A Data Symphony of Image Segmentation

"Data Analysis on Image Segmentation Data"

Syed Muhammad Taqi Raza Kazmi
Aamir Khan
Irfan Mahmood

Università degli Studi dell'Aquila

July 4, 2023

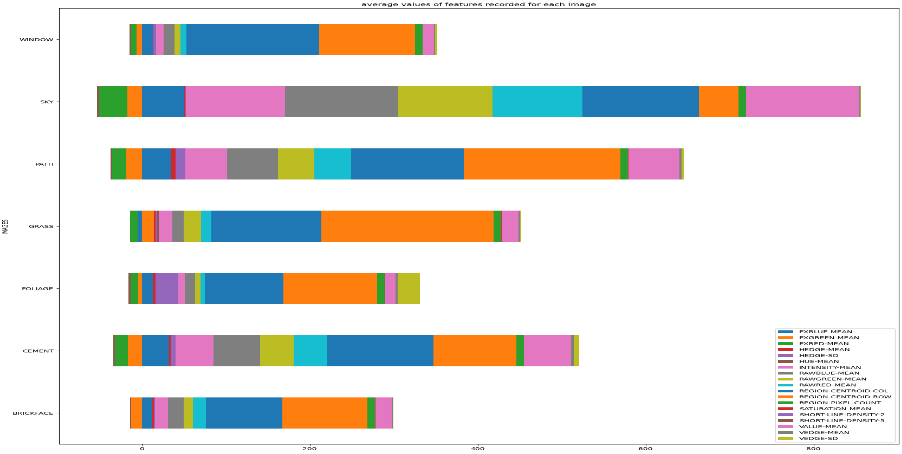
UNIVERSITÀ
DEGLI STUDI
DELL'AQUILA

Contents

1. Introduction
2. Data Cleaning
3. Analyzing relationships and patterns
4. Main Analysis
5. Conclusion

Data Structure and Visualization

The image segmentation data is the combination of two data sets. One data set consists of 210 rows and 20 columns. Another data set consists of 2100 rows and 20 columns. The plot below show the average values of features recorded for each Image:

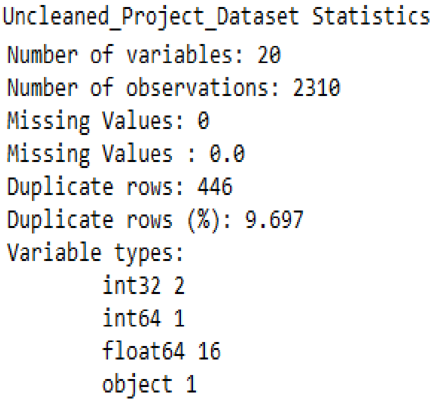


Aim and objective

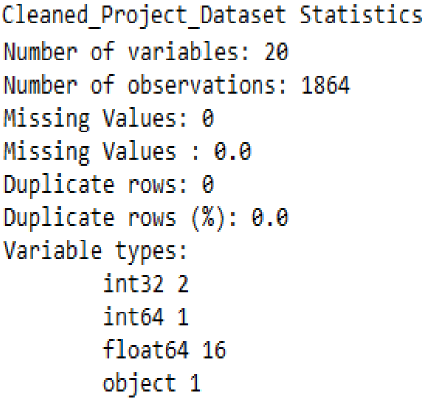
The image segmentation data is a complex and large data set. The instances were drawn randomly from a database of 7 outdoor images. The images were hand segmented to create a classification for every pixel. Each instance is a 3x3 region. The seven outdoor images are brickface, sky, foliage, cement, window, path, grass. In this project, we applied first implement K-Means clustering to find intrinsic groups within this dataset, getting it misfit, we applied an ML algorithm (SVC) to train an SVM classification model with default options for hyperparameters and computed different classification scores for the training and testing set. Finally, we applied the ensemble algorithm GradientBoostingClassifier also and compared the accuracies.

