PROJECT REPORT

(Project 1 – Clustering Analysis of the Land Mine Data Set)

(EAS 509 – Statistical Learning and Data Mining II)

Team Members:

1. Sujay Shrivastava (50496221) (sujayshr)
2. Utkarsh Mathur (50495131) (umathur)
3. Venkata Lakshmi Krishna Tejaswi Gudimetla (50496378) (vgudimet)

Instructor: Dr. Leonid Khinkis

Note: Please refer to the knit HTML file to better visualize 3D plots.

---

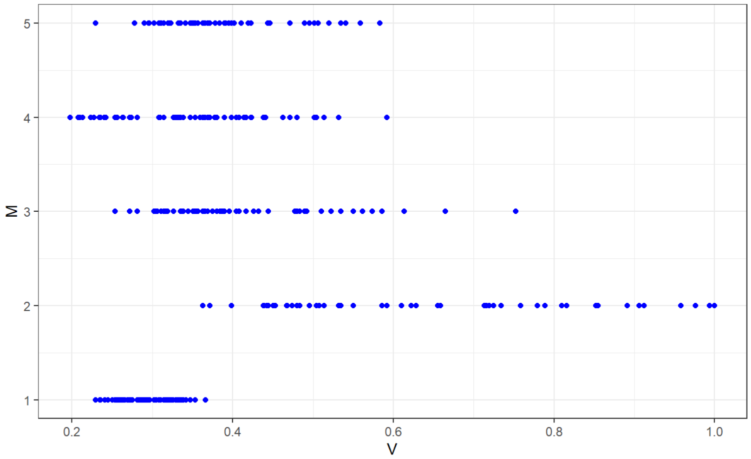
1. INTRODUCTION

- The aim of the project is to perform clustering analysis on the Land Mine data set of 338 observations and 4 features and derive insights to help us perform multiclass classification.

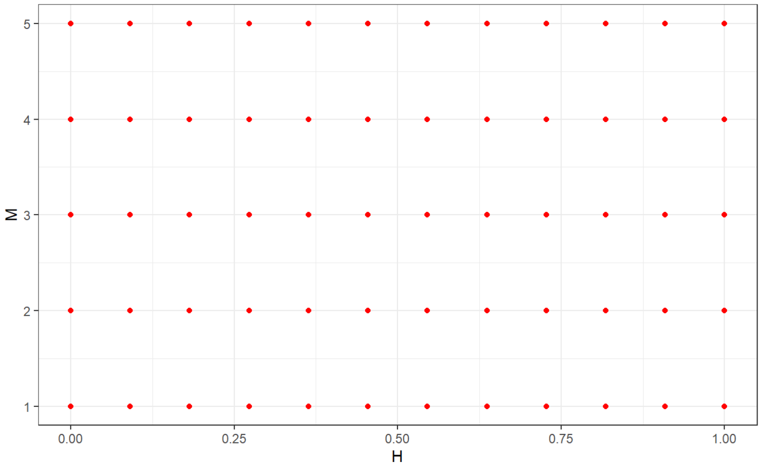
- This dataset has 4 features:
  1. V (Voltage) – Output voltage of the FLC sensor due to magnetic distortion.
  2. H (High) – The height of the sensor from the ground.
  3. S (Soil Type) – 6 soil type based on moisture condition (Dry and Sandy, Dry and Humus, Dry and Limy, Humid and Sandy, Humid and Humus, and Humid and Limy)
  4. M (Mine Type) – 5 mine types commonly encountered on land.

- The feature M is been addressed as the output or “**dependent variable**” by the authors so we have performed clustering analysis based on the “**independent variables**” (V, H, and S) and compared the results with the observed dependent variable. Furthermore, based on the insights from the clustering analysis we’ve trained classification models to validate our insights.

