Cats and Dogs Dataset Analysis

Elínborg Ásbergsdóttir
İpek Korkmaz
Luca Modica
Patrícia Marques

Group 27
Room SB-L111

23.05.2024

Part 1: Summary

Classifiers

- Random Forest (RF)
- Logistic Regression (LR)
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Multilayer Perceptron (MLP)

Important Pixels Selection

- Random Forest (RF) feature importances
- Logistic Regression (LR) coefficients
- ANOVA F-Test scores

Classifiers Analysis

- Accuracy
- F1 score
- ROC AUC score

Clustering Methods

- K-Means
- Hierarchical Clustering

Clustering Preprocessing

- Standard scaling
- Dimensional reduction
    - Kernel PCA

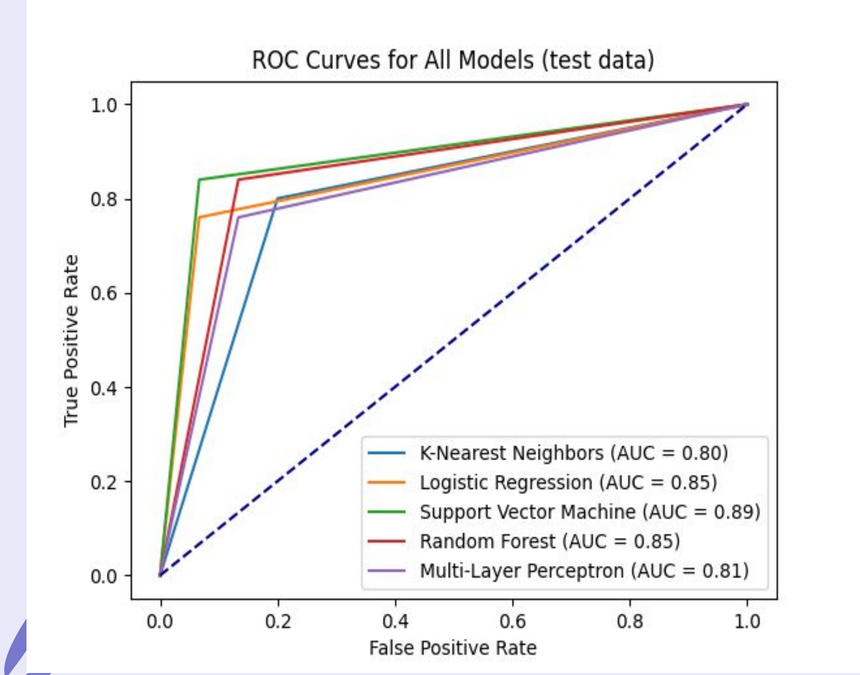
Cluster Analysis

- Internal evaluation
- External evaluation

Hyperparameter Fine-tuning

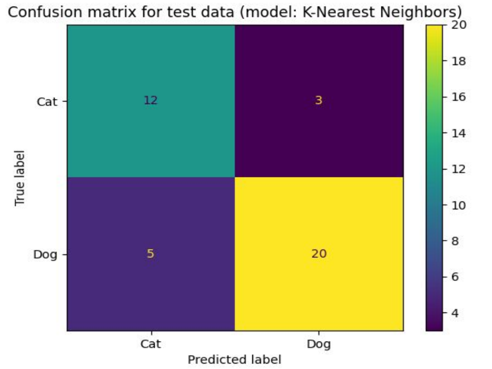
- Number of clusters
- Linkage method

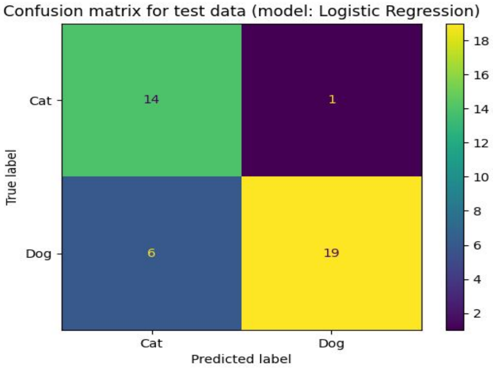
Classifiers Performance

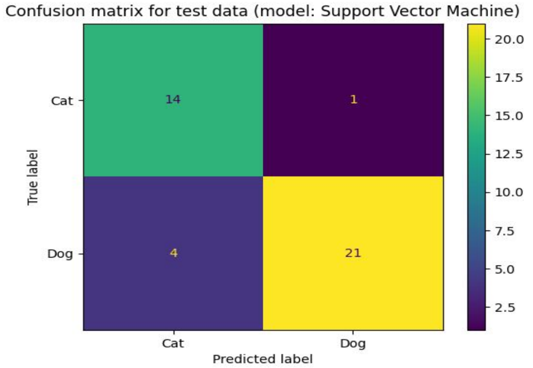


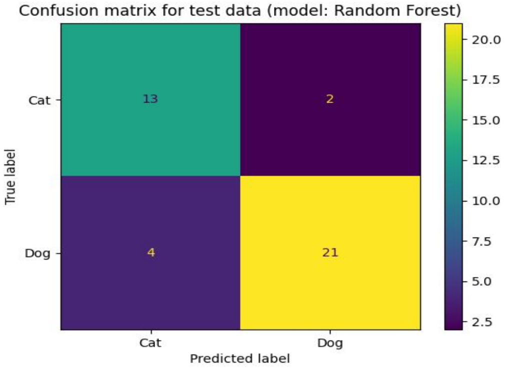
Model	KNN	LR	SVM	RF	MLP
CV Accuracy	0.80	0.87	0.90	0.70	0.85
Test Accuracy	0.80	0.83	0.88	0.85	0.80
Test F1 Score	0.79	0.82	0.87	0.84	0.80
Test ROC AUC score	0.80	0.85	0.89	0.85	0.81

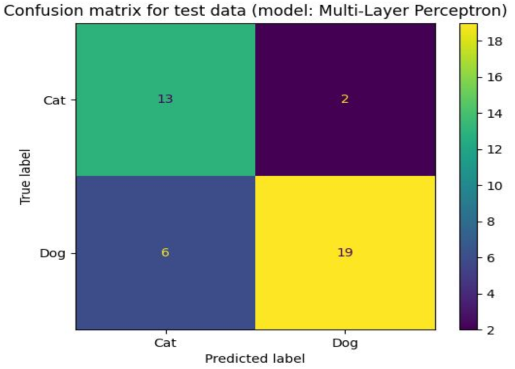
Confusion Matrices


Confusion matrix for test data (model: K-Nearest Neighbors)