Data Cleaning

The data frame has no missing values and we observed 224 duplicate rows, which we considered as unhealthy to the dataset so they were removed.
The data statistics are shown below:

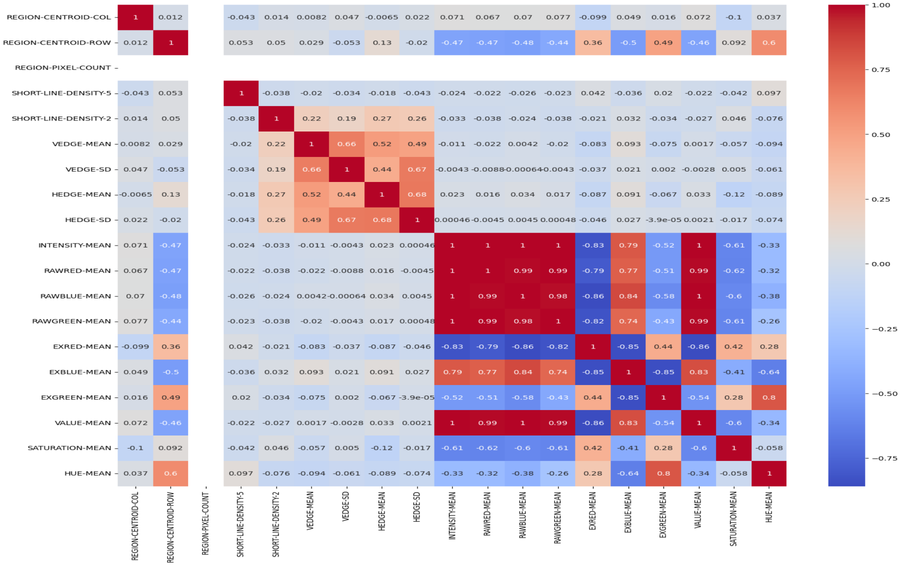


```
Uncleaned_Project_Dataset Statistics
Number of variables: 20
Number of observations: 2310
Missing Values: 0
Missing Values : 0.0
Duplicate rows: 446
Duplicate rows (%): 9.697
Variable types:
        int32 2
        int64 1
        float64 16
        object 1
```



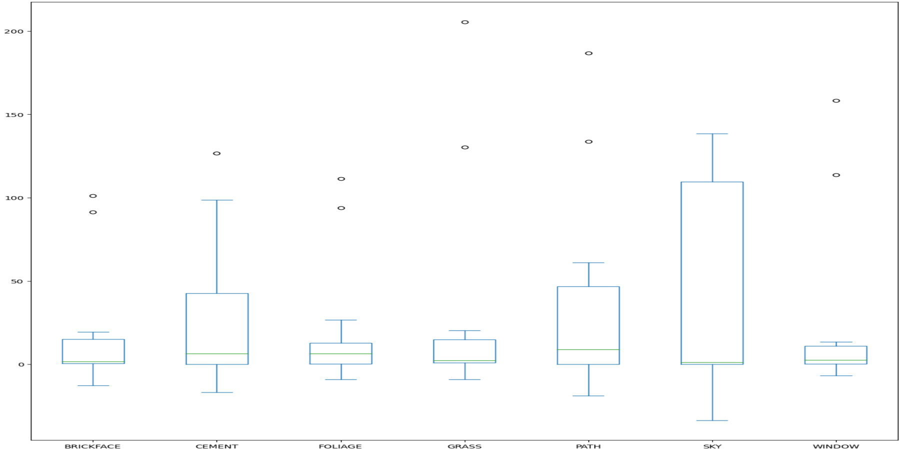
```
Cleaned_Project_Dataset Statistics
Number of variables: 20
Number of observations: 1864
Missing Values: 0
Missing Values : 0.0
Duplicate rows: 0
Duplicate rows (%): 0.0
Variable types:
        int32 2
        int64 1
        float64 16
        object 1
```