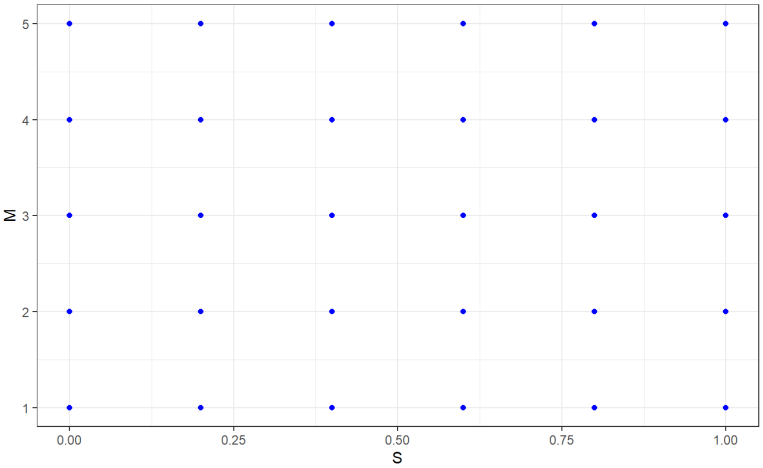
- We have performed K-means clustering and Hierarchical clustering for Clustering Analysis and used Logistic Regression and SVM (linear and radial kernels) for validation of insights.



(Figure 1: Scatter Plot between feature V and feature M)



(Figure 2: Scatter Plot between feature H and feature M)



(Figure 3: Scatter Plot between feature S and feature M)

