Course code and name:	F20DL: Data Mining and Machine Learning
Type of assessment:	Group
Coursework Title:	Data Analysis and Bayes Nets Part 2
Student Names:	Hana Khan, Neil Patrao, Yasir Rayamarakar Veetil, George Chandy, Fatima Patel
Student ID Numbers:	H00379151, H00385850, H00385488, H00377409, H00339652

Declaration of authorship. By signing and uploading this form:

□ **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

□ I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

□ I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature: *Hana Khan, Neil Patrao, Yasir Rayamarakar Veetil, George Chandy, Fatima Patel*
Date: *06/11/2023*

Data Mining and Machine Learning

Data Analysis and Bayes Nets

F20DL CW PART TWO

Group11 {Hana Khan, Neil Patrao, Yasir Rayamarakar Veetil, George Chandy, Fatima Patel}

Stop Sign Dataset Notebook:

https://colab.research.google.com/drive/1Q4n35T50CM4mKbSE46azhFmijNmS7ZPA?usp=sharing

Additional Datasets Notebook (Car Dataset):

https://colab.research.google.com/drive/1MEGh8jh_o6JPOvO3qp-EVjbITSGno3UI?usp=sharing

The intuition behind the preprocessing, clustering algorithms and evaluation metrics are below. The numbers supporting these claims are in the table ensuing. The table does NOT contain all the datasets tested only the deemed best unsupervised datasets were mentioned.

Unsupervised Datasets Preprocessing

PCA: Reduces number of dimensions, therefore improves clustering results by reducing noise/redundancy so the clusters are better separated. It was also used for visualizing clusters in 2D space by reducing data to two dimensions.

Thresholding at 160 and 200: reduces noise, simplifies features and focuses on structure. Hence gave better silhouette scores compared to just PCA. However, it is very important to pick a good threshold value and use various metrics to evaluate it. For 160 the silhouette scores and accuracy were better than with just PCA. However, thresholding at 200 using K Means gave the highest silhouette score of 0.7 but there was no jump in accuracy and in fact reduced to 11% compared to 15% using PCA on original dataset.

Supervised Datasets Preprocessing

Top 5/10 Features Correlated with True Labels: kept most relevant features and therefore produced better results than the unsupervised data preprocessing techniques.

Clustering Algorithms Used

1- Vanilla K-Means Clustering: EM clustering algorithm used for hard clustering. The goal of K-Means clustering is to group similar data points into the same cluster and separate dissimilar data points into different clusters. Since our dataset has a lot of class overlaps, clustering is not ideal to classify the images into 10 classes. However, clustering can be used to give more insights about which classes have the most overlap and therefore which classes need more further preprocessing to be effectively separated. K Means using Binary Threshold 160 gave the best silhouette score and accuracy compared to all other models and datasets tried.

2- Hierarchical clustering

2.1- Dendrogram: uses a single cutoff to determine the number of clusters in hierarchical clustering. The ideal number of clusters visualized was 3/4/5 depending on the linkage method and dataset used.

2.2- Agglomerative: starts off each data point as a cluster and merges the most similar clusters. It is useful for understanding the hierarchical structures with the dataset but not for achieving high accuracy classifying. Our data did have defined hierarchy between the classes but there were greater distinction of clusters for 5/6. It performed slightly worser than K Means but still relatively good compared to the other clustering algorithms.

2.3- Birch: BIRCH works better with larger datasets compared to agglomerative. It generates a small summary of the large dataset retaining as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset. Birch performed similarly to agglomerative but with slightly less accuracy and silhouette scores. This is because our dataset is not that huge.

3- GMM: was tried since it is well suited to datasets where they are relationships between features as in our case with the pixels and it can detect more sophisticated clustering structures. However, the silhouette scores, v measure and accuracy were lower compared to using K Means. This could be because K Means makes fewer assumptions about the data distribution and additionally, since there's a lot of class overlap, the model increases the covariance which doesn't necessarily reflect the true structure of the data.

4- Spectral Clustering: captures more complex structures and does not confine the clustering to a specific shape like K-means. It can better handle clusters of unequal sizes, noise and outliers. However, when tried on our dataset, it gave slightly lower silhouette scores and accuracy compared to K Means on the original dataset with PCA. This could be because it's based-on graph theory, and therefore, performs well when the clusters form compact, non-linearly separable groups in the feature space. When clusters have a lot of overlap, the affinity matrix might not capture the true relationships between different data points, leading to less effective clustering.

Metrics for Evaluation

Silhouette Score evaluates how well a data point fit into its assigned cluster and how distinct it is from other clusters.

V-Measure uses 2 criteria: homogeneity (each cluster has members of the same class) and completeness (all members of a given class are assigned to the same cluster). It is independent of the exact assignment of labels and works well with imbalanced datasets since it doesn't depend on the support of the classes.