Correlation matrix



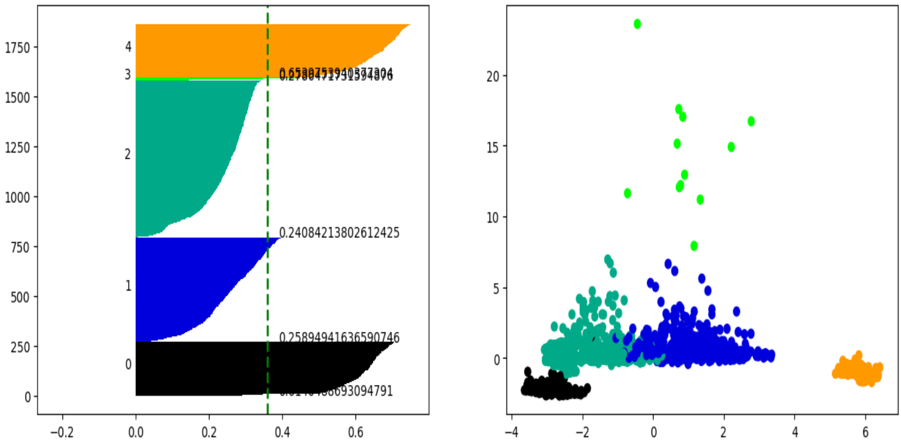
	REGION-CENTROID-COL	REGION-CENTROID-ROW	REGION-PIXEL-COUNT	SHORT-LINE-DENSITY-5	SHORT-LINE-DENSITY-2	VEDGE-MEAN	VEDGE-SD	HEDGE-MEAN	HEDGE-SD	INTENSITY-MEAN	RAWRED-MEAN	RAWBLUE-MEAN	RAWGREEN-MEAN	EXRED-MEAN	EXBLUE-MEAN	EXGREEN-MEAN	VALUE-MEAN	SATURATION-MEAN	HUE-MEAN
REGION-CENTROID-COL	1	0.012		-0.043	0.014	0.0082	0.047	-0.0065	0.022	0.071	0.067	0.07	0.077	-0.099	0.049	0.016	0.072	-0.1	0.037
REGION-CENTROID-ROW	0.012	1		0.053	0.05	0.029	-0.053	0.13	-0.02	-0.47	-0.47	-0.48	-0.44	0.36	-0.5	0.49	-0.46	0.092	0.6
REGION-PIXEL-COUNT																			
SHORT-LINE-DENSITY-5	-0.043	0.053		1	-0.038	-0.02	-0.034	-0.018	-0.043	-0.024	-0.022	-0.026	-0.023	0.042	-0.036	0.02	-0.022	-0.042	0.097
SHORT-LINE-DENSITY-2	0.014	0.05		-0.038	1	0.22	0.19	0.27	0.26	-0.033	-0.038	-0.024	-0.038	-0.021	0.032	-0.034	-0.027	0.046	-0.076
VEDGE-MEAN	0.0082	0.029		-0.02	0.22	1	0.66	0.52	0.49	-0.011	-0.022	0.0042	-0.02	-0.083	0.093	-0.075	0.0017	-0.057	-0.094
VEDGE-SD	0.047	-0.053		-0.034	0.19	0.66	1	0.44	0.67	-0.0043	-0.0088	-0.00064	-0.0043	-0.037	0.021	0.002	-0.0028	0.005	-0.061
HEDGE-MEAN	-0.0065	0.13		-0.018	0.27	0.52	0.44	1	0.68	0.023	0.016	0.034	0.017	-0.087	0.091	-0.067	0.033	-0.12	-0.089
HEDGE-SD	0.022	-0.02		-0.043	0.26	0.49	0.67	0.68	1	0.00046	-0.0045	0.0045	0.00048	-0.046	0.027	-3.9e-05	0.0021	-0.017	-0.074
INTENSITY-MEAN	0.071	-0.47		-0.024	-0.033	-0.011	-0.0043	0.023	0.00046	1	1	1	1	-0.83	0.79	-0.52	1	-0.61	-0.33
RAWRED-MEAN	0.067	-0.47		-0.022	-0.038	-0.022	-0.0088	0.016	-0.0045	1	1	0.99	0.99	-0.79	0.77	-0.51	0.99	-0.62	-0.32
RAWBLUE-MEAN	0.07	-0.48		-0.026	-0.024	0.0042	-0.00064	0.034	0.0045	1	0.99	1	0.98	-0.86	0.84	-0.58	1	-0.6	-0.38
RAWGREEN-MEAN	0.077	-0.44		-0.023	-0.038	-0.02	-0.0043	0.017	0.00048	1	0.99	0.98	1	-0.82	0.74	-0.43	0.99	-0.61	-0.26
EXRED-MEAN	-0.099	0.36		0.042	-0.021	-0.083	-0.037	-0.087	-0.046	-0.83	-0.79	-0.86	-0.82	1	-0.85	0.44	-0.86	0.42	0.28
EXBLUE-MEAN	0.049	-0.5		-0.036	0.032	0.093	0.021	0.091	0.027	0.79	0.77	0.84	0.74	-0.85	1	-0.85	0.83	-0.41	-0.64
EXGREEN-MEAN	0.016	0.49		0.02	-0.034	-0.075	0.002	-0.067	-3.9e-05	-0.52	-0.51	-0.58	-0.43	0.44	-0.85	1	-0.54	0.28	0.8
VALUE-MEAN	0.072	-0.46		-0.022	-0.027	0.0017	-0.0028	0.033	0.0021	1	0.99	1	0.99	-0.86	0.83	-0.54	1	-0.6	-0.34
SATURATION-MEAN	-0.1	0.092		-0.042	0.046	-0.057	0.005	-0.12	-0.017	-0.61	-0.62	-0.6	-0.61	0.42	-0.41	0.28	-0.6	1	-0.058
HUE-MEAN	0.037	0.6		0.097	-0.076	-0.094	-0.061	-0.089	-0.074	-0.33	-0.32	-0.38	-0.26	0.28	-0.64	0.8	-0.34	-0.058	1

