SOUND OF ML

Music Genre and Popularity Prediction and Classification Using Machine Learning

Jack, Abdul, Talia

AGENDA

- The Story
- Data Collection
- Questions:
  - Clustering Genres (Clustering)
  - Prediction of Genres (Classification)
  - Prediction of Popularity (Regression)
- Results

- Our system classifies new songs into specific genres based on their characteristics.
- Mislabeled or unlabeled songs can also be properly classified using our system.
- We clustered genres based on song characteristics using unsupervised learning.
- Our system can also predict the popularity of a song.

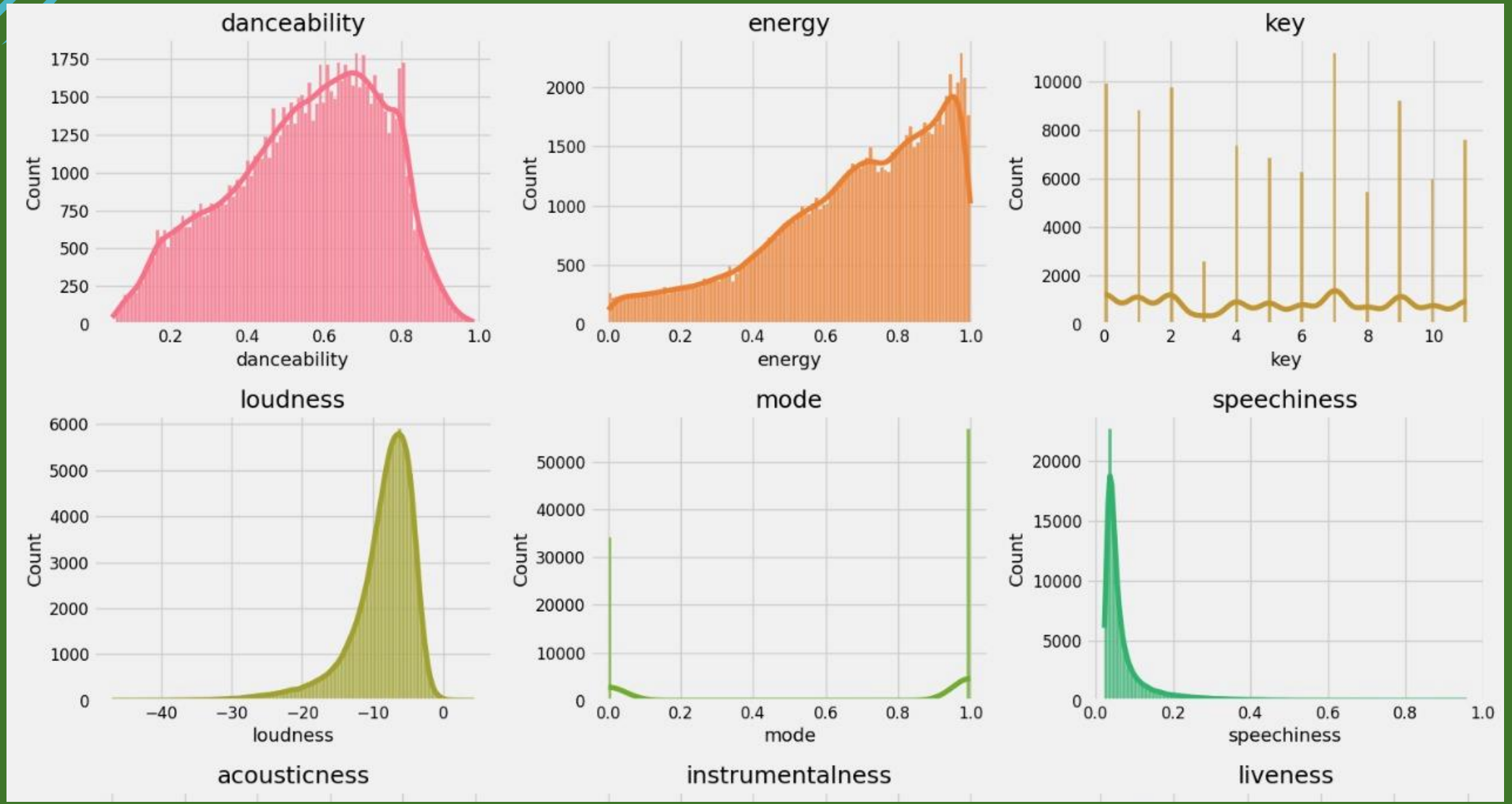
The story

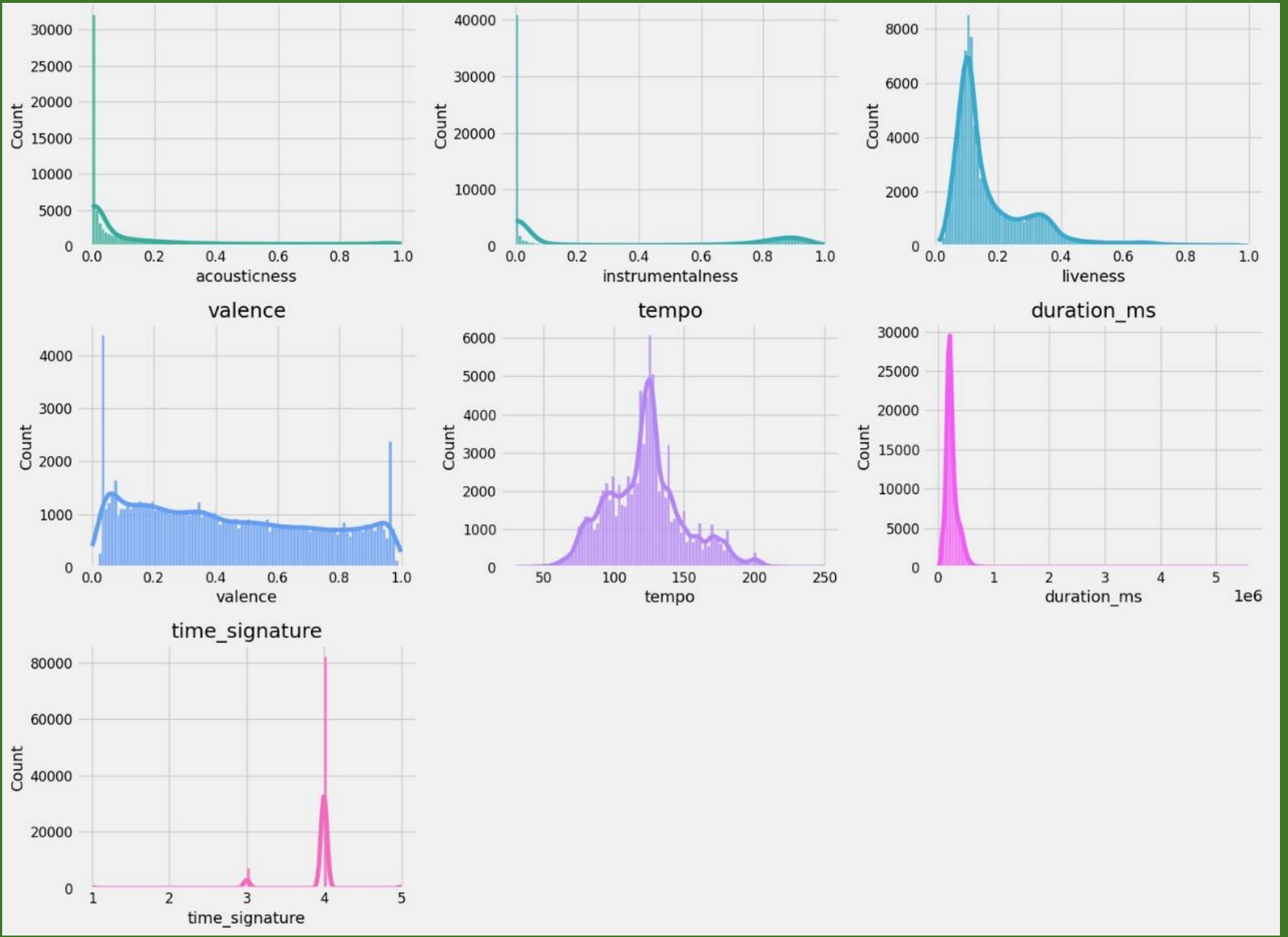
DATA COLLECTION

- We queried data from the Spotify API to assemble a dataset of ~500,000 unique tracks.
- The tracks spanned 132 genres and about 50,000 artists.- This gave us a dataset of different metrics compiled by Spotify for every track, including a popularity score, a genre, and audio features.
- Our audio features include musical characteristics like a key, whether major/minor chords were dominant, tempo, and length.
- We also have scores assigned by Spotify, like the songs' "danceability", "energy," "liveness" (which measures presence of crowd noise), and "speechiness" (which measures the amount of human vocals in the track)

DATA CLEANING

- Because of the way the data was collected from the API, there were a significant amount of duplicate songs, which we removed.
- We identified loudness, speechiness, liveness, tempo, and duration as having a large number of outliers as defined by IQR.
- We investigated these outliers to see if there was any bad data.
- We found that the songs which were outliers in a large number of features were actually just weird songs. Because this was legitimate data, we kept it.


danceability
Count
danceability
energy
Count
energy
key
Count
key
loudness
Count
loudness
mode
Count
mode
speechiness
Count
speechiness
acousticness
instrumentalness
liveness


Count
acousticness
instrumentalness
liveness
valence
tempo
duration_ms
1e6
time_signature

QUESTIONS?

