## Abstract

Ensemble Learning is a popular Machine Learning technique that combines multiple models to achieve better accuracy and generalization than individual models. This project aims to investigate the use of Ensemble Learning to improve the accuracy of Cluster Analysis. Traditional cluster algorithms may have limitations when dealing with datasets that have multiple attributes, some of which may not be related. By applying Ensemble Learning techniques, we hope to improve the clustering accuracy and identify hidden patterns in the data.

The proposed project involves selecting a dataset that has 8 or more attributes, with related and non-related attributes, and has been shown to have low accuracy with traditional clustering algorithms. We will then apply Ensemble Learning methods to the dataset to improve the clustering accuracy. To validate our model, we will compare the results with those obtained from various datasets and traditional clustering algorithms.

The role of Ensemble Learning in cluster analysis has significant importance in data analysis. It can improve the accuracy of clustering algorithms and identify hidden patterns in the data that may not be visible with traditional methods. Therefore, the success of our project can have a significant impact on improving the accuracy of clustering analysis in real-world applications.

## Introduction

Cluster analysis is a widely used data exploration technique that groups similar data points together based on their similarities. It plays a crucial role in pattern recognition, data mining, and information retrieval. However, traditional clustering algorithms may face challenges when dealing with datasets that contain multiple attributes, some of which might not be directly related to the clusters. This can lead to less accurate clustering results and obscure meaningful patterns present in the data.

Ensemble Learning offers a solution to these challenges by combining multiple clustering algorithms to create a stronger and more robust model. The ensemble model can effectively mitigate the weaknesses of individual algorithms and improve overall clustering accuracy. By applying Ensemble Learning to cluster analysis, we aim to achieve better insights and identify hidden structures within the data.

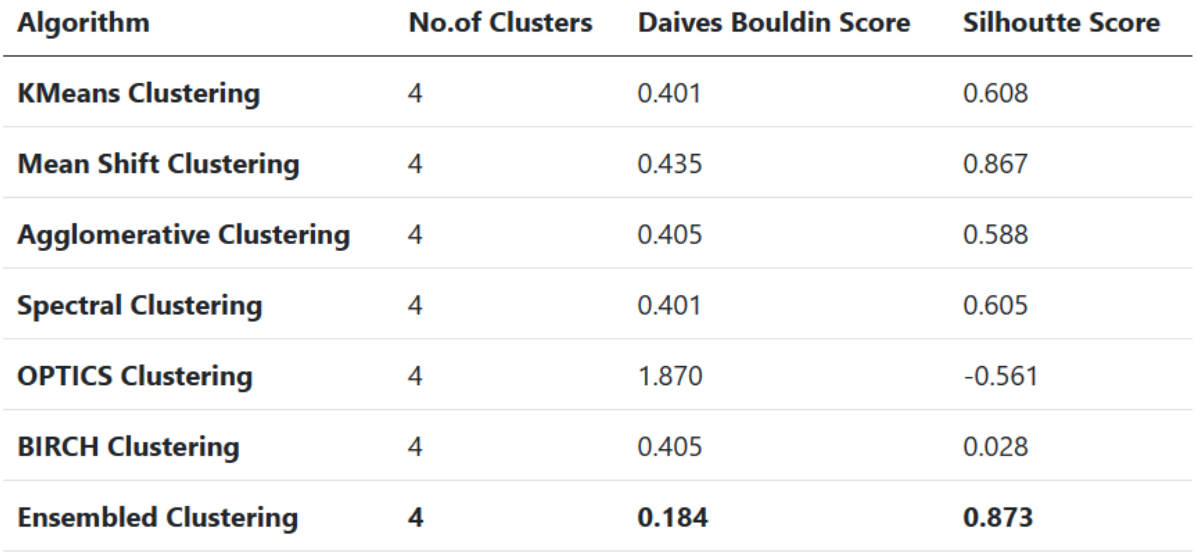
## Objectives

The primary objectives of this project are as follows:

- Investigate the use of Ensemble Learning techniques for improving clustering accuracy in datasets with diverse attributes.
- Select a suitable dataset with 8 or more attributes, including both related and non-related attributes.
- Apply various Ensemble Learning methods to the dataset and compare their performance.
- Validate the accuracy improvement of the ensemble model against traditional clustering algorithms.
- Identify hidden patterns and meaningful clusters within the data using the ensemble approach.

