Crafting BeatWave's Genre Tapestry:

An Iterative Approach to Classification Modeling

Data classification modeling is a machine learning technique that involves categorizing data points to predefined ‘classes’ based on their features and characteristics.

The Goal

Through the iterative process of data classification modeling, we can begin to sift through popular electronic songs and give BeatWave a better understanding of which sonic metrics predict which genres.

This will allow for a streamlined approach to classifying titles into their appropriate categories.

The Dataset

Data Source:

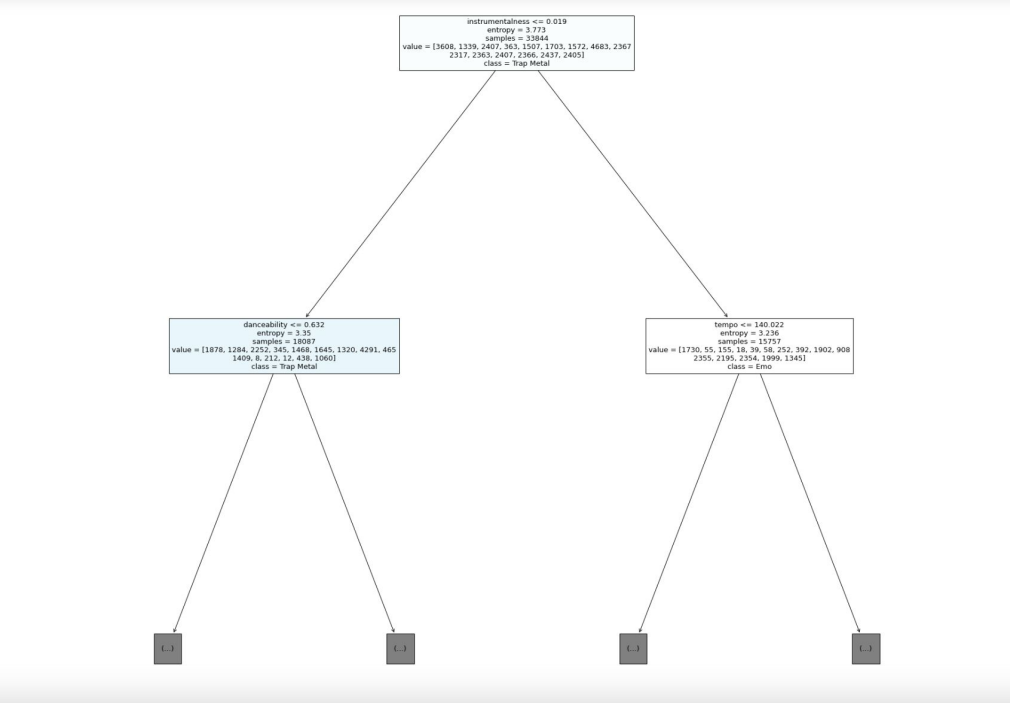
https://www.kaggle.com/datasets/mrmorj/dataset-of-songs-in-spotify

We are using an open source dataset that displays sample data from popular streaming platform Spotify.

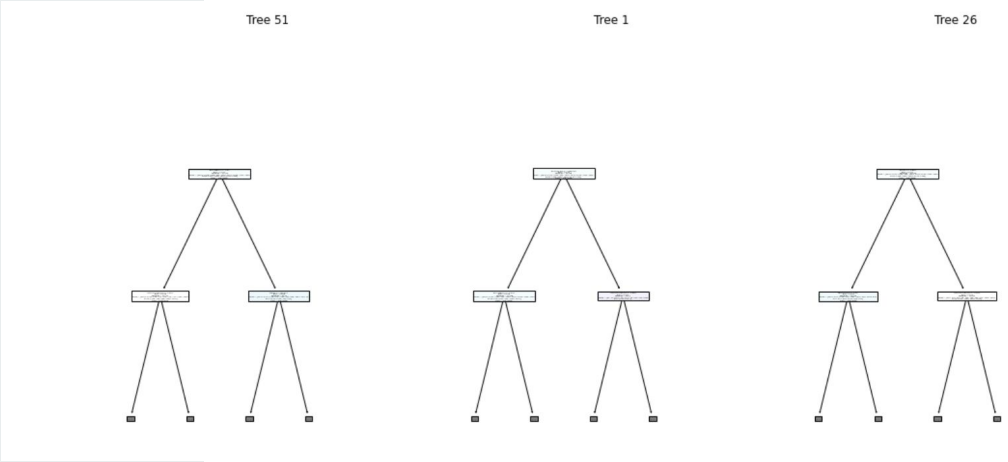
Using this data will help us gauge our place in competition with the world's leading streaming service.