How can the clustering of song genres together based on common characteristics be useful in-terms of identifying broader song categories?

How is it possible to predict the genre of a song only based on its audio characteristics and title?

How can the song popularity be predicted based on common characteristics of a song?

CLUSTERING GENRES

MODELS USED

Kmeans	Hierarchical
• Scalers • PCA components • Num clusters	• Affinities • Num clusters

Gaussian Mixture Models	Spectral Clustering (KNN)
• Covariances • Num clusters	• Num Neighbors • Num Clusters

NOVEL CLUSTERING METHODS

Gaussian Mixture Models

Spectral Clustering

METRICS USED

Inertia

Calinski-Harabasz

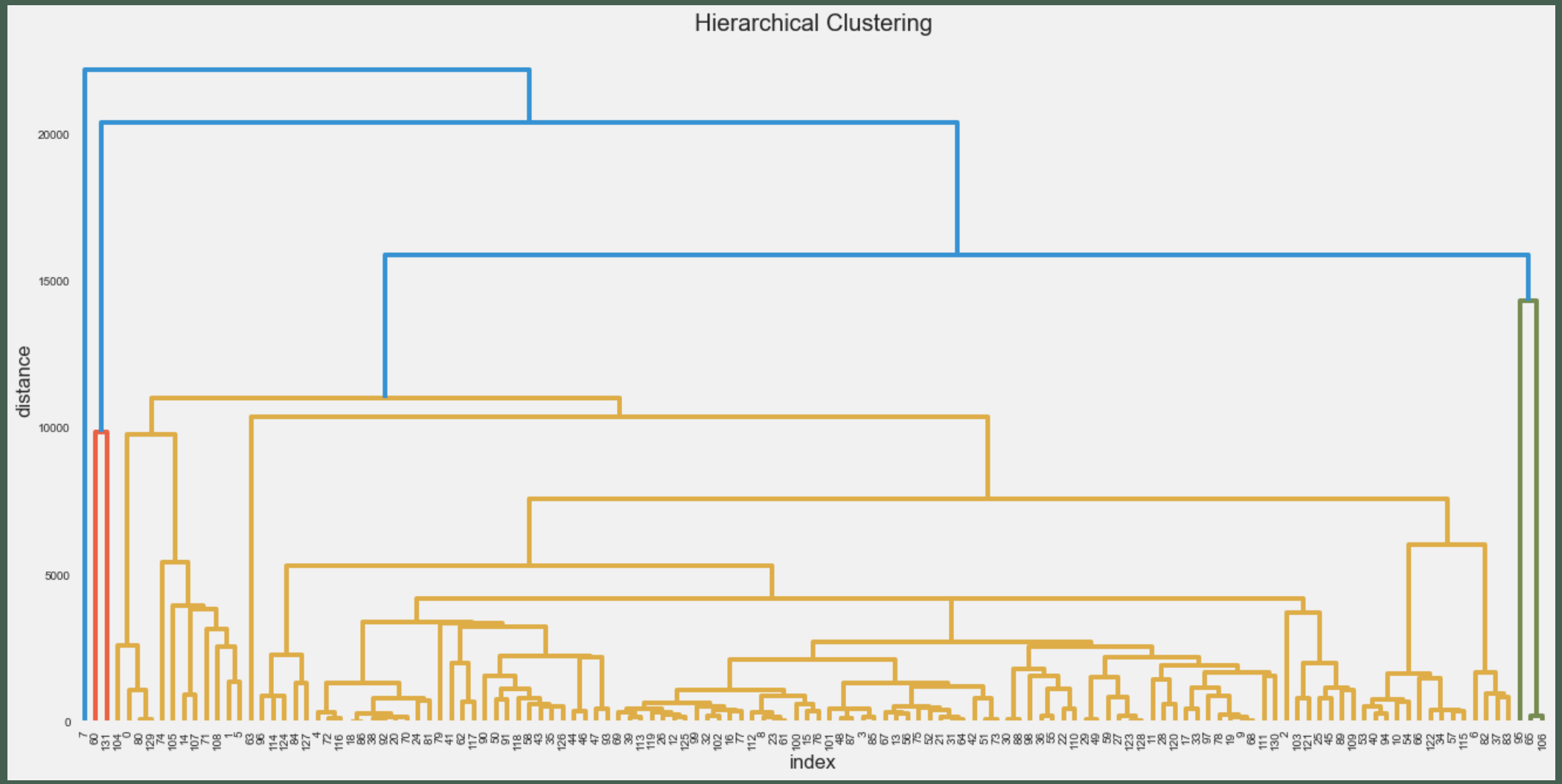
Silhouette

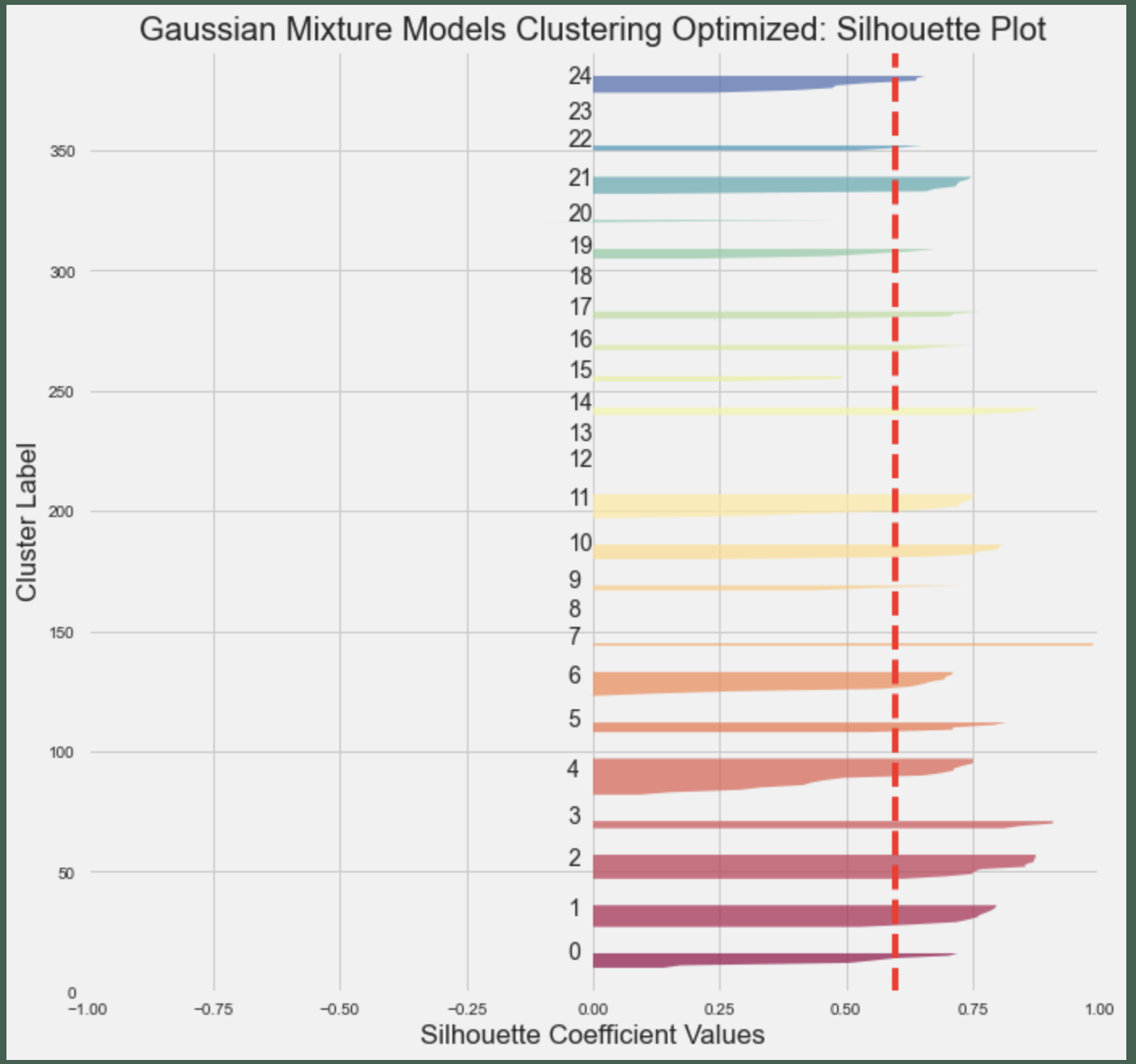
Davies-bouldin

RESULTS

Hierarchical

general model	specific model	inertia	calinski-harabasz	silhouette	davies-bouldin
KMeans	Kmeans_baseline	4714891416.236173	805.575336	0.55764	0.483585
KMeans	Kmeans_optimal	1.112751	176.382386	0.398355	0.626794
Hierarchical Clustering	Hierarchical Clustering Baseline	-	54.591585	0.355454	0.255994
Hierarchical Clustering	Hierarchical Clustering Optimal	-	594.579145	0.484443	0.243415
GMM	GMM_Baseline	-	791.625524	0.590777	0.413616
GMM	GMM Optimal	-	2765.156409	0.595639	0.340321