Outliers



Except for Sky, every image have outliers. Cement has one outlier above the upper whisker. On the other hand, Brickface, Foliage, Grass, Path and Window have several outliers above the upper whisker.

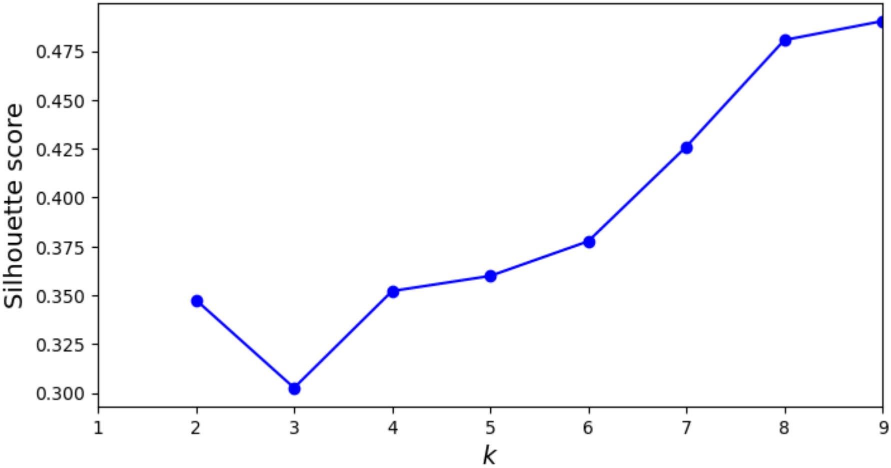
Unsupervised learning: Clustering



Shown is the K-means clustering with K=5

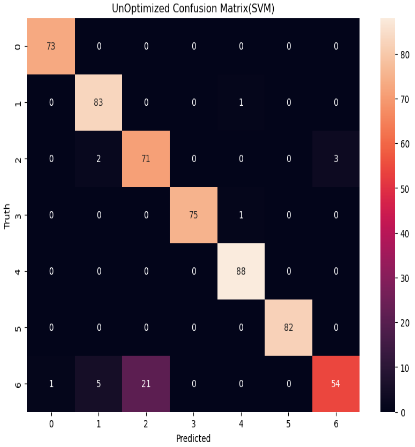
We have repeated the experiment many times with different clusters. Computing the silhouette scores

