CS777 Project

Airport Data Analysis and Predictive Modeling

Author

Zhang Xizhao

Date

10/17/2023

Introduction

The objective of this project is to analyze a dataset from openflights containing information about various airports around the world. The dataset includes features such as latitude, longitude, altitude, and timezone, and attached below. I aim to perform clustering to understand the geographical distribution of these airports and also attempt classification to predict the type of each airport.

Airport ID	Unique OpenFlights identifier for this airport.
Name	Name of airport. May or may not contain the **City** name.
City	Main city served by airport. May be spelled differently from **Name**.
Country	Country or territory where airport is located. See Countries to cross-reference to ISO 3166-1 codes.
IATA	3-letter IATA code. Null if not assigned/unknown.
ICAO	4-letter ICAO code. Null if not assigned.
Latitude	Decimal degrees, usually to six significant digits. Negative is South, positive is North.
Longitude	Decimal degrees, usually to six significant digits. Negative is West, positive is East.
Altitude	In feet.
Timezone	Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5.
DST	Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown). *See also: Help: Time*
Tz database time zone	Timezone in "tz" (Olson) format, eg. "America/Los_Angeles".
Type	Type of the airport. Value "airport" for air terminals, "station" for train stations, "port" for ferry terminals and "unknown" if not known. *In airports.csv, only type=airport is included.*
Source	Source of this data. "OurAirports" for data sourced from OurAirports, "Legacy" for old data not matched to OurAirports (mostly DAFIF), "User" for unverified user contributions. *In airports.csv, only source=OurAirports is included.*

Methodology

Data Preprocessing

Encoded the 'Type' column to numerical labels for classification. Removed rows with missing values in key columns like Latitude, Longitude, Altitude, Timezone and label.

Machine Learning Models

Two clustering models were used: K-Means and Gaussian Mixture Model (GMM). These models were run three times using different features: *Timezone*, *Altitude*, and a combination of *Latitude and Longitude*. For each time the optimal number of clusters for both model was

determined to be three by comparing the Silhouette Score.

Logistic Regression and Random Forest models were used for classifying the airports based on their types. These models were trained on features including *Latitude, Longitude, Timezone, and Altitude*.