Decision Tree

Useful for tasks like genre classification of music due to their intuitive nature and ability to handle complex decision boundaries.



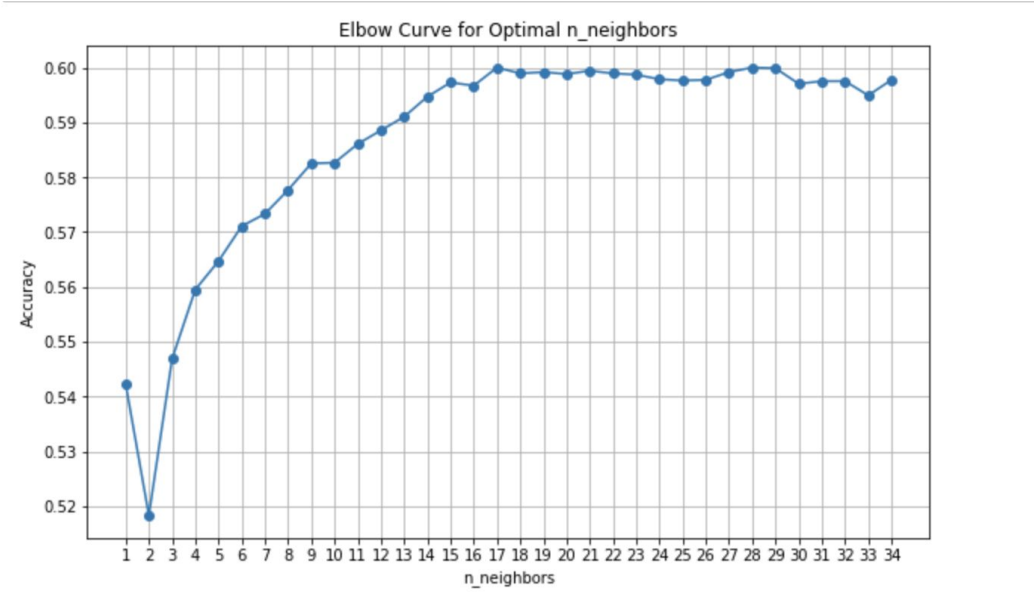
Random Forest



Each decision tree in the ensemble, or ‘forest’, is constructed using a random subset of the training data and features to reduce ‘overlearning.’

k-Nearest Neighbors

Operates on the principle that similar data points tend to belong to the same class. For a given data point, k-NN identifies the (k) nearest neighbors in the training dataset and predicts the class label by majority vote.



Precision, Recall & Accuracy

- **Precision** is the ratio of true positive predictions to the total number of positive predictions, which helps assess the accuracy of positive predictions.
- **Recall**, also known as True Positive Rate (TPR) or Sensitivity, measures the proportion of *correctly predicted* positive instances out of all *actual* positive instances, while taking into a False Negative Rate (FNR).
- **Accuracy** measures the overall correctness of predictions by calculating the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset.

Random Forest Metrics

Random Forest Accuracy: 0.6696607965961471

	precision	recall	f1-score	support
Dark Trap	0.54	0.45	0.49	970
Emo	0.72	0.74	0.73	341
Hiphop	0.42	0.39	0.41	621
Pop	0.25	0.08	0.12	98
Rap	0.48	0.32	0.39	341
RnB	0.41	0.36	0.38	396
Trap Metal	0.36	0.27	0.31	384
Underground Rap	0.39	0.53	0.45	1192
dnb	0.96	0.98	0.97	599
hardstyle	0.89	0.93	0.91	619
psytrance	0.94	0.93	0.93	598
techhouse	0.87	0.90	0.88	568
techno	0.87	0.85	0.86	590
trance	0.82	0.90	0.86	562
trap	0.86	0.86	0.86	582
accuracy			0.67	8461
macro avg	0.65	0.63	0.64	8461
weighted avg	0.67	0.67	0.66	8461

Next Steps

Find or compile a much larger dataset, ensuring that majority of data points contain data for all features to avoid class imbalances.

Invest in feature engineering, like feature interactions, aggregate features, temporal patterns, harmonic/melodic features, and data transformations