Confusion matrix for test data (model: Logistic Regression)

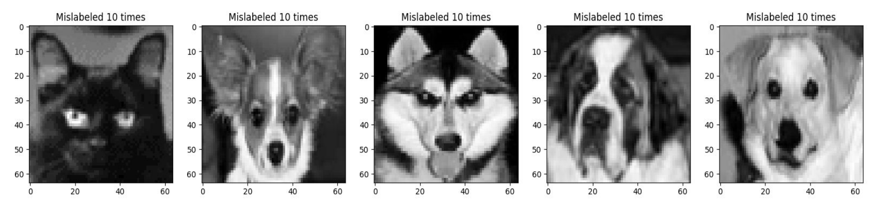

Confusion matrix for test data (model: Support Vector Machine)


Confusion matrix for test data (model: Random Forest)

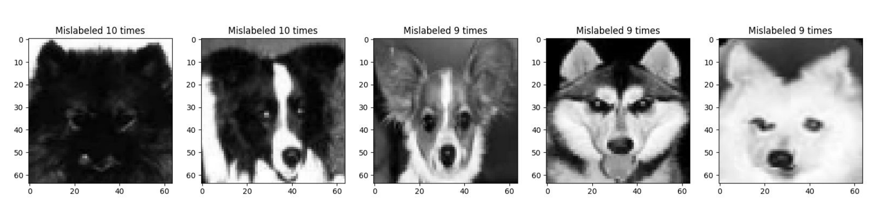

Confusion matrix for test data (model: Multi-Layer Perceptron)