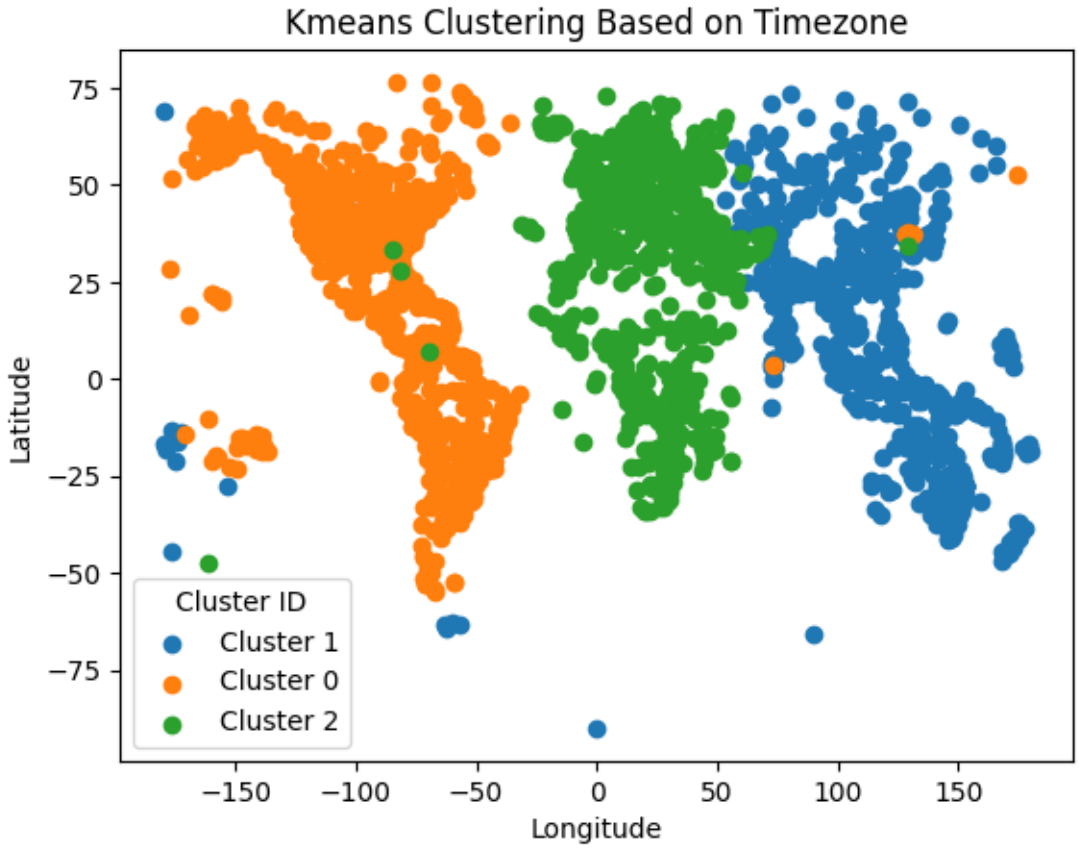
Results

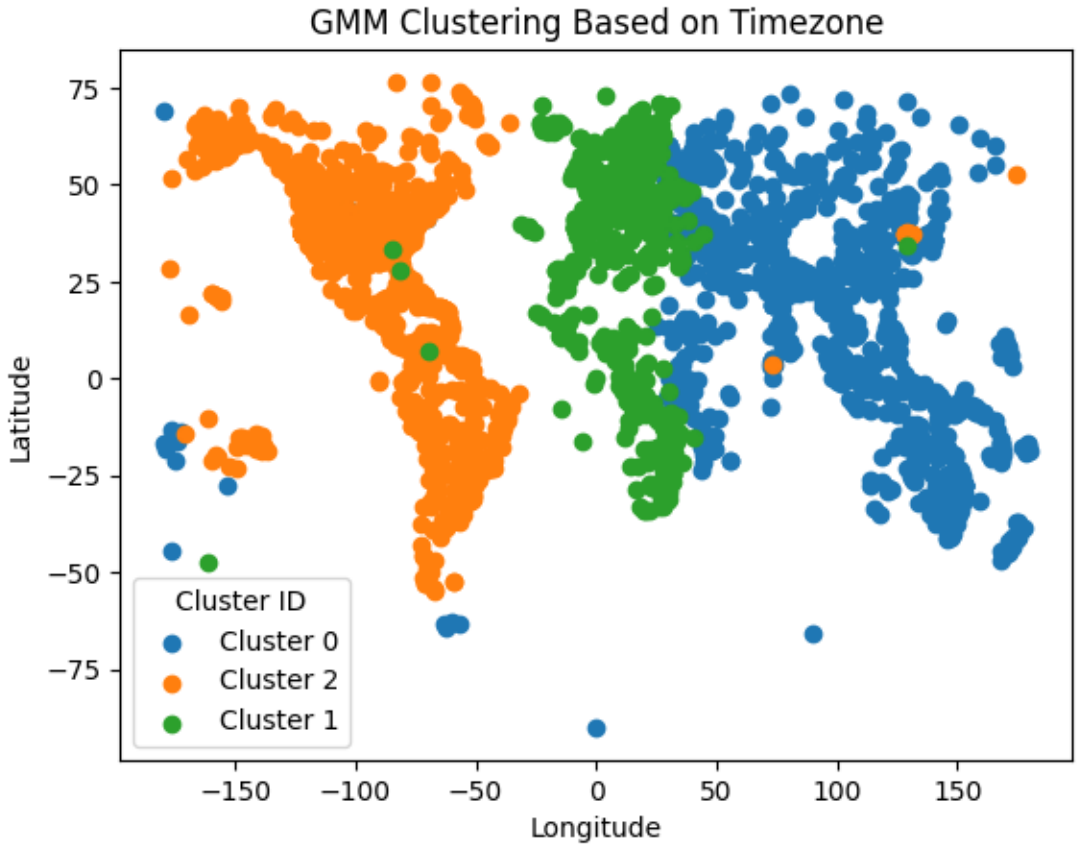
Silhouette Score was used for clustering models and accuracy for classification models.