Silhouette Analysis by Elbow Method

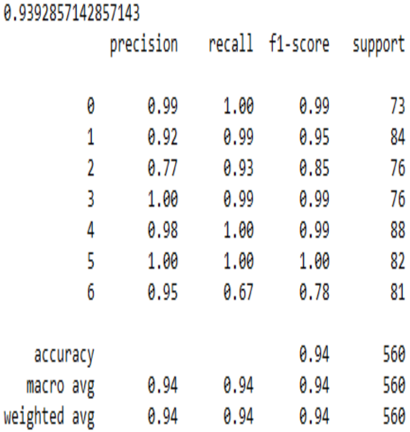


Here we can see that Silhouette's score is increasing with the change in the number of clusters which clearly shows that Clustering is not a better strategy for this data set.

Unoptimized SVM Classifier: Confusion matric with Classification Report



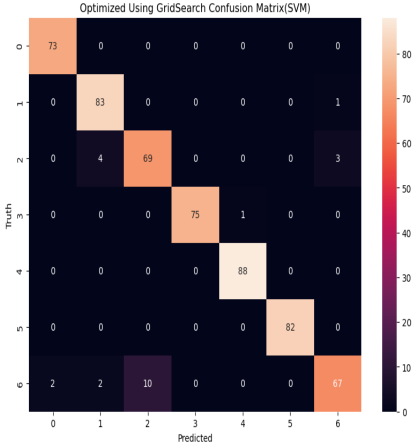
	0	1	2	3	4	5	6
0	73	0	0	0	0	0	0
1	0	83	0	0	1	0	0
2	0	2	71	0	0	0	3
3	0	0	0	75	1	0	0
4	0	0	0	0	88	0	0
5	0	0	0	0	0	82	0
6	1	5	21	0	0	0	54



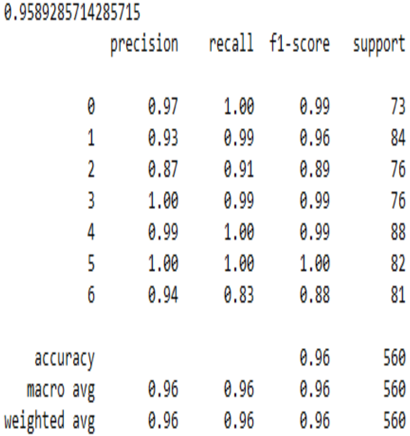
```
0.9392857142857143
              precision    recall  f1-score   support

           0       0.99      1.00      0.99        73
           1       0.92      0.99      0.95        84
           2       0.77      0.93      0.85        76
           3       1.00      0.99      0.99        76
           4       0.98      1.00      0.99        88
           5       1.00      1.00      1.00        82
           6       0.95      0.67      0.78        81

    accuracy                           0.94       560
   macro avg       0.94      0.94      0.94       560
weighted avg       0.94      0.94      0.94       560
```

Prediction accuracy is 94 percent

Optimized SVM Classifier: Confusion matric with Classification Report



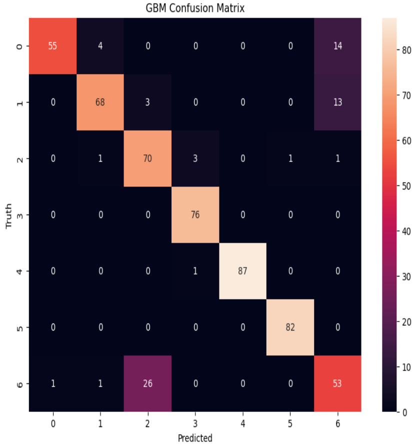
Truth \ Predicted	0	1	2	3	4	5	6
0	73	0	0	0	0	0	0
1	0	83	0	0	0	0	1
2	0	4	69	0	0	0	3
3	0	0	0	75	1	0	0
4	0	0	0	0	88	0	0
5	0	0	0	0	0	82	0
6	2	2	10	0	0	0	67