Misclassified Images

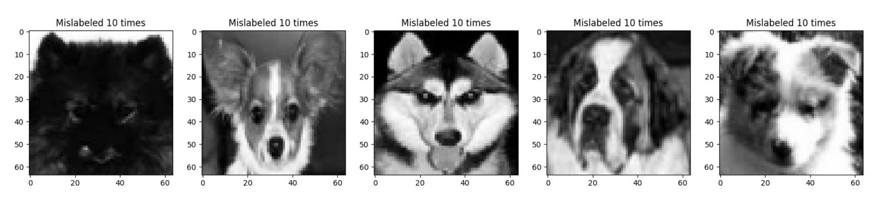
KNN



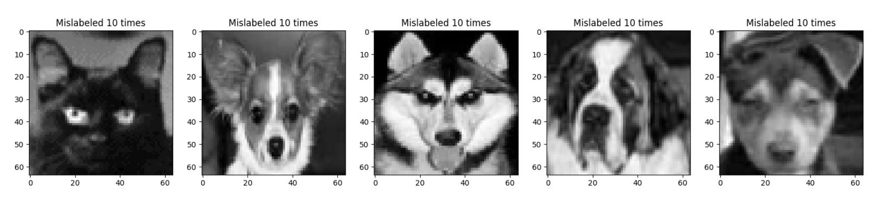
MLP



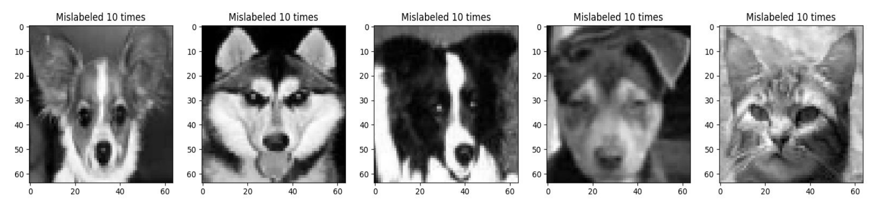
Logistic Regression



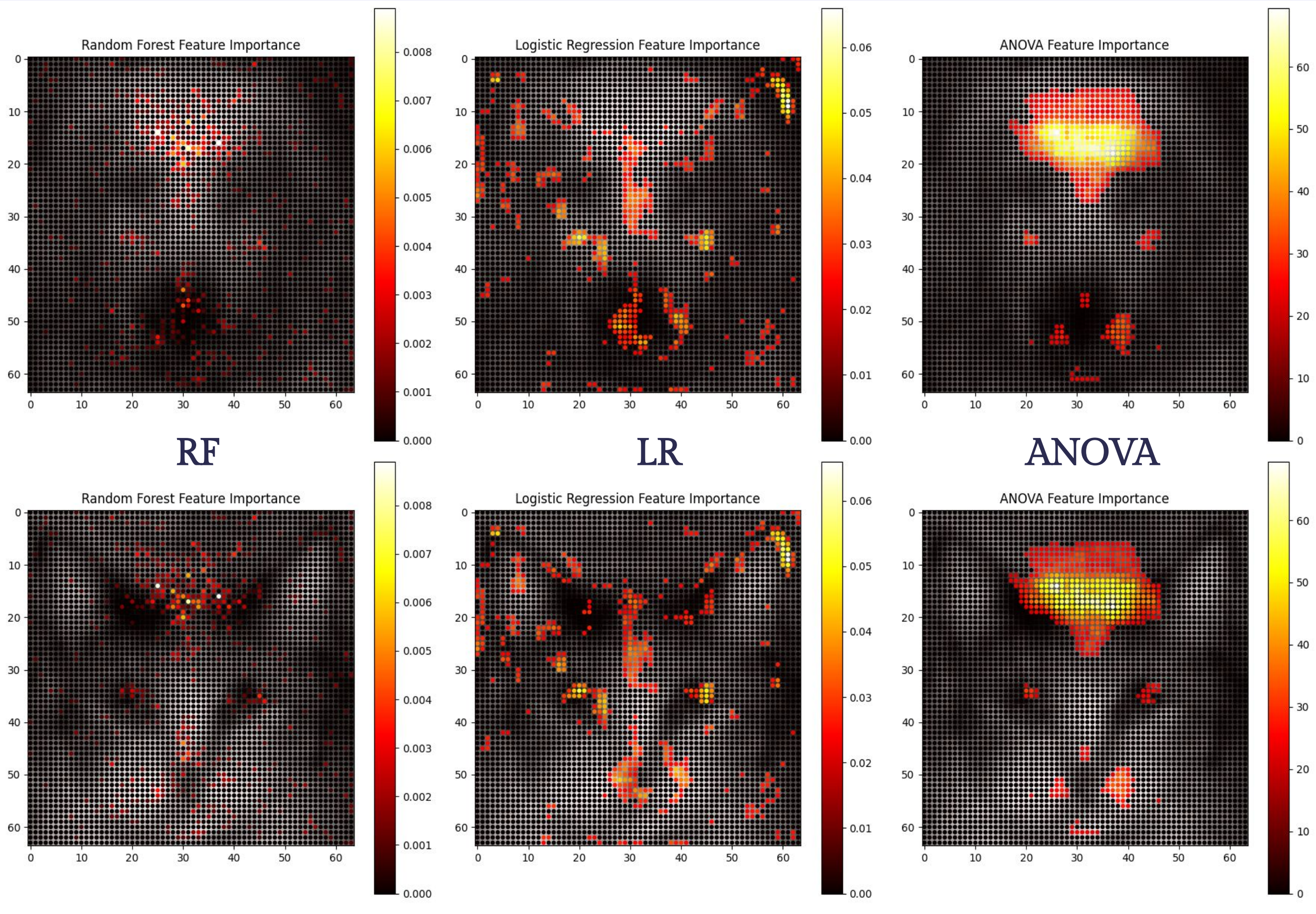
SVM



Random Forest



Top Pixels

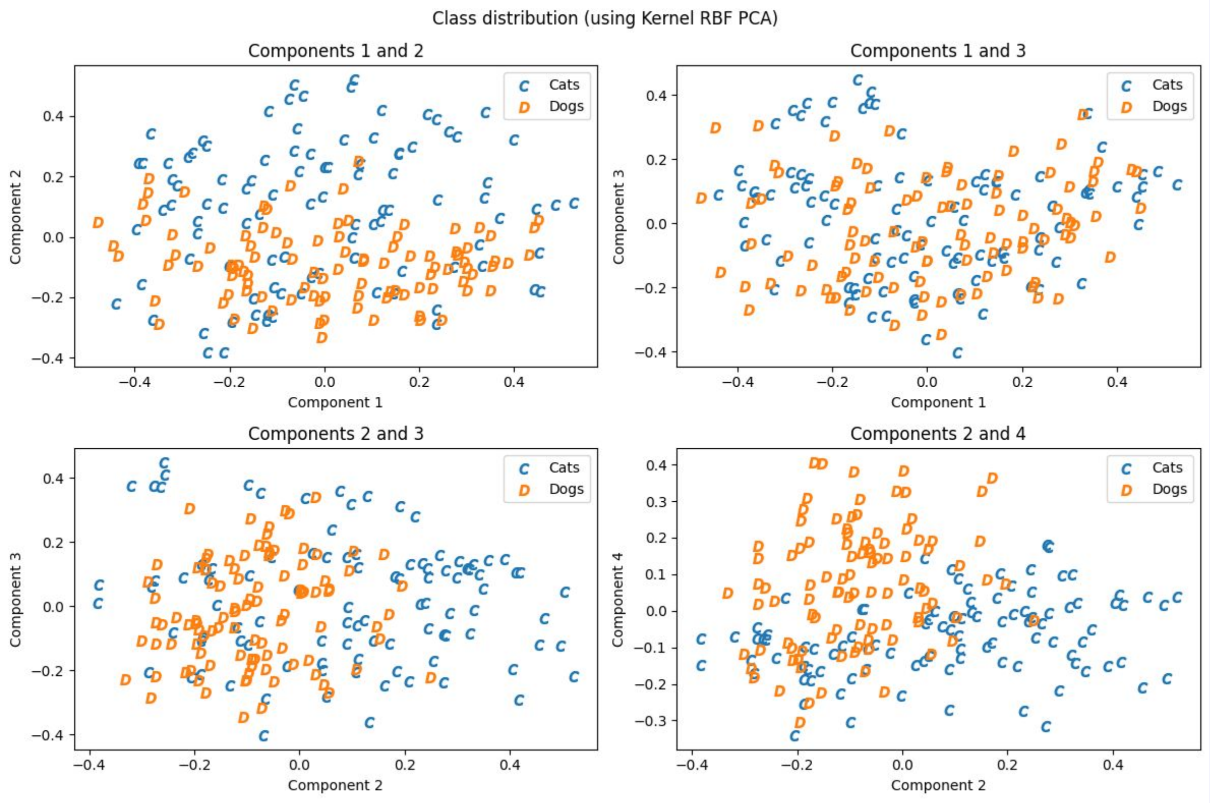


RF

LR

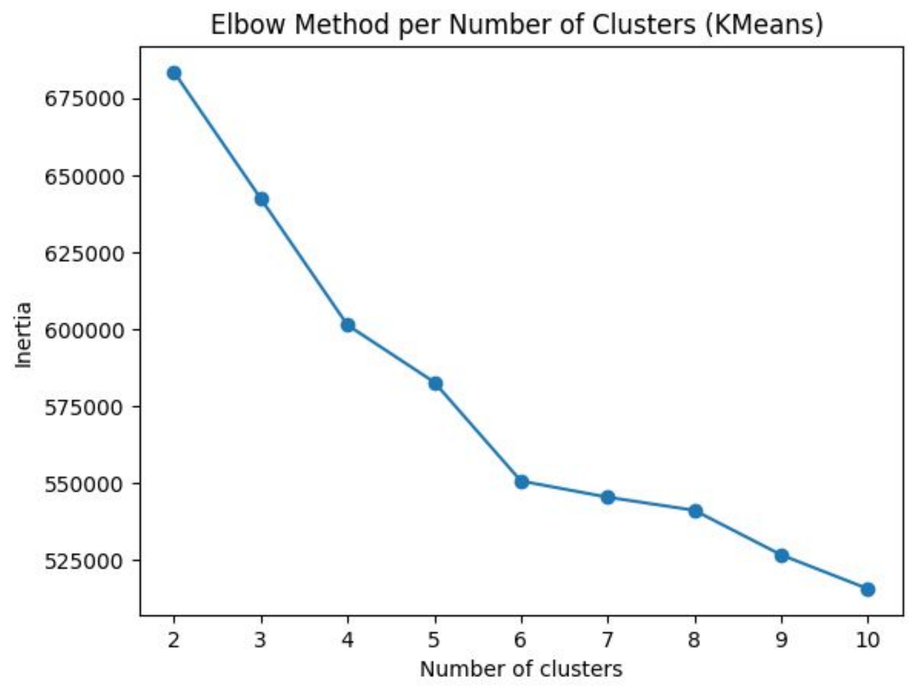
ANOVA

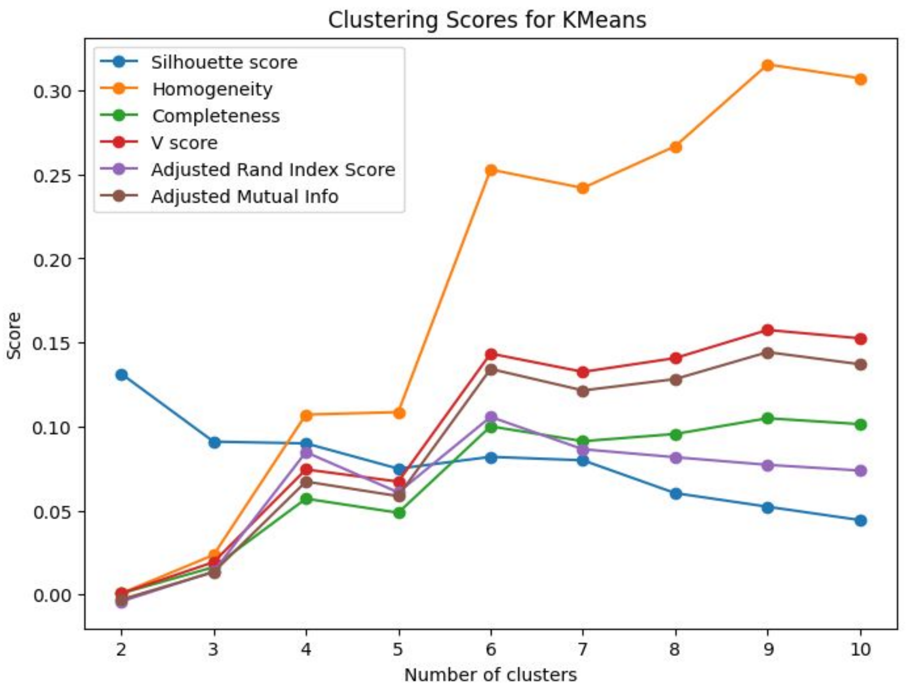
Clustering Preliminaries



