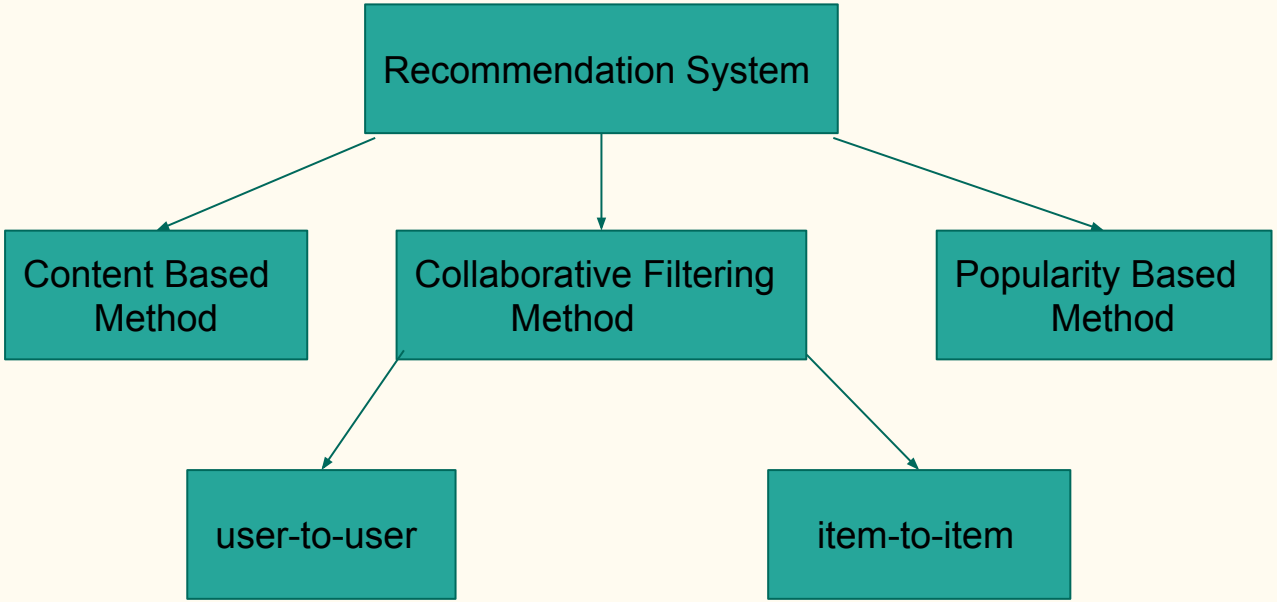
Team 9

Spotify Music Recommendation

K
rina Shah

Raj Shah
Vaishnavi Shirsath

What is Recommendation System?