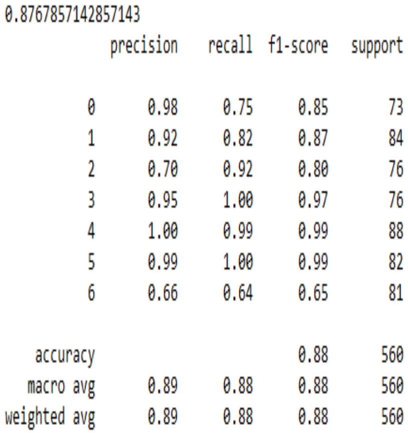


```
0.9589285714285715
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        73
           1       0.93      0.99      0.96        84
           2       0.87      0.91      0.89        76
           3       1.00      0.99      0.99        76
           4       0.99      1.00      0.99        88
           5       1.00      1.00      1.00        82
           6       0.94      0.83      0.88        81

    accuracy                           0.96       560
   macro avg       0.96      0.96      0.96       560
weighted avg       0.96      0.96      0.96       560
```

Prediction accuracy is 96 percent

Gradient Boosting Classifier

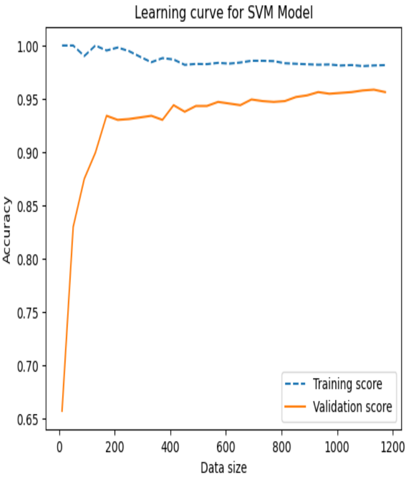


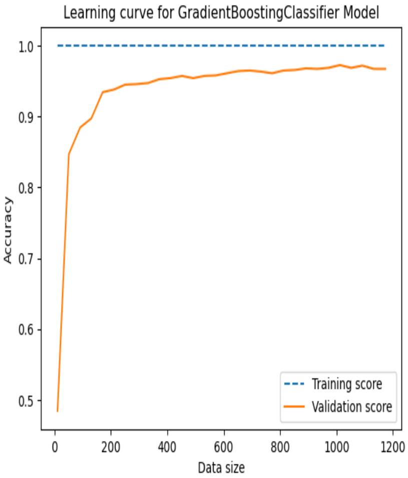


```
0.8767857142857143
              precision    recall  f1-score   support

           0       0.98      0.75      0.85        73
           1       0.92      0.82      0.87        84
           2       0.70      0.92      0.80        76
           3       0.95      1.00      0.97        76
           4       1.00      0.99      0.99        88
           5       0.99      1.00      0.99        82
           6       0.66      0.64      0.65        81

    accuracy                           0.88       560
   macro avg       0.89      0.88      0.88       560
weighted avg       0.89      0.88      0.88       560
```

Overall, the prediction accuracy is above 80 percent, which is considered a good performance. The results suggested that F1-score is high in "Window" and "Path" features (Class 4 and 5), which had the largest sample size. The opposite is true for "grass" (Class 6).

Learning Curve for GBC Model and SVM Model





The learning curve suggests that the SVM model achieves higher accuracy than GBC Model and is ready to predict new datasets.

Results

In this exercise, we have built a successfully optimized Support Vector Machine classifier, of which we were able to improve the accuracy slightly from 0.93 to 0.95 by tuning the model's parameters. The optimum set of parameters for this SVM model is 'C': 100, 'kernel': 'rbf'. We also compared it with the accuracy of the Gradient Boosting Classifier and found that it has accuracy of 0.875, which less than that of the SVM Classifier

Thank you!