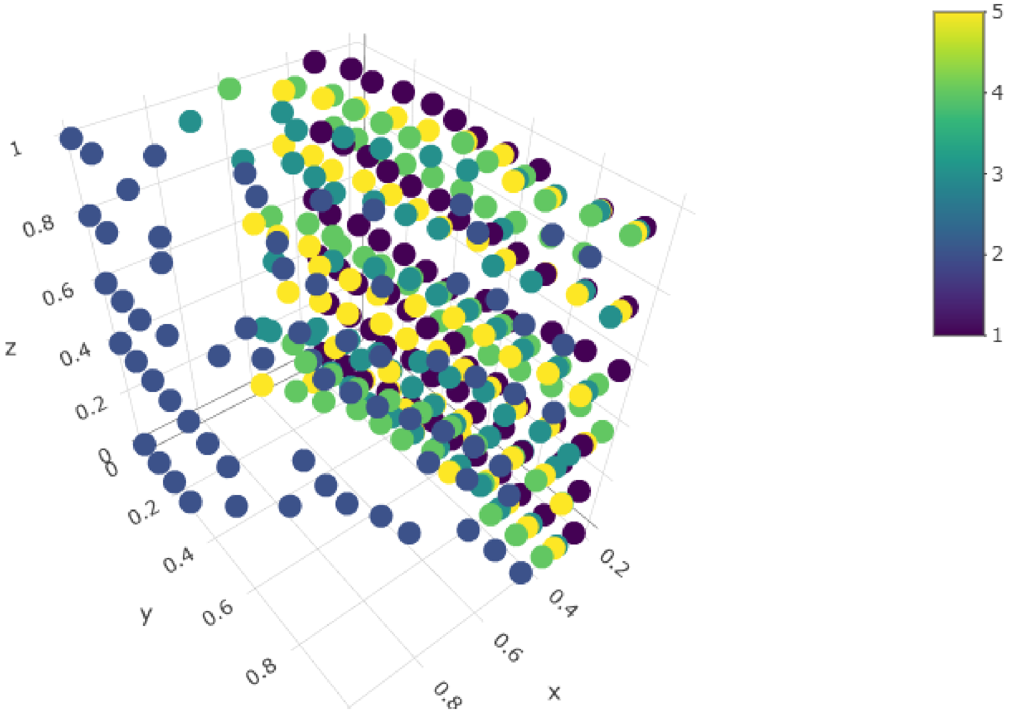
2. RESULTS & ANALYSIS

2.1. Clustering Analysis

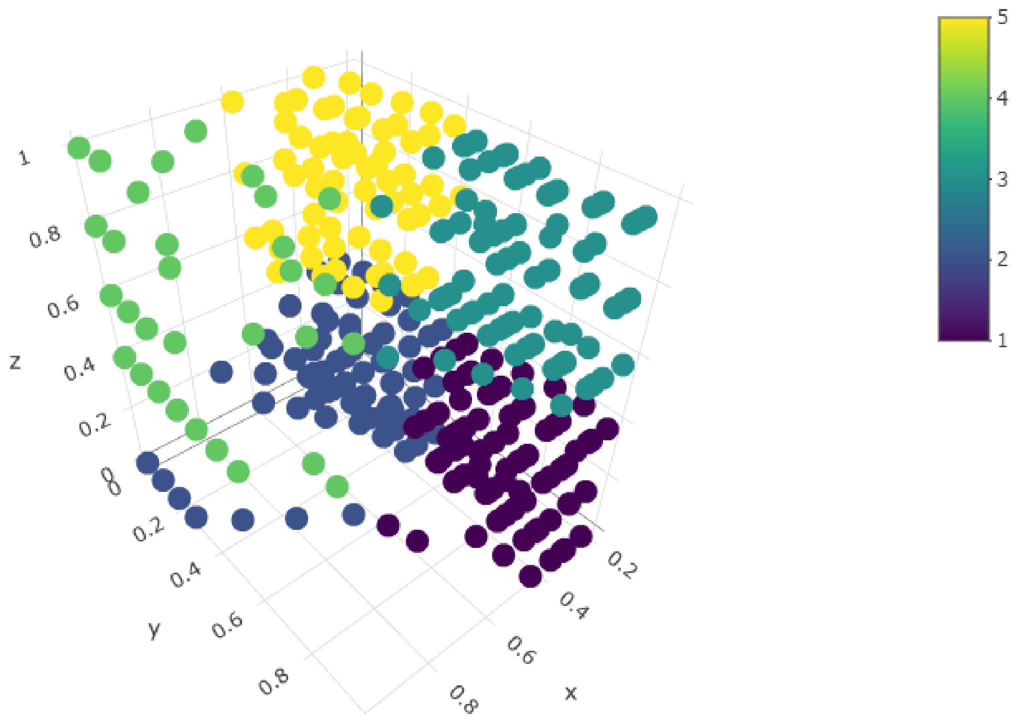
- Clustering Analysis is performed to investigate two aspects of the classification dataset:
  a. How does K-Means clustering perform in classifying the dataset based on V, H, and S for the given number of classes (i.e. 5)?
  b. What is the appropriate number of clusters based on K-Means and Hierarchical clustering in the dataset based on V, H, and S?

2.1.1. K-Means Clustering

- Upon performing the K-Means clustering for 5 clusters we obtained the below plots (refer to the HTML file to pan the 3D plots):

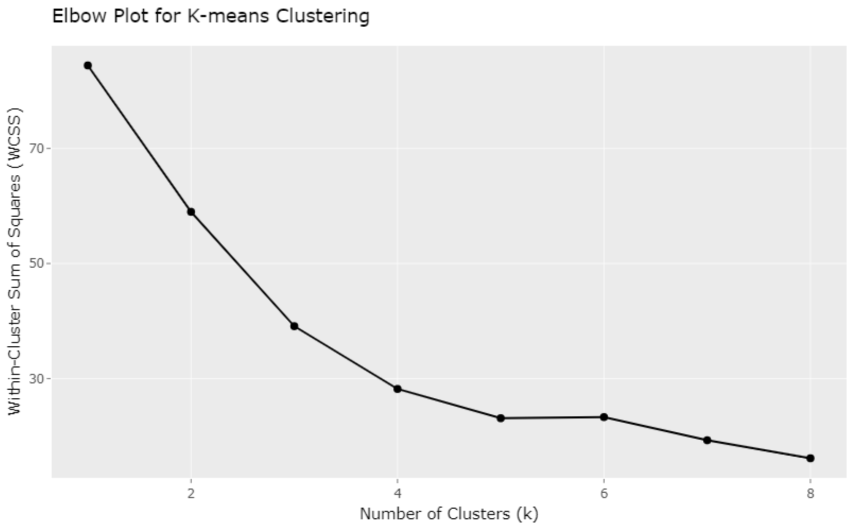


(Figure 4: V v/s H v/s S plot with points labelled based on classes in M)