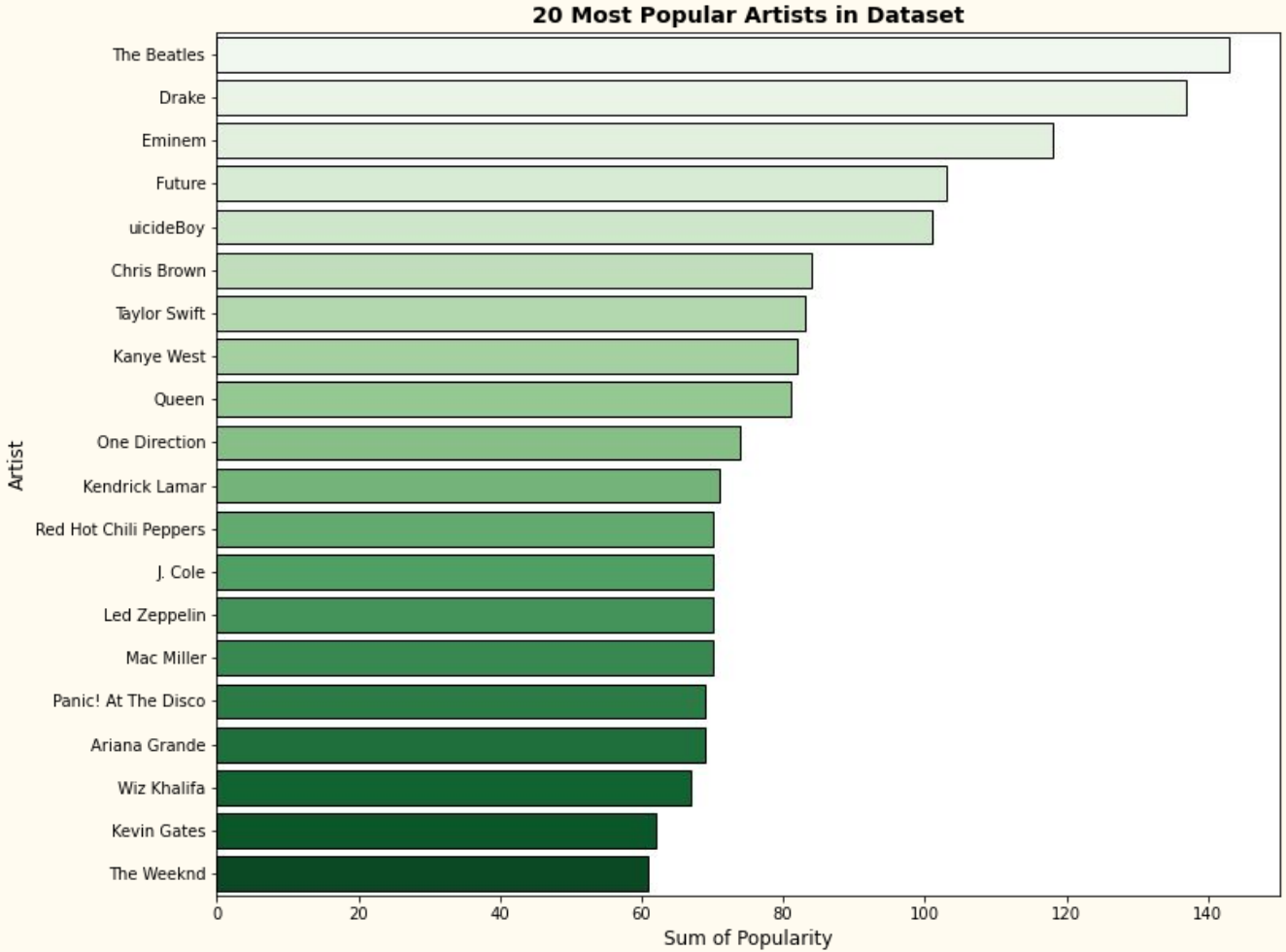


Data Cleaning

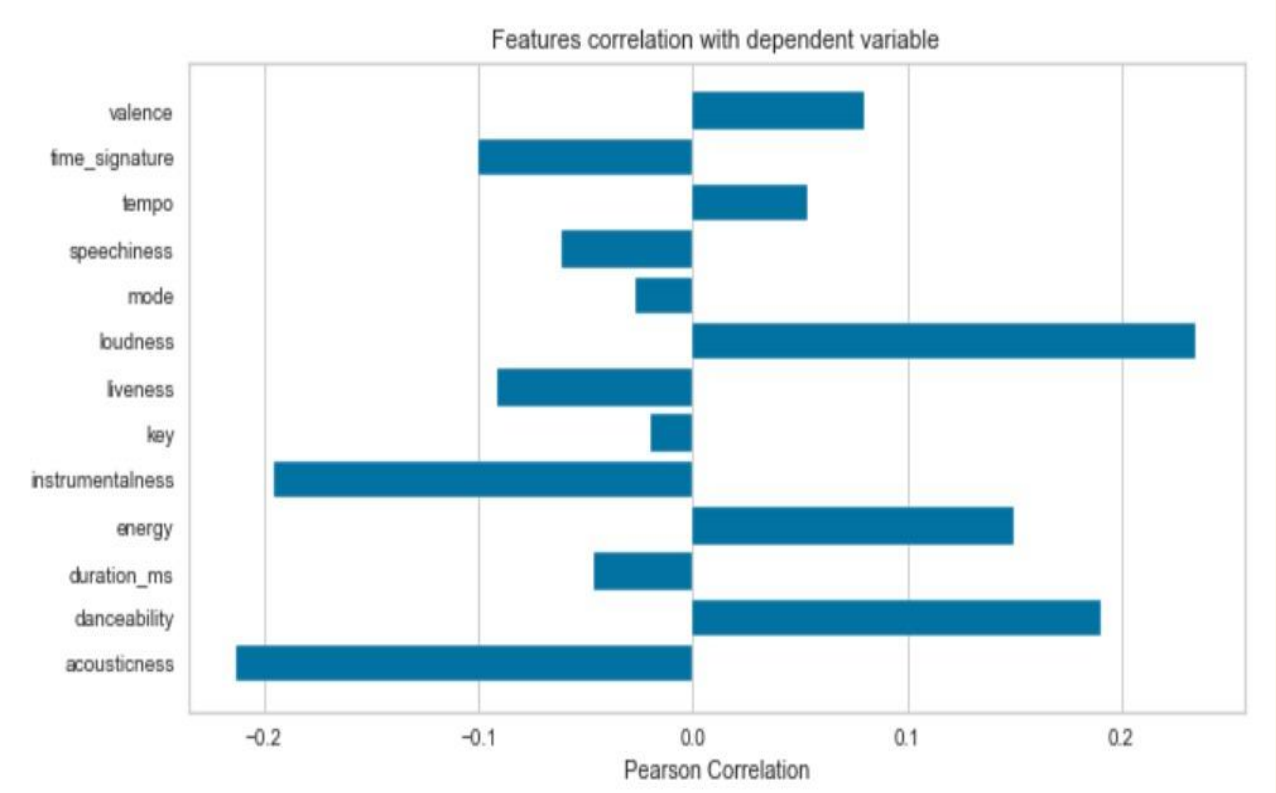
- Checking for Null values
- Removing duplicates
- Removing Special Characters
- Converting Character data into numerical
- Converting continues value to discrete value