Timezone:

K-Means: Train (0.868), Test (0.859)

GMM: Train (0.740), Test (0.746)

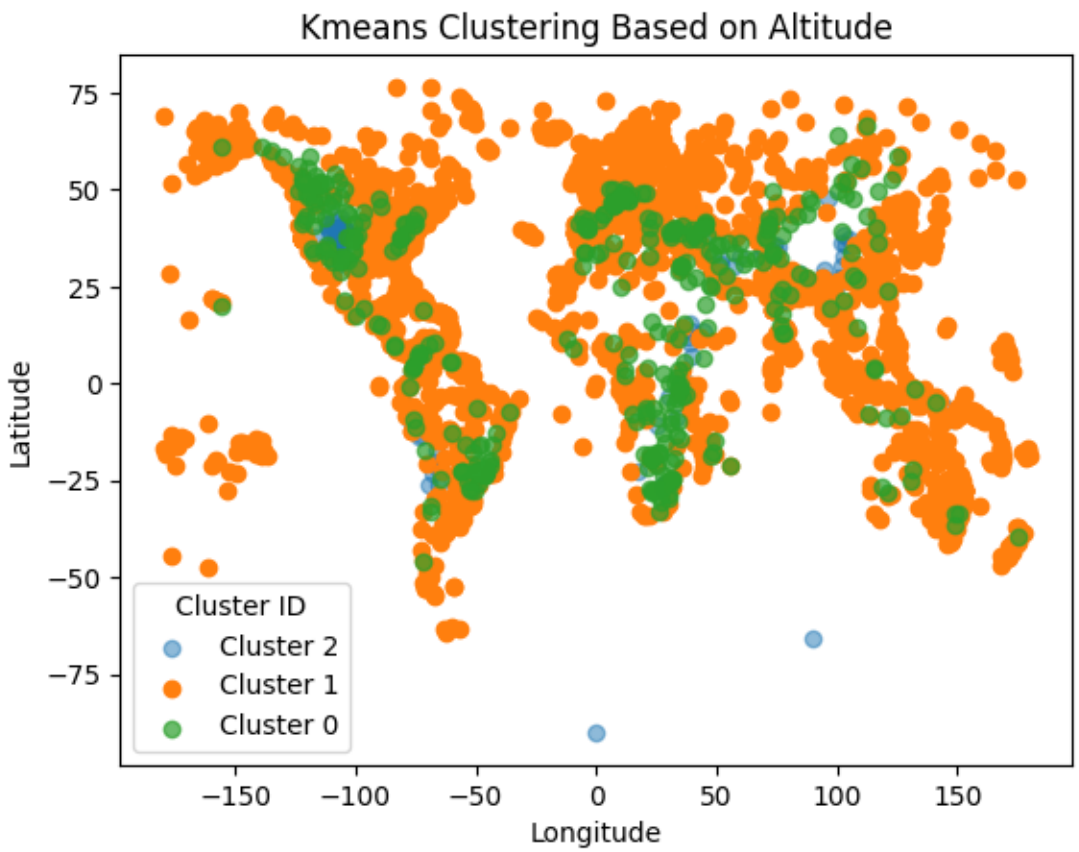