Accuracy measures how many instances were correctly identified out of all instances. Since clustering is unsupervised, the labels of the predicted values do not reflect the classes in the cluster, therefore, we relabeled the clusters to the class with the greatest *number of occurrences in cluster : support* ratio within that cluster. Ratio of the class within the class was used instead of number of occurrences since the dataset was imbalanced and therefore highly skewed towards majority classes.

Elbow the sum of squared distances between each point and the centroid in a cluster. The lesser the better but for more clusters, the value will always decrease, hence we take the value before convergence. Most of the elbow values were around 5/6 for k means suggesting significant overlap.

BIC/AIC are mostly used for model selection. AIC is the measure relative quality of model for a given dataset. It balances the goodness of fit of the model (how well it fits the data) and the complexity of the model (the number of parameters). Similar to AIC, BIC places a stronger penalty on model complexity. Lower AIC and BIC score indicates a better model fit.

Comparison with Naïve Bayes

To compare clustering and Gaussian Naïve Bayes (unsupervised vs supervised learning techniques), accuracy and f measure was used. The accuracies using the same datasets for the Naïve Bayes was significantly better compared to clustering. This is because Naïve Bayes is a supervised learning algorithm that learns from labeled training data and then makes predictions. In contrast, clustering is an unsupervised learning technique that tries to group similar instances together without prior knowledge of the group labels. The best Gaussian Naïve Bayes classifier was for PCA on top 5 features with 60% accuracy, 51% f-measure and similarly for this data set the best clustering results were for PCA on top 5 features with 20% and 15% respectively.

Main Conclusions

When evaluating the clustering, the main metrics used were silhouette score, V-measure, and accuracy. Clustering is not very accurate when trying to classify into the 10 target classes but useful for gaining insights on the structure of the data, class overlap and visualizing the class distribution on scatter plots. K Means performed the best and the supervised preprocessed datasets performed better than the unsupervised ones. Thresholding worked best for unsupervised preprocessing techniques.

Best Evaluated Dataset	Model	No of Clusters	Silhouette Score	Elbow	V-Measure	Accuracy	BIC	AIC	Optimal Clusters
Original Dataset	Vanilla K-means	2	0.48061	3.3e+10	N/A	N/A	N/A	N/A	6
		3	0.46248	2.2e+10					
		4	0.46688	1.7e+10					
		5	0.40761	1.3e+10					
		6	0.40970	1.0e+10					
		7	0.37904	8.8e+09					
		8	0.38366	7.6e+09					
		9	0.38604	6.8e+09					
		10	0.38161	6.1e+09	0.111	15%			
		11	0.38223	5.5e+09					
		12	0.38189	5.0e+09					
Binary Threshold at 160	Vanilla K-means	2	0.59396	3.3e+10	N/A	N/A	N/A	N/A	6
		3	0.62873	2.2e+10					
		4	0.62997	1.7e+10					
		5	0.60825	1.3e+10					
		6	0.63444	1.0e+10					
		7	0.63202	8.8e+09					
		8	0.62743	7.6e+09					
		9	0.62944	6.8e+09					
		10	0.62610	6.1e+09	0.111	18%			
		11	0.62487	5.5e+09					
		12	0.61906	5.0e+09					
Original Dataset With PCA 2	Gaussian Mixture Model (GMM)	2	0.4045	N/A	N/A	N/A	344563	344527	2
		3	0.3309				337421	337342	
		4	0.3453				336251	336129	
		5	0.3375				335802	335637	
		6	0.3353				335074	334866	
		7	0.3198				334819	334567	
		8	0.3387				334675	334381	
		9	0.3329				334574	334237	
		10	0.3186		0.117	13%	334347	333786	
		11	0.3207				334210	333649	
		12	0.3290				334116	333609	
Original Dataset With PCA 2 then Normalized	Spectral Clustering	2	0.4394	N/A	N/A	N/A	N/A	N/A	3
		3	0.4732						
		4	0.4362						
		5	0.3548						
		6	0.3216						
		7	0.3159						
		8	0.3066						
		9	0.2821						
		10	0.2951		0.079	13%			
		11	0.2993						
		12	0.2959						
Threshold Images at 160 with PCA 2	Birch	2	0.51690	N/A	N/A	N/A	N/A	N/A	6
		3	0.57342						
		4	0.55002						
		5	0.58940						
		6	0.59837						
		7	0.58669						
		8	0.58653						
		9	0.59347						
		10	0.57879		0.088	15%			
		11	0.57778						
		12	0.58198						
Threshold Images at 160 with PCA 2	Agglomerative	2	0.53954	N/A	N/A	N/A	N/A	N/A	8
		3	0.59067						
		4	0.57878						
		5	0.58728						
		6	0.60054						
		7	0.61596						
		8	0.61661						
		9	0.60570						
		10	0.59946		0.091	14%			
		11	0.59576						
		12	0.56949						