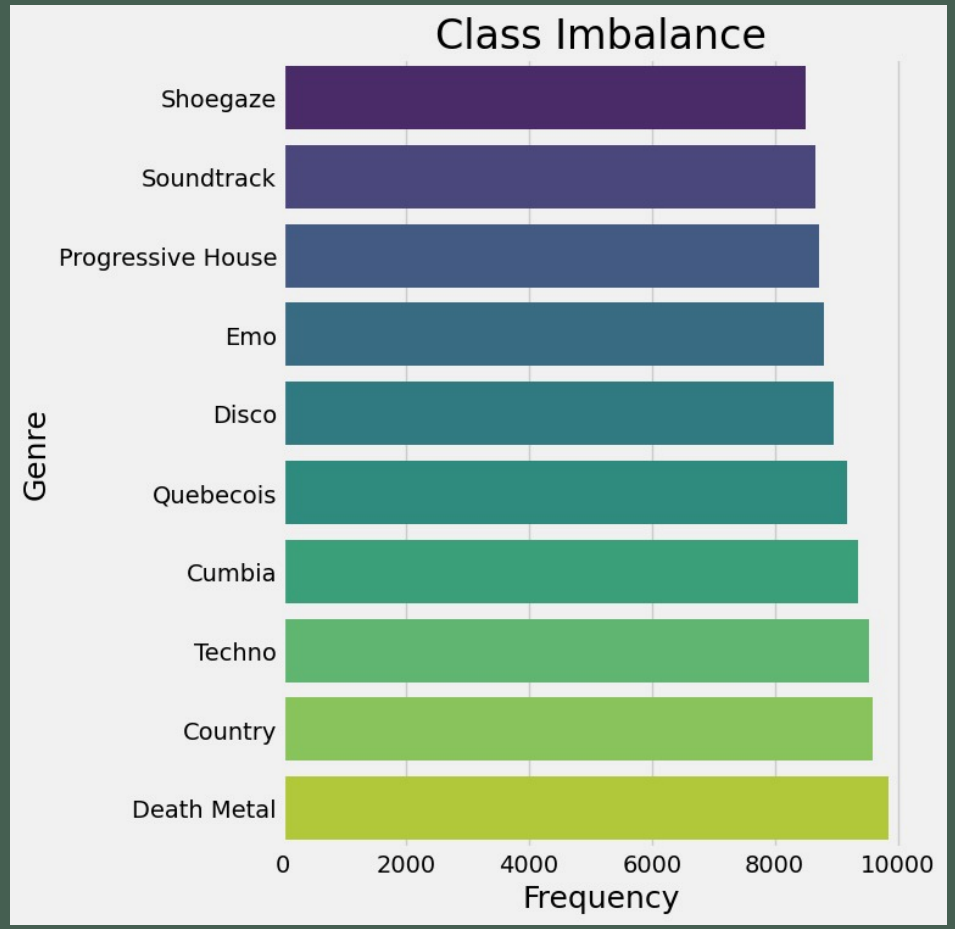

Hierarchical Clustering
distance
20000
15000
10000
5000
0
index


Gaussian Mixture Models Clustering Optimized: Silhouette Plot
Cluster Label
Silhouette Coefficient Values
0
50
100
150
200
250
300
350
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0