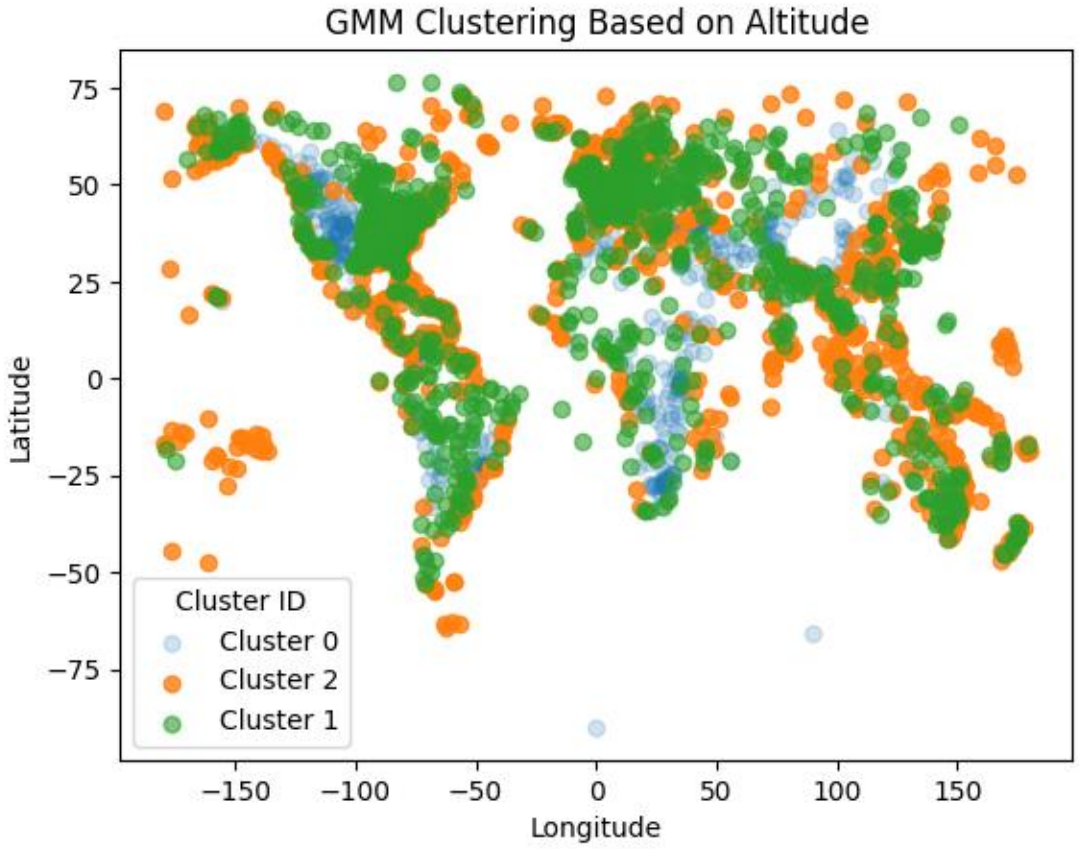


Altitude:

K-Means: Train (0.871), Test (0.875)

GMM: Train (0.368), Test (0.391)

