categorial and a few numerical columns.

After the feature selection I checked for null values in each column and there aren't any:

```
NaN count per column:
site_type                                 0
city                                      0
X_ITM                                     0
Y_ITM                                     0
transmission_technologies                 0
Maximum_theoretical_radiation_intensity   0
related_intensity                         0
radiation_approval_status                 0
dtype: int64
```

My goal target will be to classify the site_type according to the other features. In order to get a since of the data distribution I plotted the histogram of the target column:



One can see that 'אתר זעיר חיצוני' has the smallest amount of samples (only 2).

Then I splitted the feature into numerical features and categorical features. I used the `StandardScaler()` for numerical features and `OneHotEncoder()` to encode the categorical features.

Models

I splitted the data into train set and test set, trained five different classification models and used grid search for hyperparameters optimization.
The compared models are:

1. Decision Tree
2. Random Forest
3. Logistic regression
4. Support Vector Machine
5. K-nearest neighbors

Results

For each one of the models I printed the precision, recall, f1-score and support e.g.

```
Model: Decision Tree
Best Parameters: {'classifier__max_depth': 10,
'classifier__min_samples_split': 2}
accuracy 0.67
```

	precision	recall	f1-score	support
אנטנה משתפלת	0.08	0.02	0.04	41
אנטנת עוקץ	0.12	0.02	0.04	140
אתר זעיר חיצוני	0	0	0	1
אתר זעיר פנימי	0.74	0.84	0.78	296
מתקן גישה אלחוטי	0.53	0.58	0.56	239
תורן על הגג	0.45	0.55	0.5	245
תורן קרקעי	0.8	0.84	0.82	680

macro avg	0.39	0.41	0.39	1642
weighted avg	0.62	0.67	0.64	1642

We can learn from this example that the minority class performs poorly due to under-fitting while classes with more data points have better results. We can see the big difference in the macro average results compared to the weighted average results. The precision and recall results are quite similar.

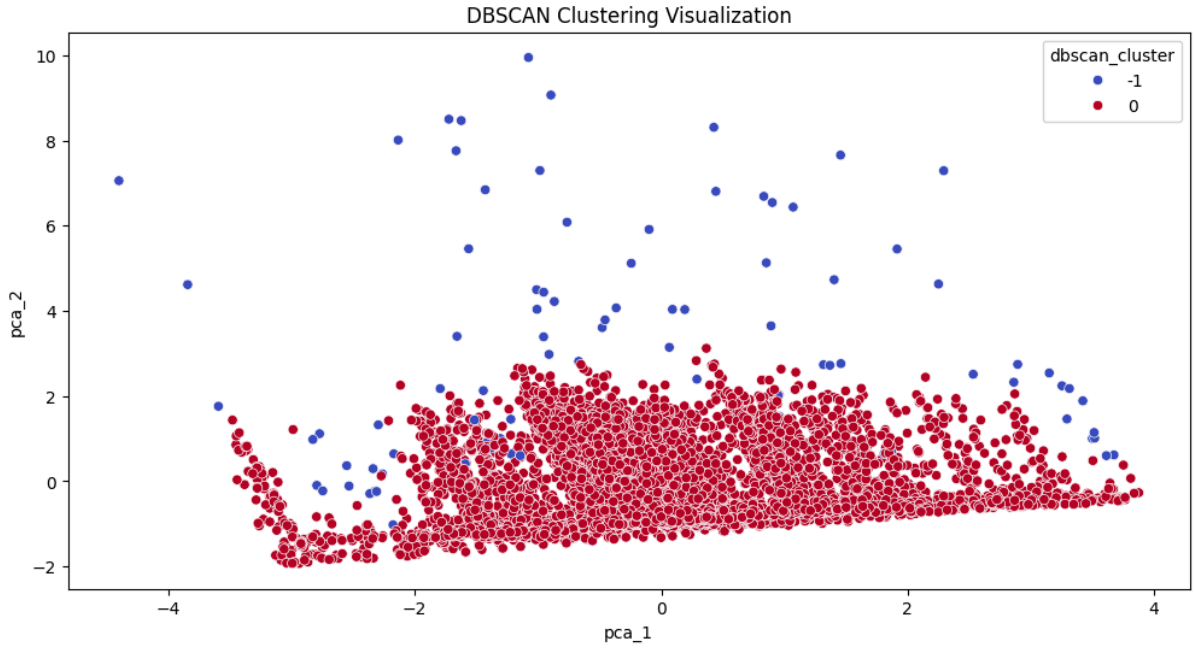
Let's compare the weighted average for the rest of the models.

Model	precision	recall	f1-score
Decision Tree	0.62	0.67	0.64
Random Forest	**0.65**	**0.69**	**0.67**
Logistic regression	0.64	0.68	0.65
Support Vector Machine	0.65	0.68	0.65
K-nearest neighbors	0.63	0.65	0.64

Random forest has the best performance for this dataset.