PREDICTING GENRES

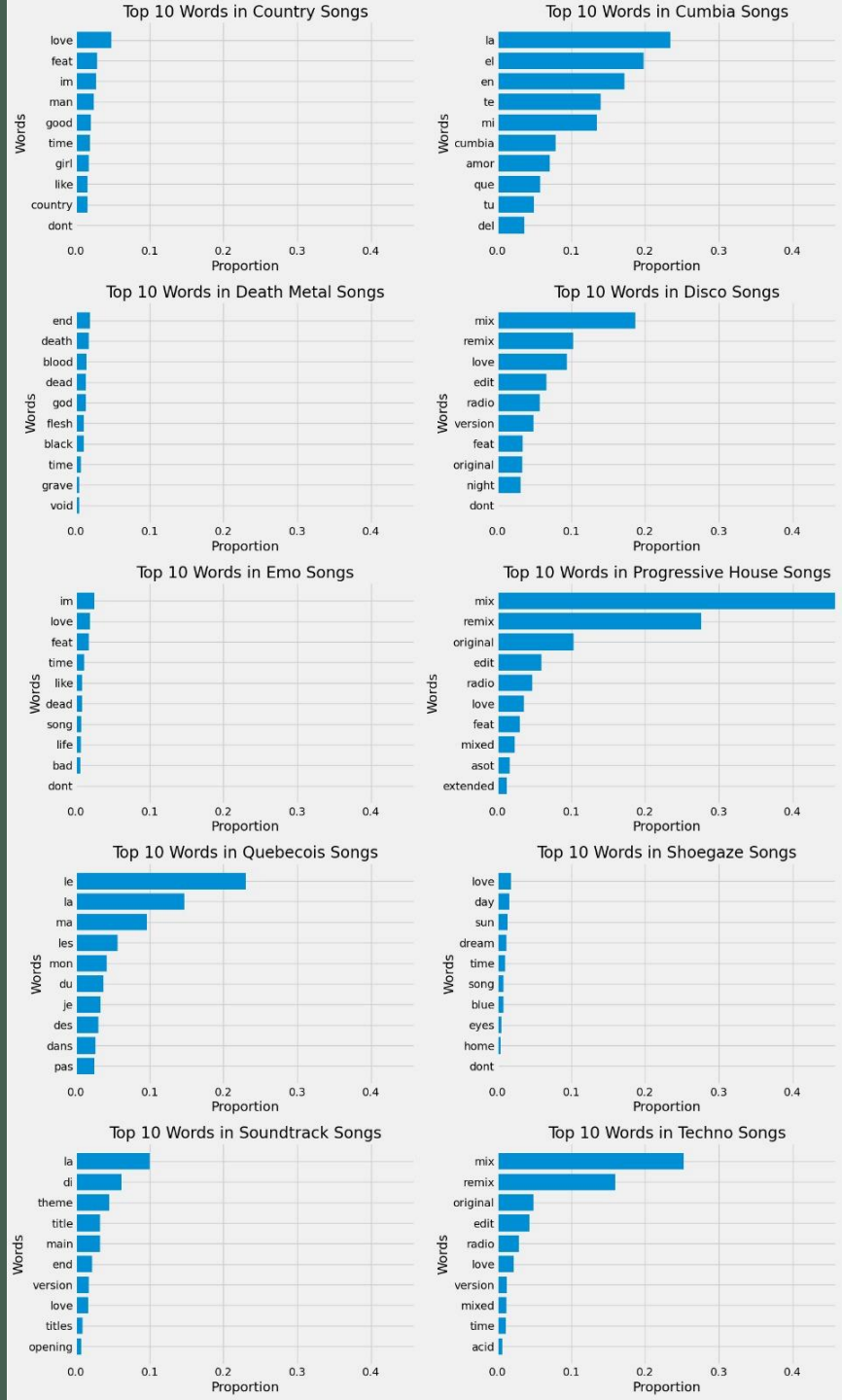
REDUCING CLASSES

- Our initial dataset contained 132 genres. We trained some initial models, which achieved a low accuracy of about 35%.
- We then limited genres to the 10 most represented in our dataset, which are displayed here.
- These genres represent a diverse range more or less typical of our dataset, while also presenting some challenges in terms of similar genres like Progressive House and Techno.