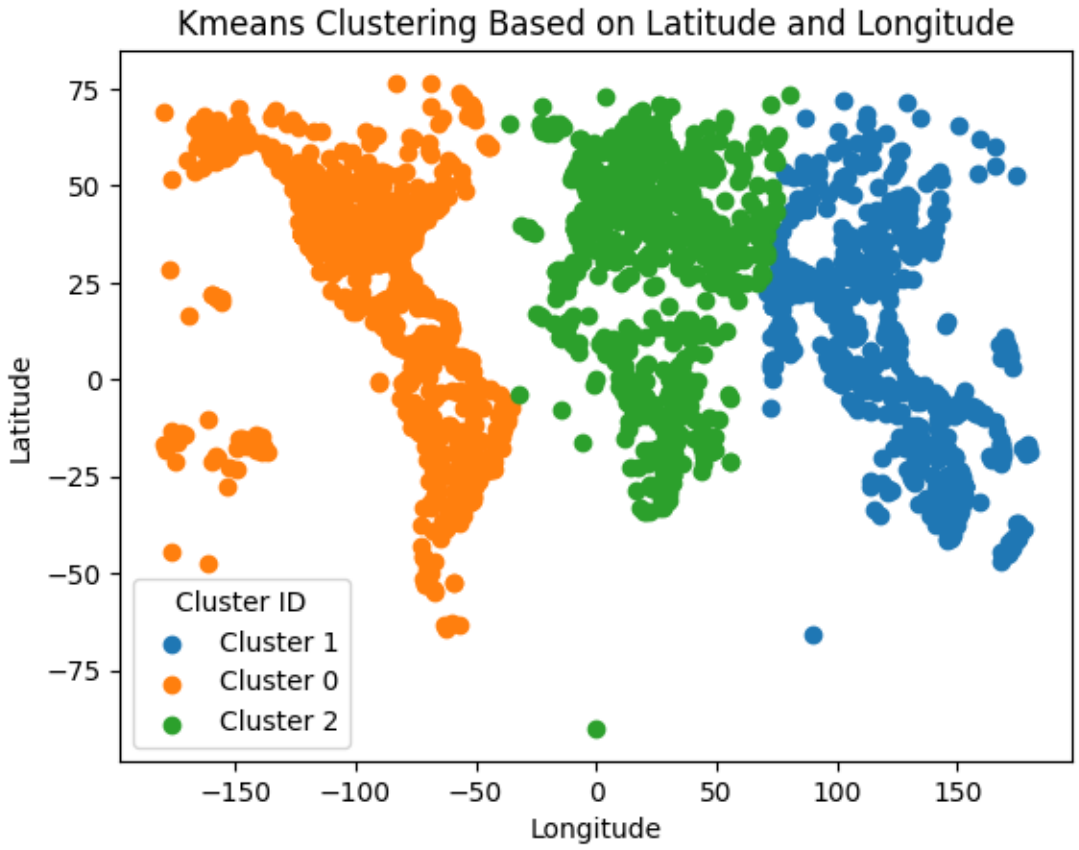

Kmeans Clustering Based on Altitude
Latitude
Longitude
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
Cluster ID
Cluster 2
Cluster 1
Cluster 0