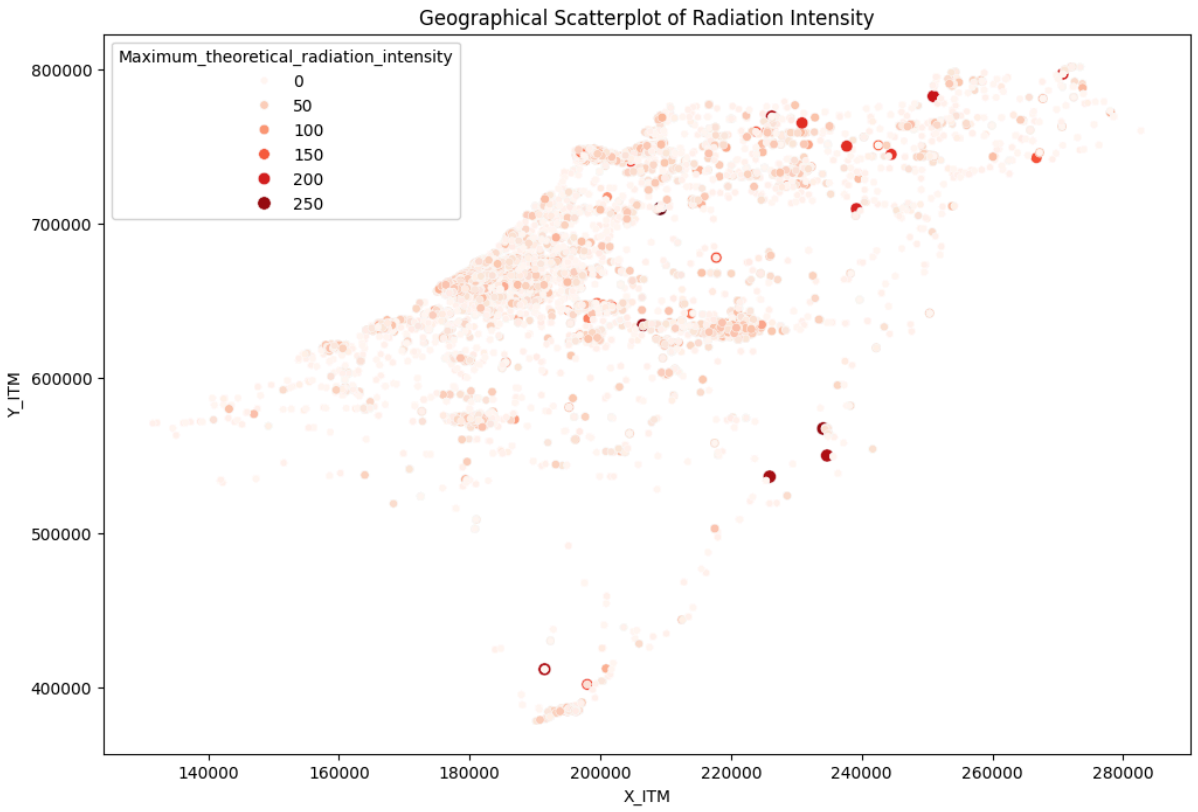
Clustering using unsupervised learning algorithms

In the second part I used two different learning algorithms K-means and DBSCAN.