K-means

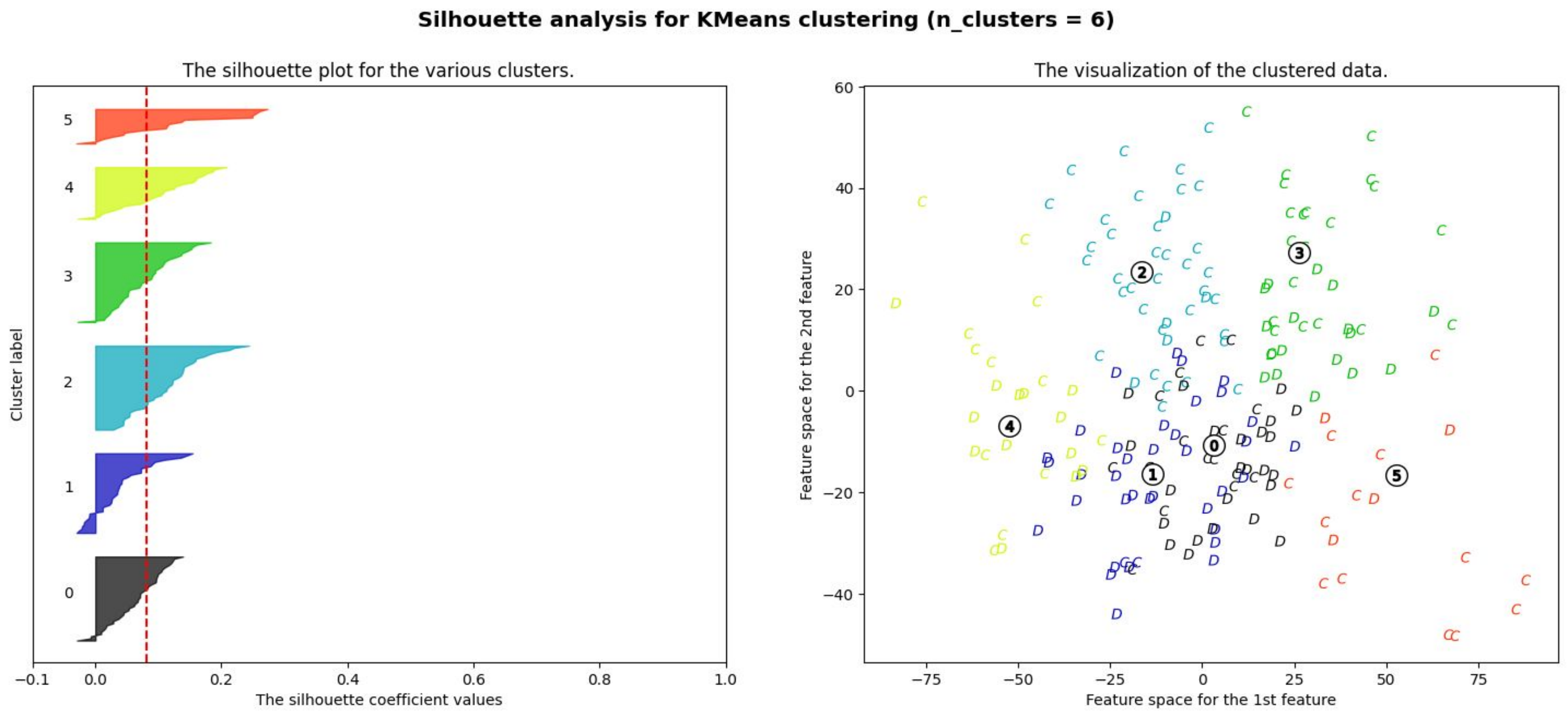
With Feature Scaling





K-means

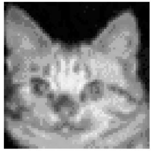
Silhouette Analysis



Representative Datapoint for Cluster 0

Representative Datapoint for Cluster 1

Representative Datapoint for Cluster 2



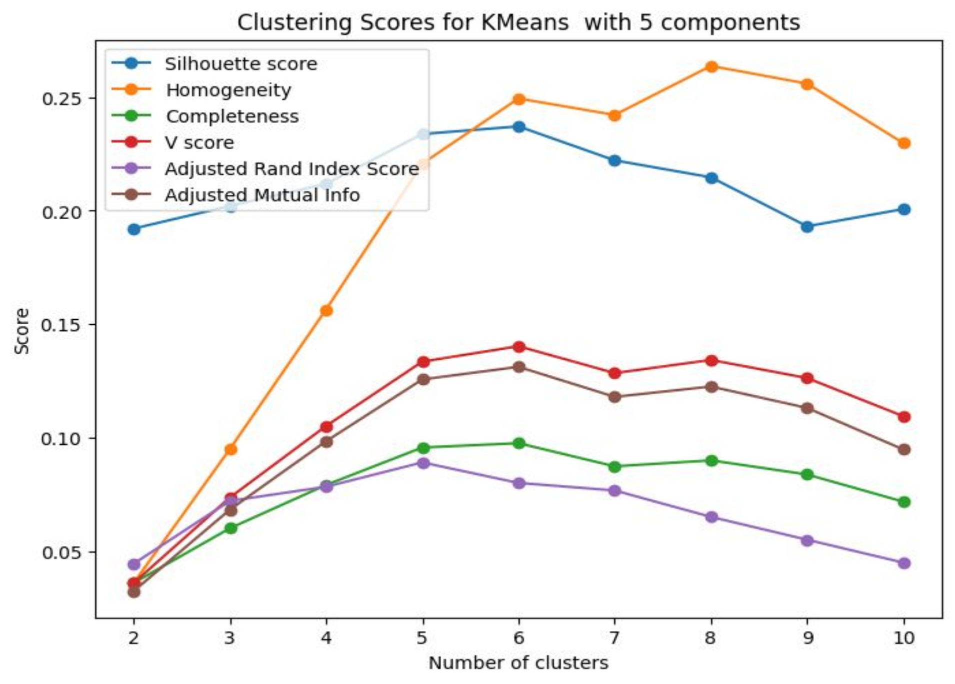
Representative Datapoint for Cluster 3

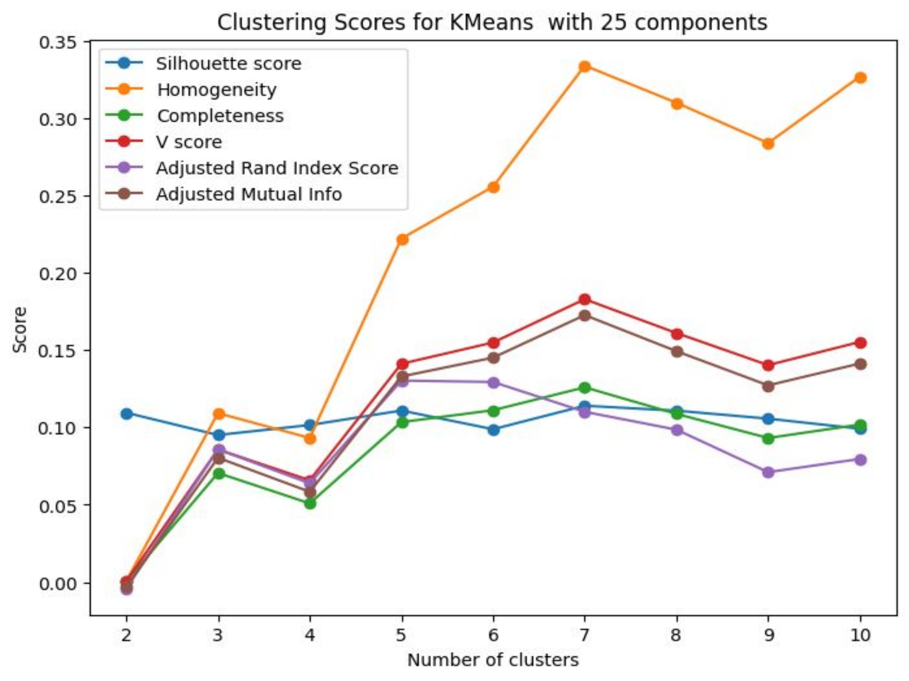
Representative Datapoint for Cluster 4