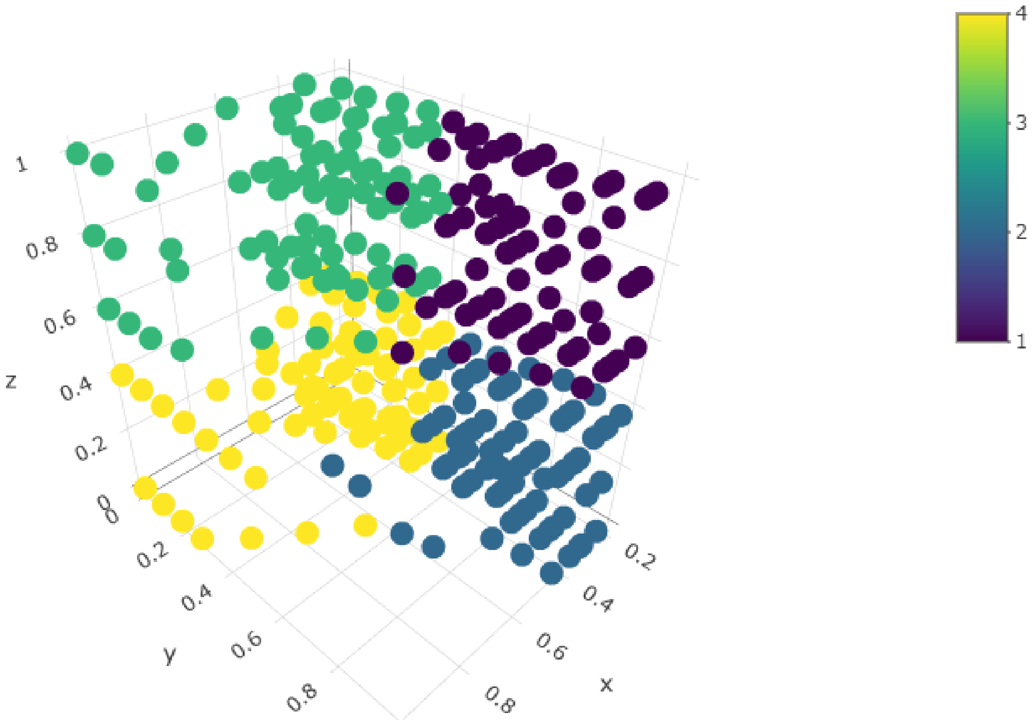
(Figure 5: V v/s H v/s S plot with points labelled based on K-Means clustering (K=5))

- It can be easily concluded that the observed classes have linearly separable boundaries and therefore radial boundaries are not the best classification technique here.
- We performed elbow method to find the appropriate number of clusters in the data based on the features V, H, and S and the results are shown in the plot below.



(Figure 6: Elbow test plot for K-Means Clustering)

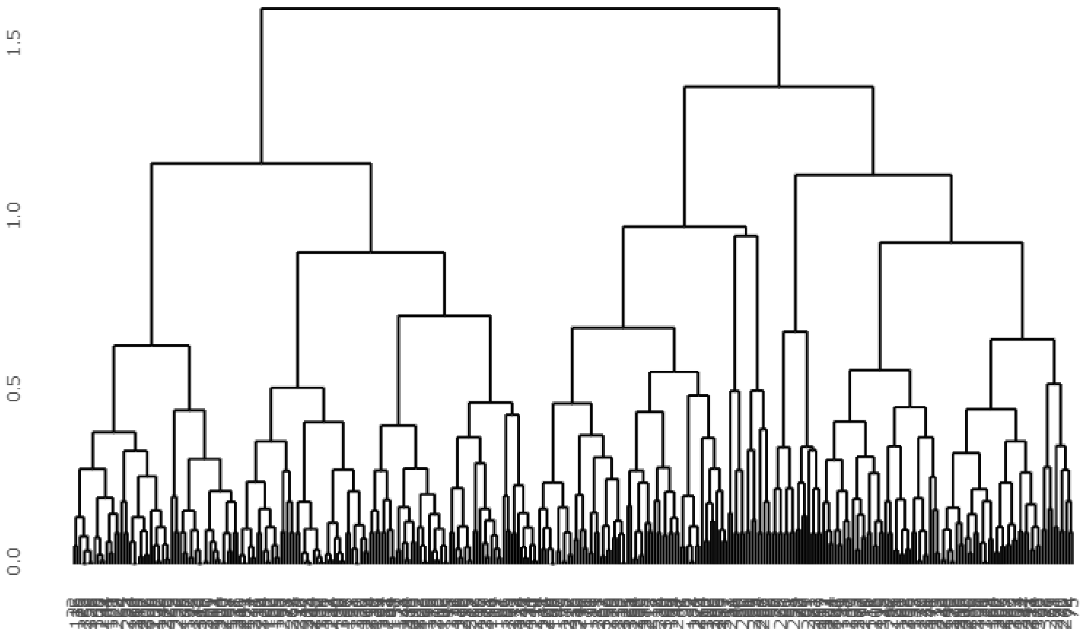
- It can be clearly seen that a simple radial K-means clustering will result yield best results with K=4 clusters and that this can be the natural classes based on the features V, H, and S as shown in the plot below.