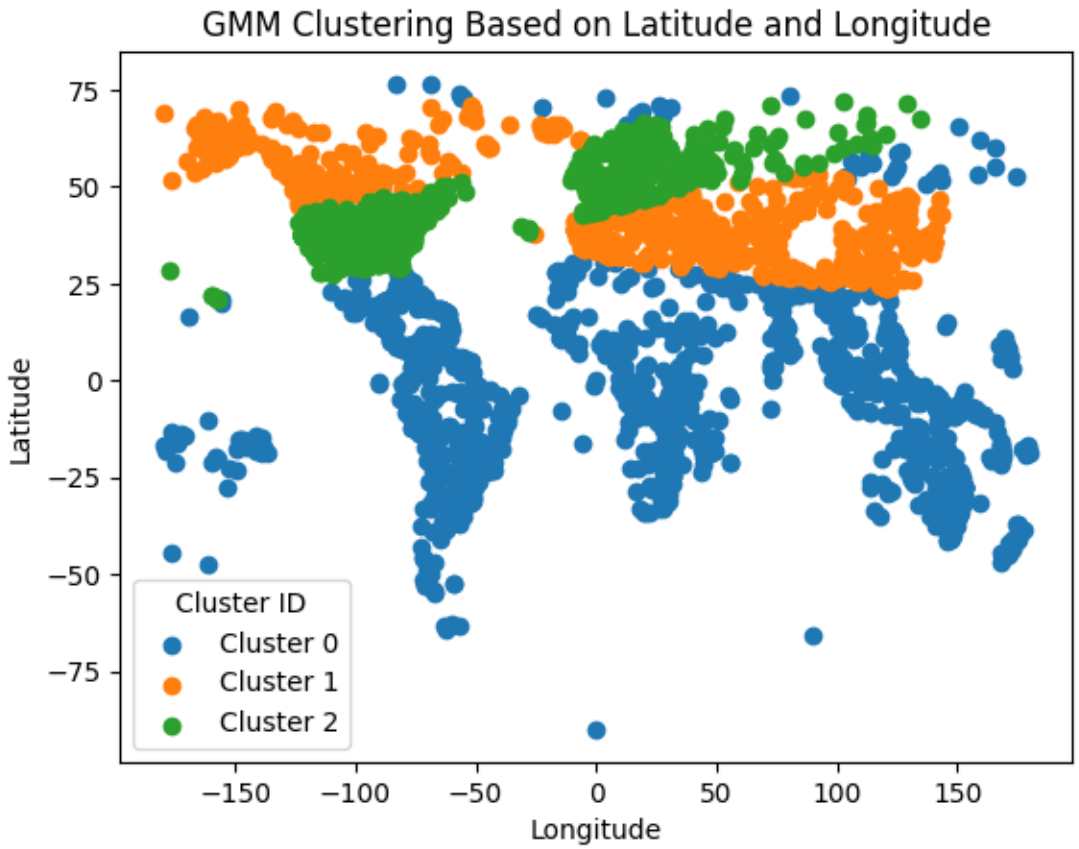

GMM Clustering Based on Altitude
Latitude
Longitude
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
Cluster ID
Cluster 0
Cluster 2
Cluster 1

Latitude & Longitude:

K-Means: Train (0.774), Test (0.761)

GMM: Train (0.081), Test (0.072)





Logistic Regression: 0.739

Random Forest: 0.762

Discussion

The K-Means algorithm demonstrated consistent and high Silhouette Scores across all feature sets, suggesting its robustness for this clustering task. On the other hand, GMM showed significant variability in its performance, indicating its sensitivity to the choice of features. Both K-Means and GMM generated similar geographical clusters when using "Timezone" and "Altitude" as features. This suggests that these features have a strong geographical correlation. Using a combination of "Latitude" and "Longitude," K-Means produced a clustering pattern

similar to when "Timezone" was used as a feature. In contrast, GMM generated a more fragmented and less interpretable clustering pattern, effectively "slicing" the map into irregular shapes. The discrepancies in GMM's performance could be attributed to its probabilistic nature, which may not align well with the geographical distribution of the data points when using a combination of "Latitude" and "Longitude."

Conclusion

This project aimed to explore the effectiveness of clustering and classification algorithms in categorizing airports based on various features such as "Timezone," "Altitude," "Latitude," and "Longitude." The K-Means and Gaussian Mixture Model (GMM) were used for clustering, while Logistic Regression and Random Forest were used for classification tasks. Various features such as "Timezone," "Altitude," "Latitude," and "Longitude" were considered to understand their impact on the model's performance and the interpretability of the results.

One of the key findings was that K-Means consistently outperformed GMM in terms of silhouette scores, regardless of the feature set used. This may suggest that K-Means is a more robust algorithm for this specific application. On the other hand, both K-Means and GMM yielded similar and meaningful clusters when the features were limited to "Timezone" and "Altitude." However, the use of "Latitude" and "Longitude" produced divergent results; while K-Means generated reasonable clusters, GMM's clusters were less interpretable. It also reveals that GMM's performance is highly sensitive to the choice of features, unlike K-Means, which remained relatively stable.

In the classification tasks, Random Forest showed a slightly better performance with an accuracy of 0.762, compared to Logistic Regression, which had an accuracy of 0.739. This indicates that Random Forest may be a more suitable model for classifying airport types based on the features considered in this study.