Representative Datapoint for Cluster 5

K-means

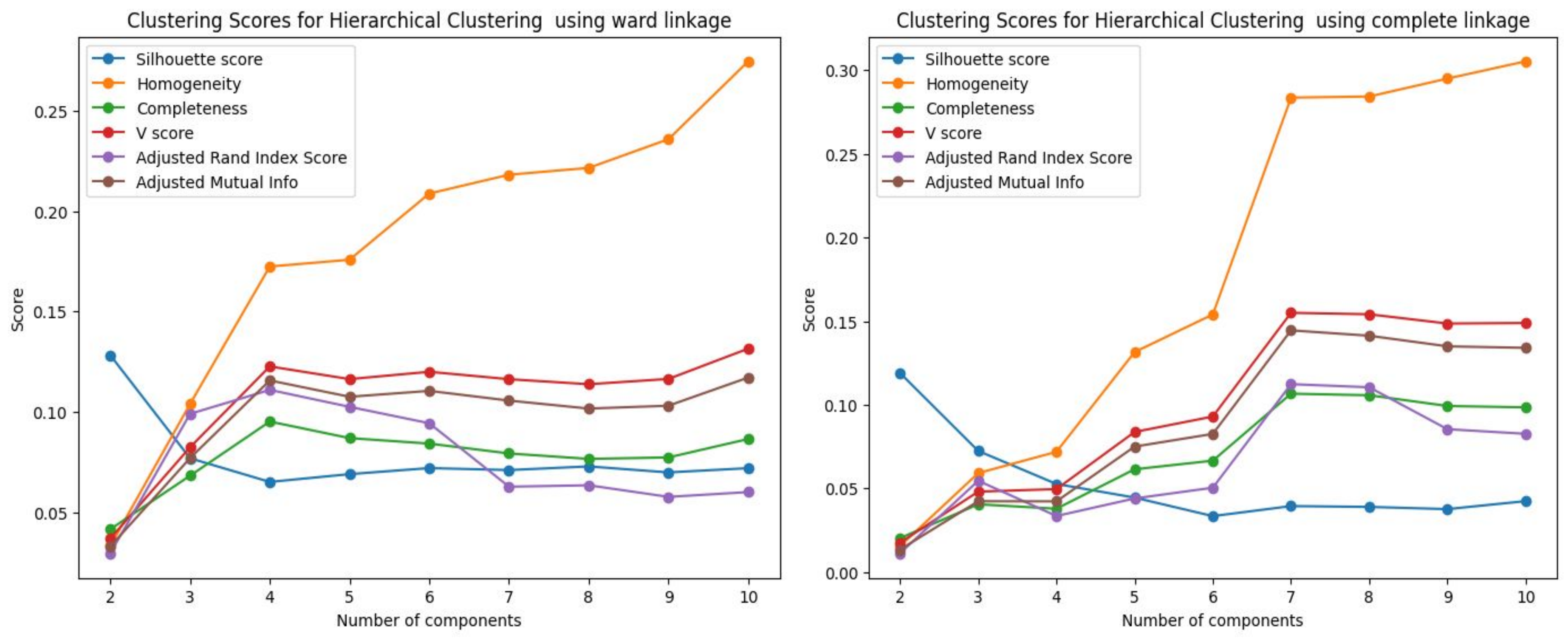
With Kernel PCA


Clustering Scores for KMeans with 5 components


Clustering Scores for KMeans with 25 components

Hierarchical Clustering

With Linkage Methods



Part 2: Summary

Theme 1

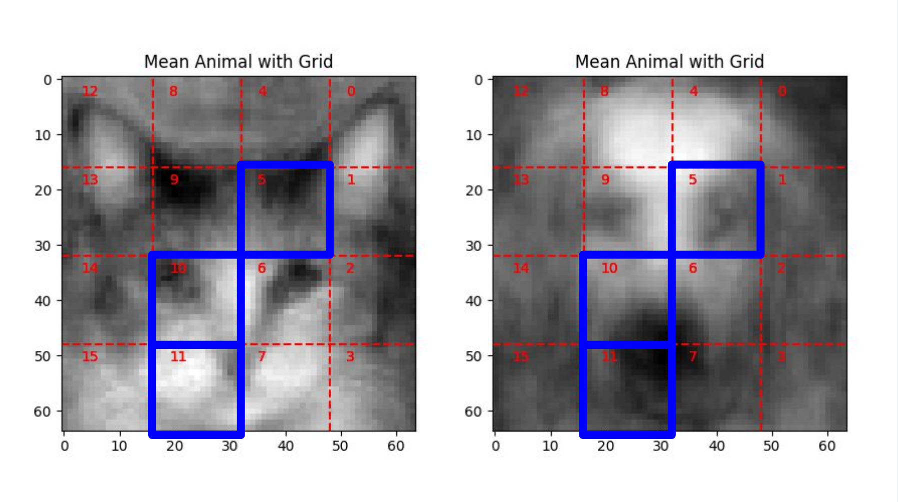
Datasets Variations

- Subset of 16x16 pixel blocks
- Half of the images untouched and half flipped