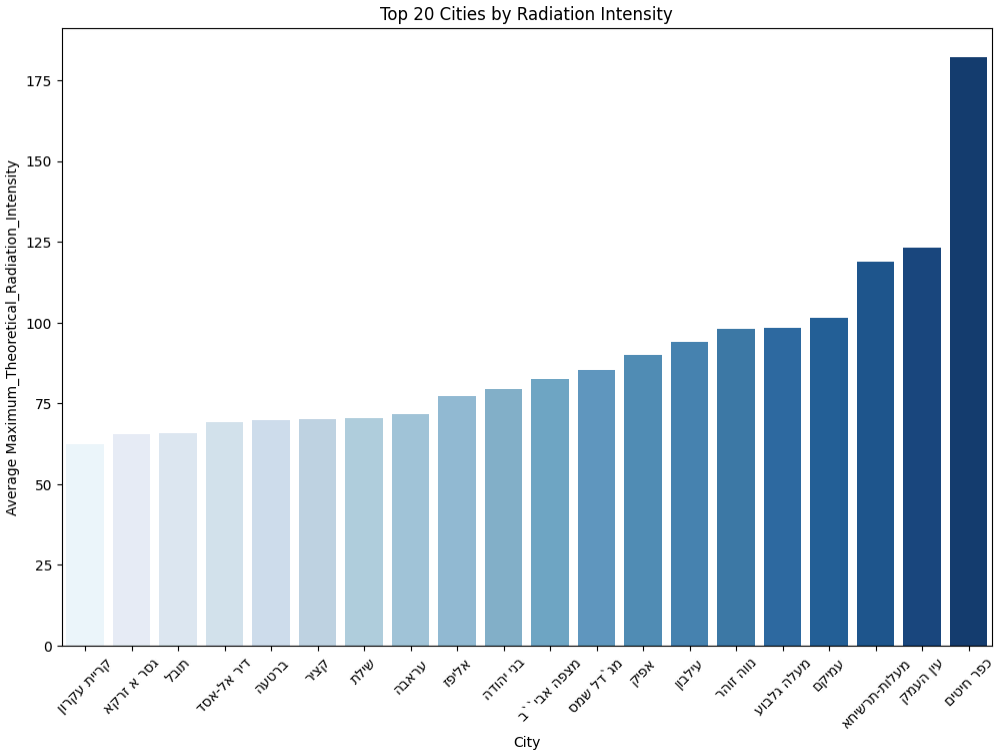


The data point does not have a very natural structure to be splitted by. Most of the data points are in a big cluster in the middle and the rest are floating in the area.

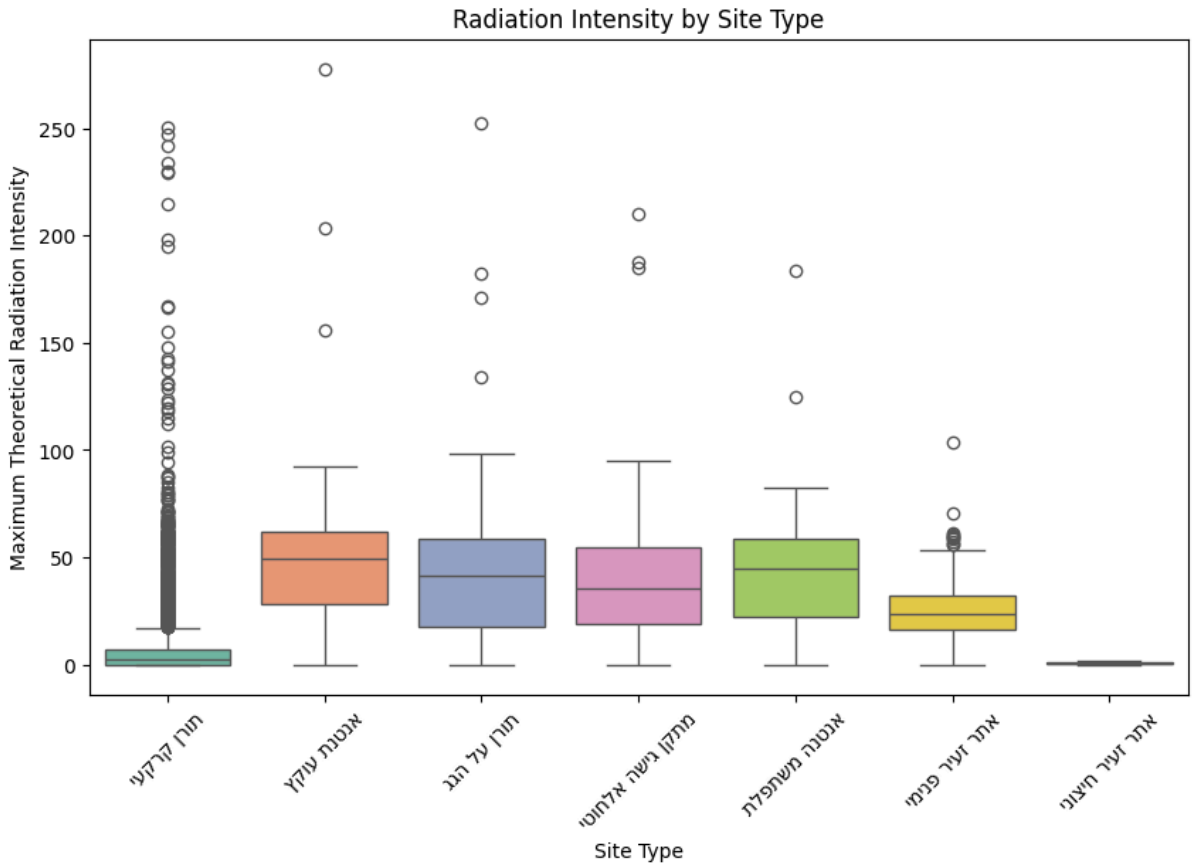
In this part I also plotted the radiation distribution according to the X and Y locations. It's interesting to see that there are some radiating sources and a big part in the middle of the image has an equal radiation intensity.



As part of the analysis I also wanted to look for most radiating cities. So I looked for the top 20.



I also tried to look for a correlation between the site type and the site intensity.



From this image we can learn that the ‘אנטנת עוקץ’ is the most radiating.