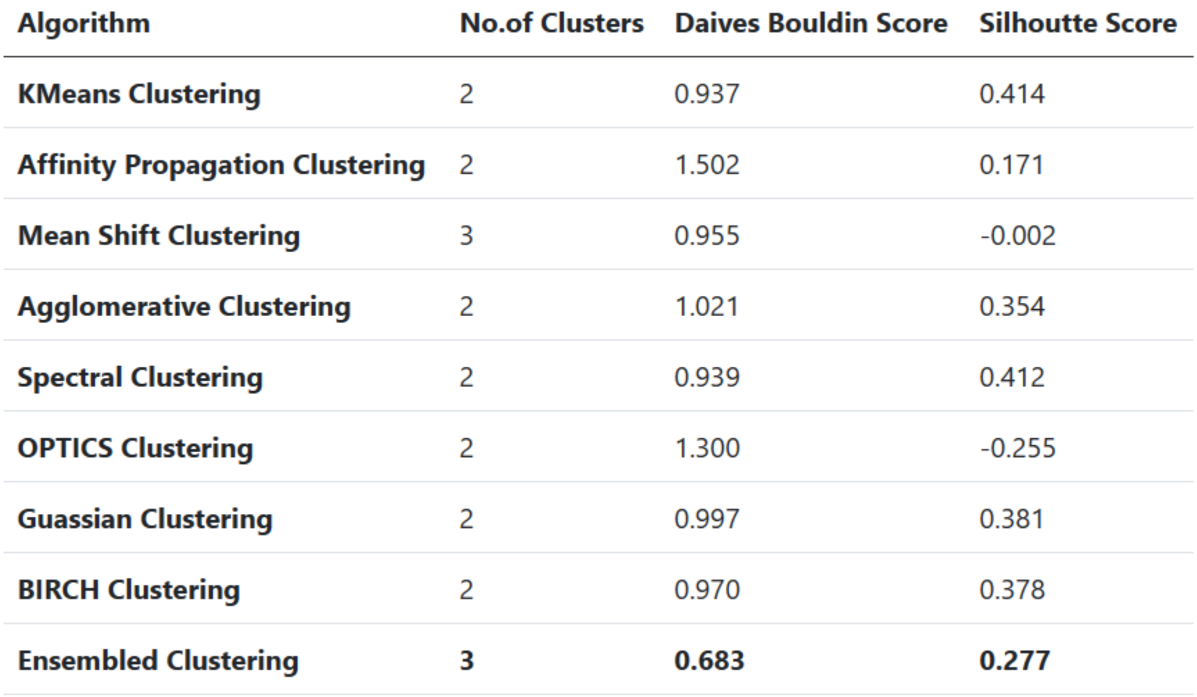
## Results

### Weather History Dataset



| Algorithm | No.of Clusters | Daives Bouldin Score | Silhoutte Score |
|---|---|---|---|
| KMeans Clustering | 4 | 0.401 | 0.608 |
| Mean Shift Clustering | 4 | 0.435 | 0.867 |
| Agglomerative Clustering | 4 | 0.405 | 0.588 |
| Spectral Clustering | 4 | 0.401 | 0.605 |
| OPTICS Clustering | 4 | 1.870 | -0.561 |
| BIRCH Clustering | 4 | 0.405 | 0.028 |
| Ensembled Clustering | **4** | **0.184** | **0.873** |

## Weather Prediction Dataset



| Algorithm | No.of Clusters | Daives Bouldin Score | Silhoutte Score |
|---|---|---|---|
| KMeans Clustering | 2 | 0.937 | 0.414 |
| Affinity Propagation Clustering | 2 | 1.502 | 0.171 |
| Mean Shift Clustering | 3 | 0.955 | -0.002 |
| Agglomerative Clustering | 2 | 1.021 | 0.354 |
| Spectral Clustering | 2 | 0.939 | 0.412 |
| OPTICS Clustering | 2 | 1.300 | -0.255 |
| Guassian Clustering | 2 | 0.997 | 0.381 |
| BIRCH Clustering | 2 | 0.970 | 0.378 |
| Ensembled Clustering | **3** | **0.683** | **0.277** |

## Conclusion

- The voting technique of ensembling with the Mean Shift and Birch clustering algorithm yielded a higher silhouette score and lower Davies-Bouldin score in the analysis of the Weather Prediction dataset and Weather History dataset.

- The **lower Davies-Bouldin score** highlights distinct and meaningful clusters with minimal overlap and high inter-cluster similarity, supporting the identification of homogeneous groups within the dataset.
- The **higher silhouette score** indicates well-separated and closely-knit data points within each cluster, showcasing the successful capture of inherent structures and patterns in the weather data.
- The effectiveness of the ensembling approach with the Mean Shift and Birch clustering algorithm demonstrates its value in improving clustering accuracy and robustness.
- Overall, these insights enhance decision-making, data exploration, and understanding of the underlying structures in the Weather datasets.