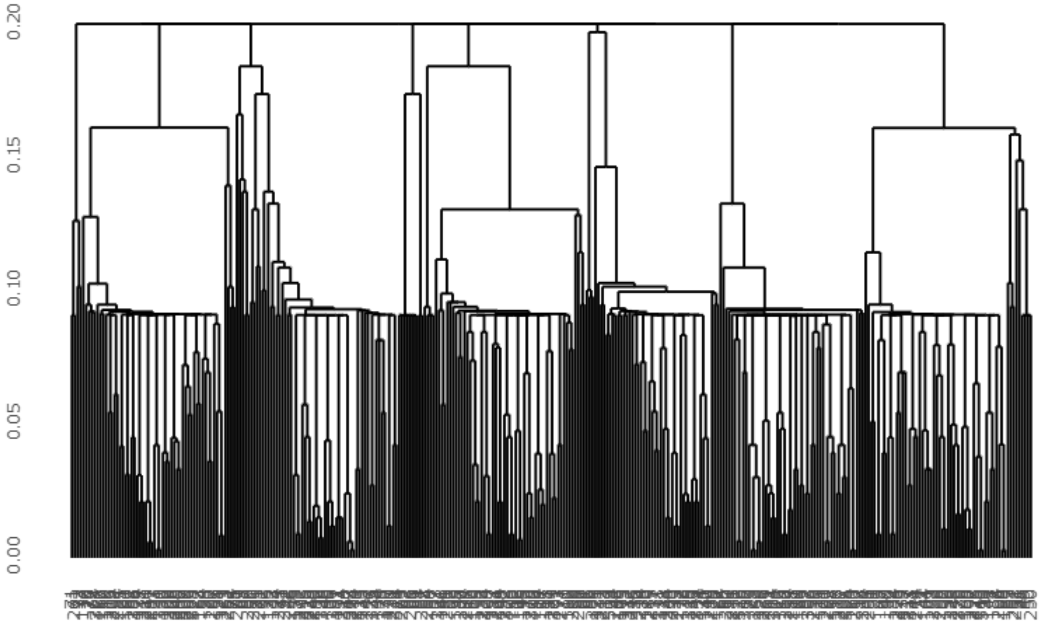
(Figure 7: V v/s H v/s S plot with points labelled based on K-means clustering (K=4))