- Transposed dataframe (pixels as observations)

Pipeline

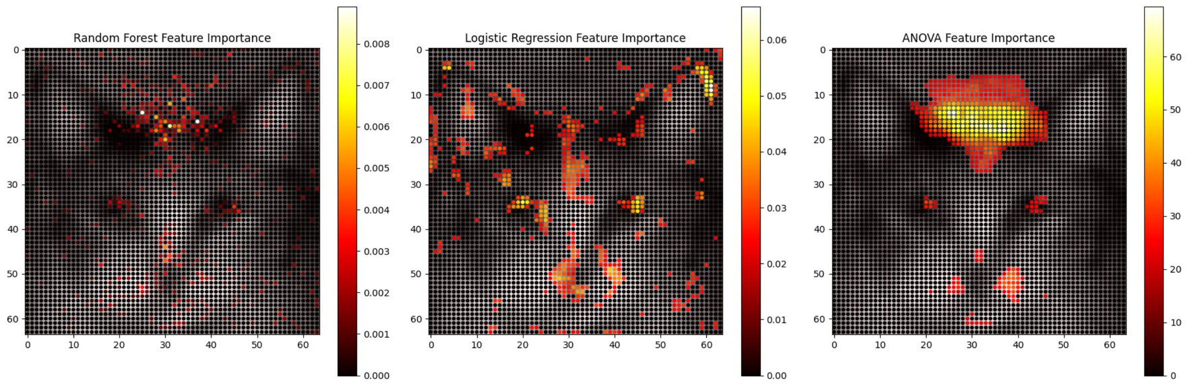
- Standard scaling
- Train (80%) and test (20%) splitting
- Stratified fold splitting (3 folds)
- Accuracy analysis (cross-validation, train and test)

Pixel Blocks



Block	10	11	5	Image
Best Model	RF	SVM	SVM	SVM
CV Accuracy	0.75	0.71	0.83	0.90
Test Accuracy	0.85	0.85	0.83	0.88
Test F1 Score	0.85	0.84	0.82	0.87