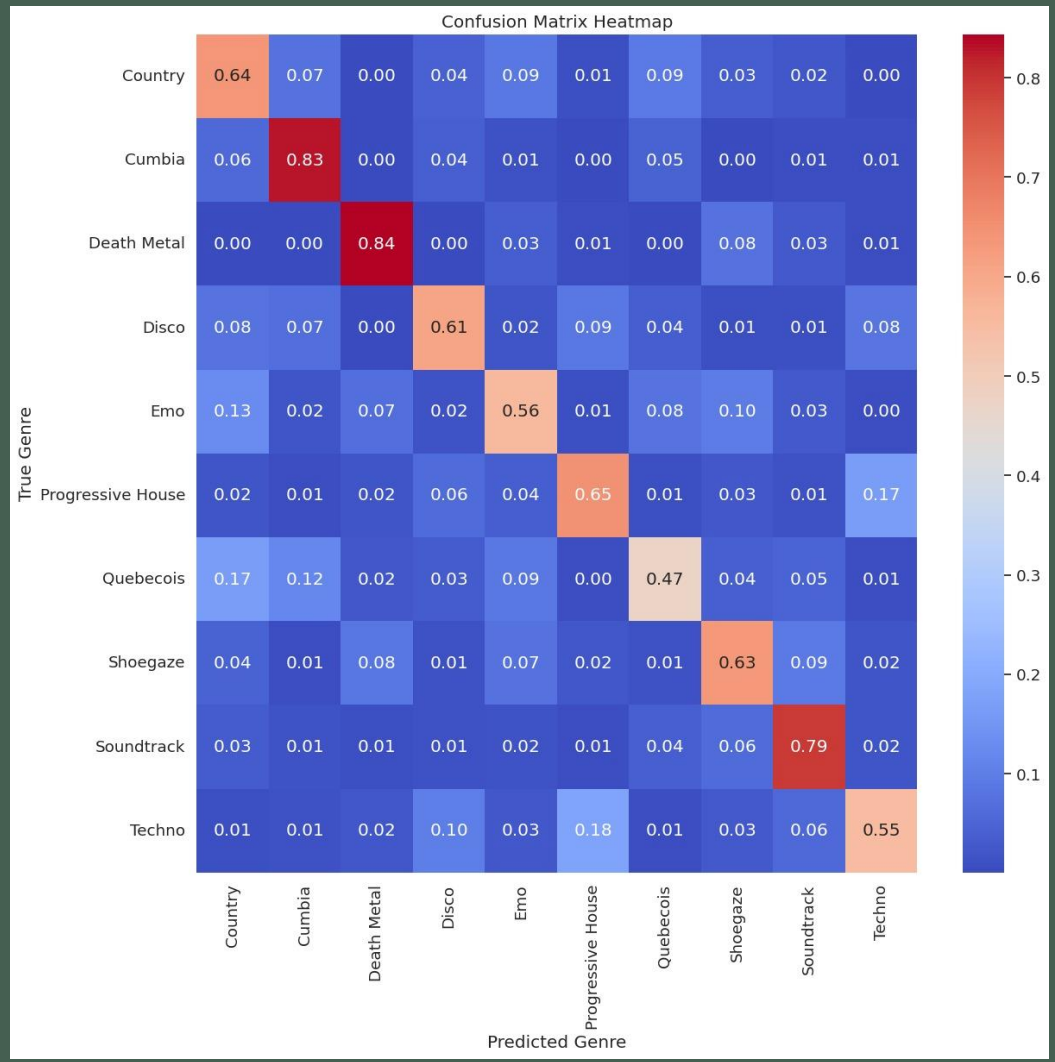
FEATURE ENGINEERING

- For each genre, we calculated a Term Frequency – Inverse Document Frequency ranking, which is a measure that combines the frequency of a token with its rarity across the whole corpus.
- We used this to generate 100 additional features, 10 top tokens for each genre. These features were then calculated as binary indicator variables across the whole dataset.
- We removed English stopwords, but left in Spanish and French stopwords on the theory that it might increase our ability to separate out non-Anglophone genres like Cumbia and Quebecois.



MODELING

- We trained four classifiers: KNN, Logistic Regression, Random Forest, and Light Gradient Boosting Machine (LightGBM).
- The two ensemble, tree-based models, Random Forest and LightGBM, performed about the same, and achieved 67% accuracy.
- Precision and recall were within 10% for most classes for most models. However, between KNN, regression, and the ensemble models, there was significant difference in which genres had higher precision or recall, indicating high sensitivity to model choice.