2.1.2. Hierarchical Clustering

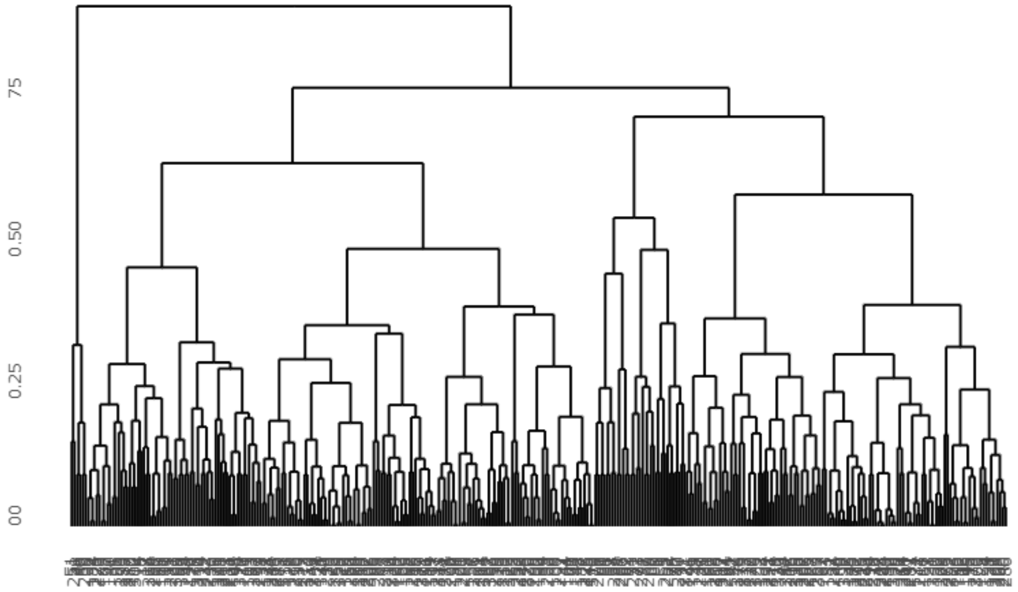
- Hierarchical clustering is performed to analyze the dendrograms obtained from three different linkages – **complete** (largest distance between any of the elements of the clusters), **single** (minimum distance between any of the elements of the clusters), and **average** (average distance between all the pair of elements of the clusters) – as shown in the plots below.



(Figure 8: Dendrogram for Complete Linkage)



(Figure 9: Dendrogram for Single Linkage)

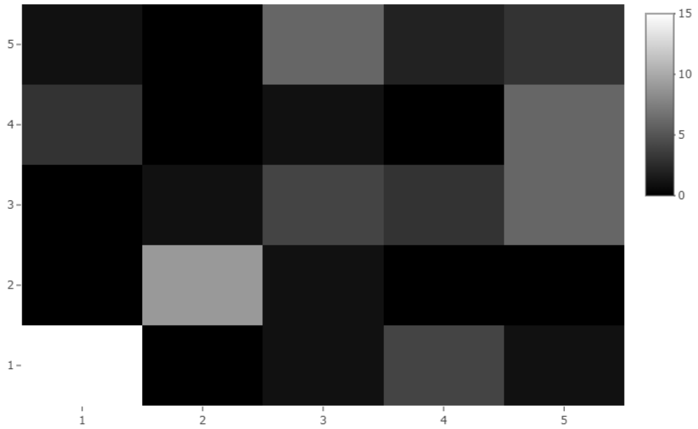


(Figure 10: Dendrogram for Average Linkage)

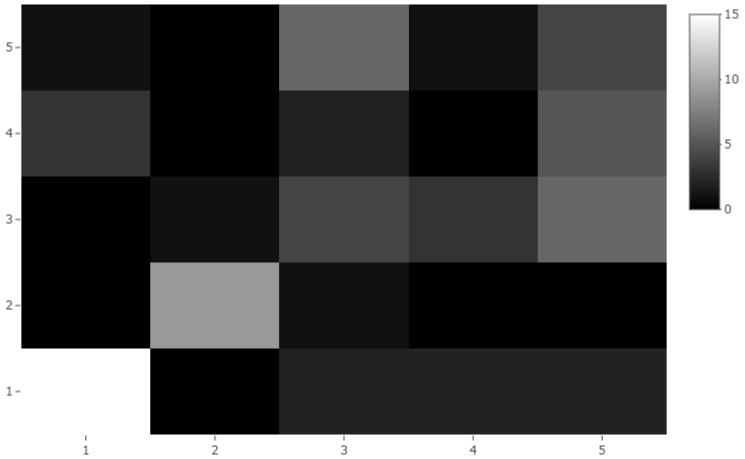
- Complete Linkage: The complete linkage suggests that the appropriate number of linkages is 4 or 5. This analysis syncs with the insights gained from K-means clustering.
- Single Linkage: The single linkage suggests that the appropriate number of clusters is 8 or 9. This analysis is different from K-means clustering and it can be pointed out that this is because of the absence of clear distinction between data points.
- Average Linkage: The average linkage suggests that the appropriate number of clusters is 4 or 5 and this analysis is also in sync with K-means clustering.
- It can be concluded that both K-means and Hierarchical clustering suggest 4 as the appropriate number of classes with the exception of Hierarchical clustering with single linkage suggesting 8 or 9 classes.