Flipping Images

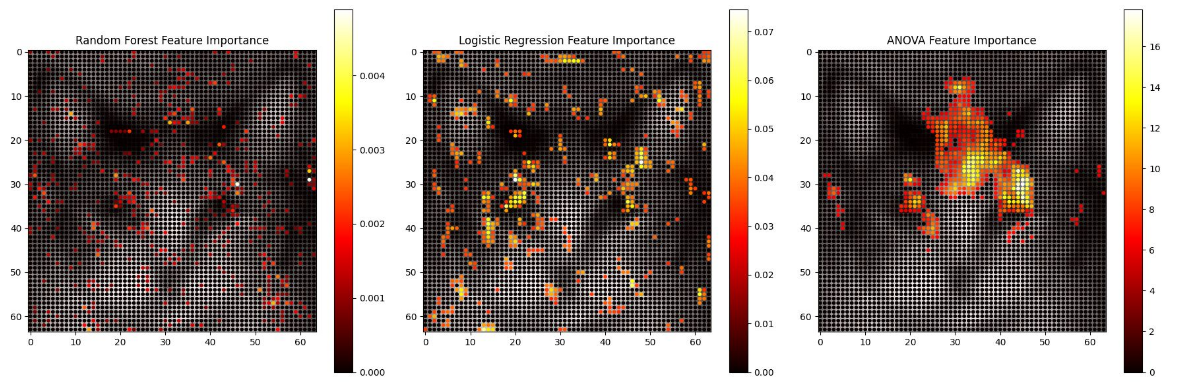


Random Forest Feature Importance

Logistic Regression Feature Importance

ANOVA Feature Importance

Original



Random Forest Feature Importance

Logistic Regression Feature Importance

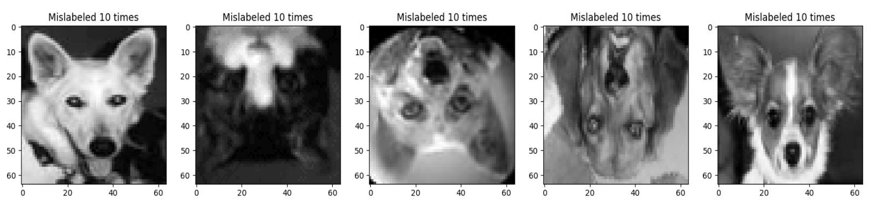
ANOVA Feature Importance

Half flipped

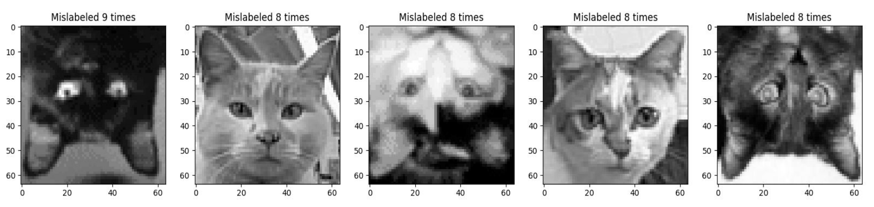
Images	Half flipped	Original (SVM)
Test Accuracy	0.81	0.88
Test F1 Score	0.80	0.87