PREDICTING POPULARITY

FEATURE ENGINEERING

Genre Encoding

Popularity Bins

Numerical Features

PREDICTING POPULARITY BASED ON SONG CHARACTERISTICS

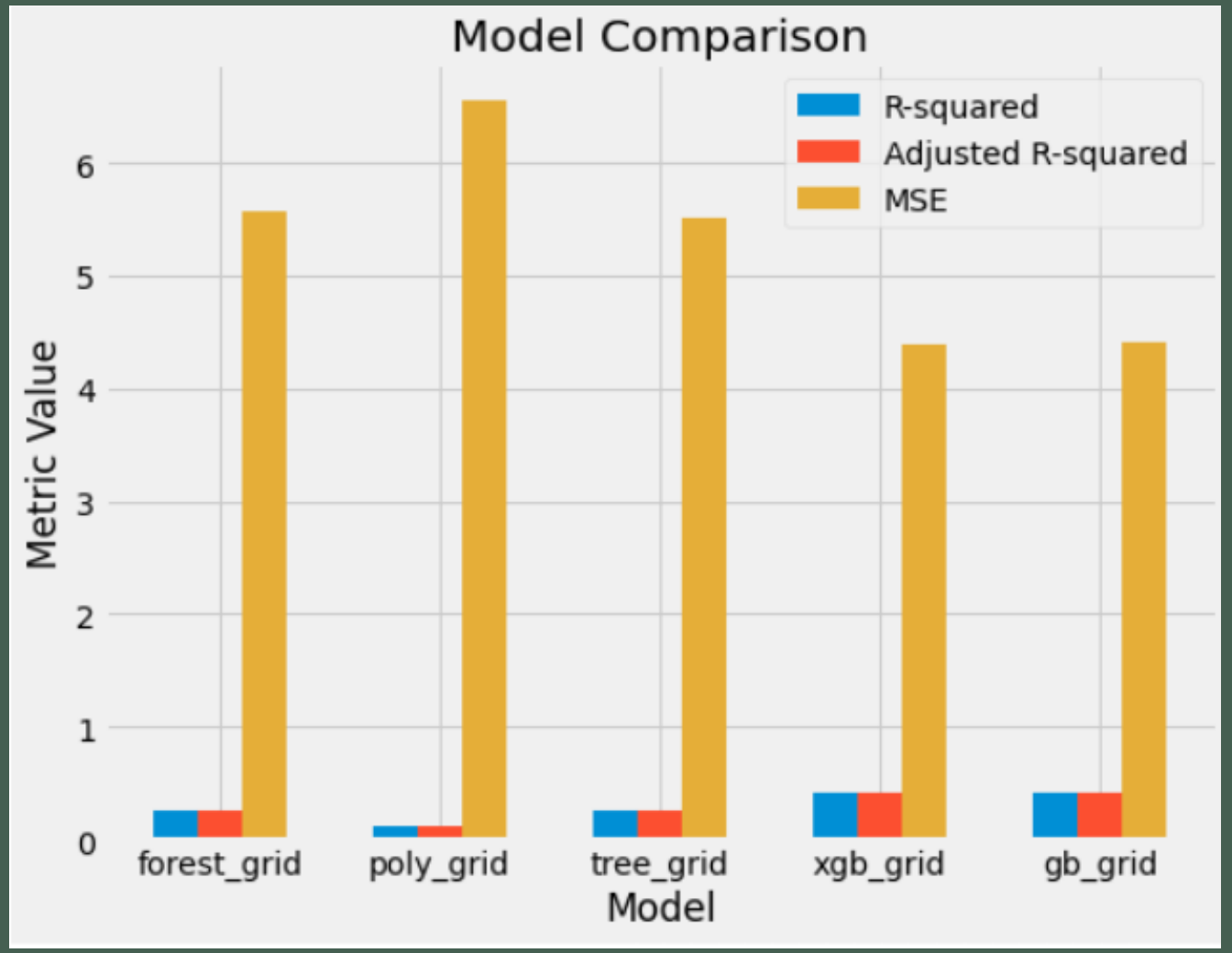
Polynomial Regression	Random Forest Regression	Decision Tree Regression	XGboost Regression	Gradient Boosting Regression
• R-squared: 12.6% • Adjusted R-squared: 12.5% • MSE: 7.45	• R-squared: 26.9% • Adjusted R-squared: 26.8% • MSE: 6.27	• R-squared: 29.3% • Adjusted R-squared: 29.2% • MSE: 6.03	• R-squared: 39.5% • Adjusted R-squared: 39.4% • MSE: 5.16	• R-squared: 39.7% • Adjusted R-squared: 39.6% • MSE: 5.14

MODELS AND METRICS

Polynomial Regression	Random Forest Regression	Decision Tree Regression	XGboost Regression	Gradient Boosting Regression
• R-squared: 12% • Adjusted R-squared: 11.9% • MSE: 6.55	• R-squared: 25.2% • Adjusted R-squared: 25.1% • MSE: 5.57	• R-squared: 26% • Adjusted R-squared: 26% • MSE: 5.5	• R-squared: 40.9% • Adjusted R-squared: 40.9% • MSE: 4.39	• R-squared: 40.2% • Adjusted R-squared: 40.8% • MSE: 5.14

IMPORTANT FEATURES

Random Forest Regression	Decision Tree Regression	XGboost Regression	Gradient Boosting Regression
• Genre • Danceability • Duration_ms	• Genre • Instrumentalness • Duration_ms	• Genre • Instrumentalness • Danceability	• Genre • Instrumentalness • Danceability


Model Comparison
R-squared
Adjusted R-squared
MSE
Metric Value
0
1
2
3
4
5
6
forest_grid
poly_grid
tree_grid
xgb_grid
gb_grid
Model

CONCLUSION