2.2. Validation by Classification

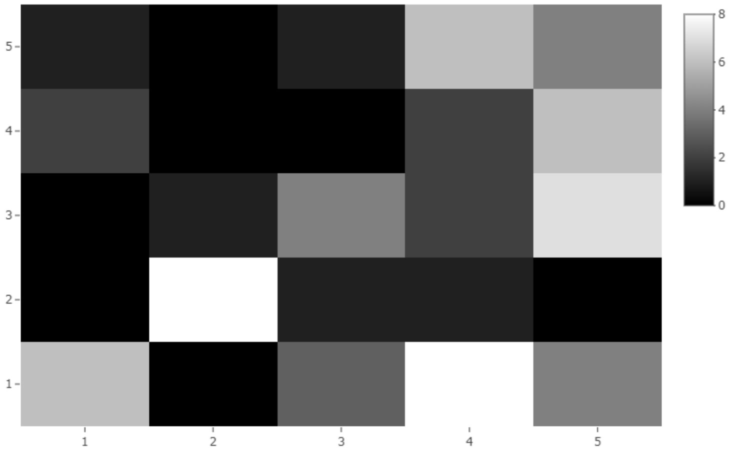
- By performing the above clustering analysis we have derived two important insights:
  1. The appropriate number of classes based on the distance between two points is 4 which is less than the number of classes defined in the observed data.
  2. The decision boundary between classes of observed data has more linear component than radial component.
- We have validated the first insight by using different techniques of clustering. To validate the second insight we've performed logistic regression, Linear-kernel SVM, and Radial-kernel SVM. The heat map of confusion matrix of the aforementioned models is shown below.



(Figure 11: Confusion Matrix for Logistic Regression Predictions)



(Figure 12: Confusion Matrix for Linear kernel SVM)



(Figure 13: Confusion Matrix for Radial kernel SVM)

Model	Accuracy
Logistic Regression	42.27%
Linear Kernel SVM	47.76%
Radial Kernel SVM	35.82%

(Table 1: Accuracy of classification models)

- It can be seen from the above heat maps that the radial kernel SVM has more misclassifications than linear kernel SVM and Logistic Regression.
- From the accuracies in Table 1 and observation around confusion matrix it can be concluded that the accuracy of the models with linear decision boundary is better than that of models with radial decision boundary. This validates our second insight.

3. CONCLUSION

- Given a multi-class classification dataset we performed Clustering Analysis to draw a few insights into the distribution of classes and nature of Bayesian Decision Boundary for classification.
- The given dataset has 5 labelled classes but clustering analysis suggests that approximately 4 classes can be best used to explain variance. This insight is validated by k-means clustering and hierarchical clustering with average and complete linkages however, hierarchical clustering with single linkage suggests 8-9 classes.
- The above clustering analysis indicates the possibility that the Bayesian decision boundary for classification has less radial component than expected. Upon plotting the data points, it can be observed that the Bayesian decision boundary may have more linear component than the radial component.
- To validate the above insight, we performed classification analysis to create Logistic Regression, Linear kernel SVM, and Radial kernel SVM models and test the performance. Upon comparing the performance of these models on our dataset, our insight was easily validated.

4. REFERENCES

1. Plotly Official Documentation - https://plotly.com/r/
2. UCI Repository of Datasets - https://archive.ics.uci.edu/dataset/763/land+mines-1
3. James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). Introduction. In: An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-0716-1418-1_1