Mislabeled Images

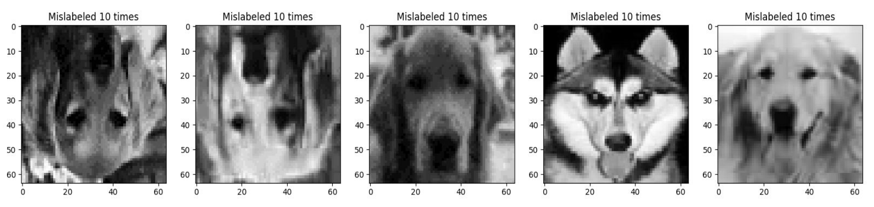
KNN



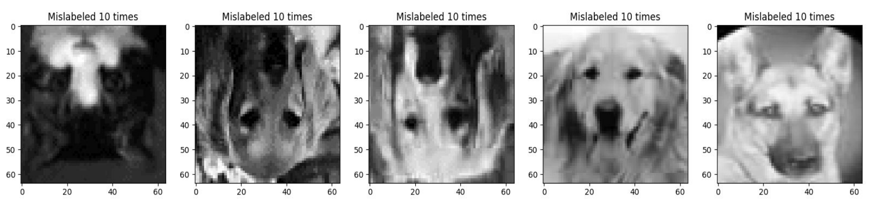
MLP



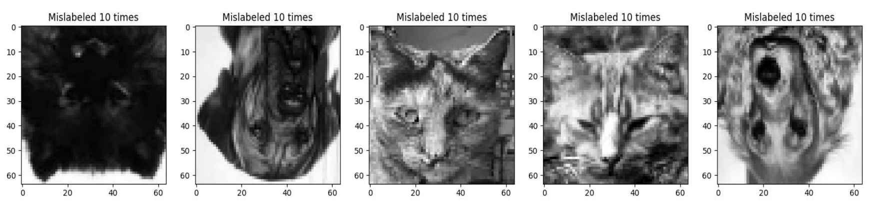
Logistic Regression



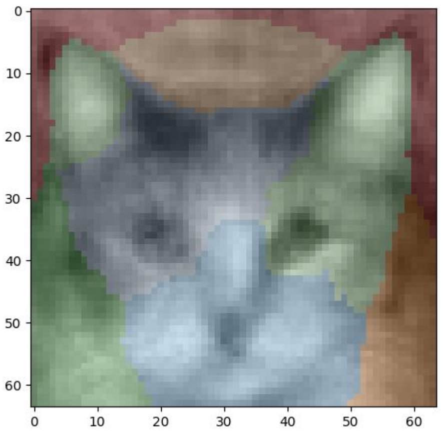
SVM

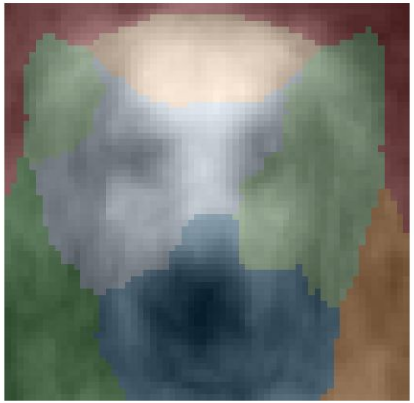


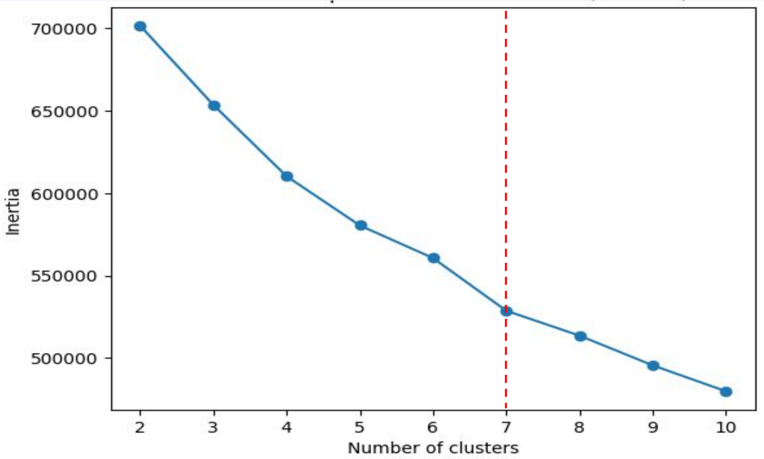
Random Forest



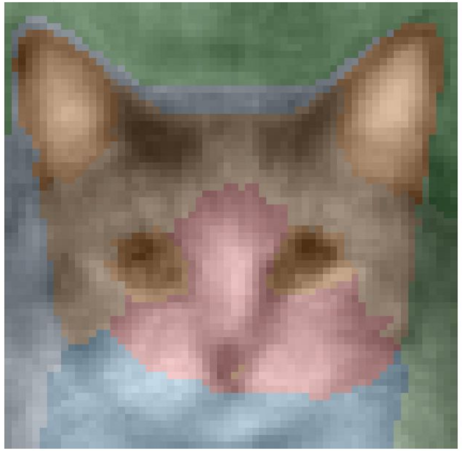
7 Clusters with All Data



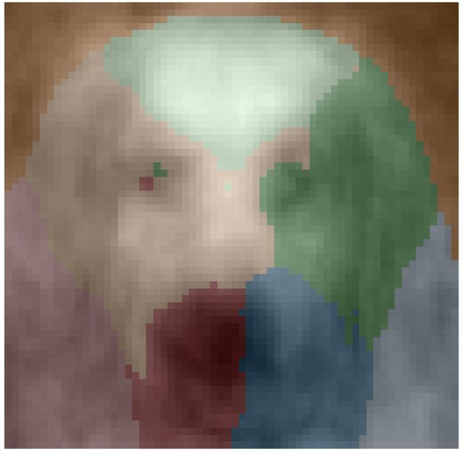




Overlay of Mean cats and Clustered cats Pixels (KMeans with 7 clusters)



Overlay of Mean dogs and Clustered dogs Pixels (KMeans with 8 clusters)

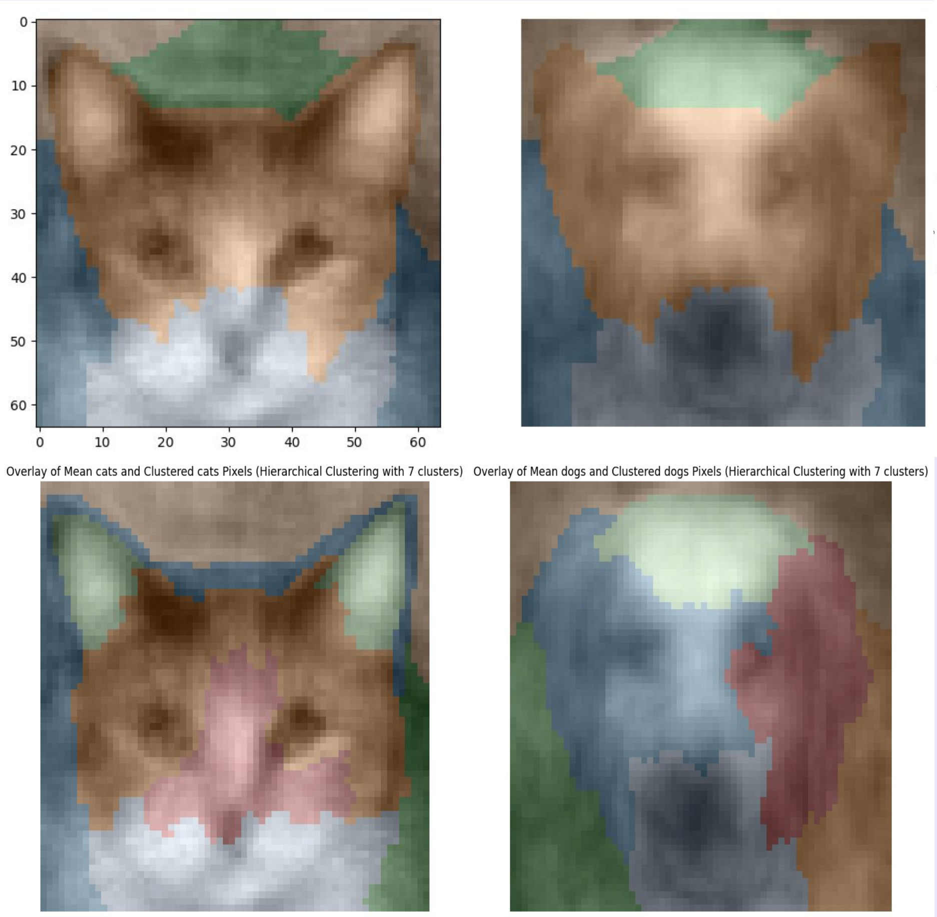


Pixels as Observations

K-means
With Feature Scaling

Differences between Cats and Dogs

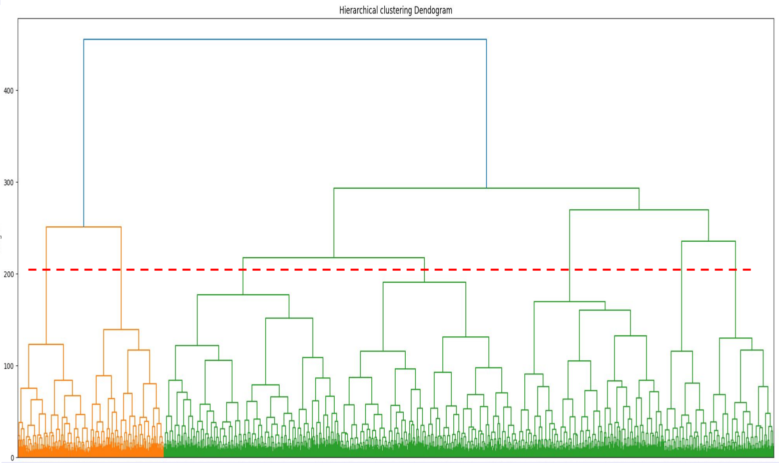
7 Clusters with All Data



Overlay of Mean cats and Clustered cats Pixels (Hierarchical Clustering with 7 clusters)

Overlay of Mean dogs and Clustered dogs Pixels (Hierarchical Clustering with 7 clusters)

Differences between Cats and Dogs



Pixels as Observations

Hierarchical Clustering With Ward's Linkage

Thanks

Does anyone have any questions?

Elínborg Ásbergsdóttir
İpek Korkmaz
Luca Modica
Patrícia Marques