	artist_name	track_name	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	mode	key	time_signature	speechiness	tempo	valence
84288	Ryuichi Sakamoto	Oto	0	0.947	0.184	229933	0.104	0.946	0.0989	-23.917	Minor	C#	3/4	0.0457	80.063	0.137
202964	Ryuichi Sakamoto	Oto	0	0.947	0.184	229933	0.104	0.946	0.0989	-23.917	Minor	C#	3/4	0.0457	80.063	0.137
228145	Ryuichi Sakamoto	Oto	0	0.947	0.184	229933	0.104	0.946	0.0989	-23.917	Minor	C#	3/4	0.0457	80.063	0.137

EDA


20 Most Popular Artists in Dataset
The Beatles
Drake
Eminem
Future
uicideBoy
Chris Brown
Taylor Swift
Kanye West
Queen
One Direction
Kendrick Lamar
Red Hot Chili Peppers
J. Cole
Led Zeppelin
Mac Miller
Panic! At The Disco
Ariana Grande
Wiz Khalifa
Kevin Gates
The Weeknd
Artist
0
20
40
60
80
100
120
140
Sum of Popularity

Correlation



Popularity Based Recommendation

Decision Tree

- The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).
- A decision tree construction is concerned with identifying splitting attributes and splitting criterion at every level of the tree.
- Accuracy : 0.68

K -Nearest Neighbor

- K stands for the number of data set items that are considered for classification.
- K-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure in this case distance between closely grouped points.
- Accuracy : 0.68

XGBoost

- XGBoost is a decision-tree based ensemble Machine Learning algorithm that uses a gradient boosting framework.
- XGBoost provides a parallel tree boosting that solves many data science problems in a fast and efficient way.
- Accuracy : 0.75

Random Forest Classifier

- The random forest is a classification algorithm consisting of many decisions trees.
- It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.
- Accuracy : 0.76

Cosine Similarity

$$cosine(x, y) = \frac{x \cdot y^T}{||x|| \cdot ||y||}$$

Cosine similarity is a method used to compute the similarity between two vectors.

We have used this method of SVD to find similarity between new unheard songs and already heard and rated popular songs to recommend new songs to user.

This method gives us similarity (cos theta) of new songs with popular songs. We then sort these values in descending order and return top 5 similar songs.

Target popularity of recommended songs from test set:

```
Top 5 new tracks for users with their similatirity to popular songs :

1. Powers  ,Target: 1

2. On And On  ,Target: 1

3. Two Tickets to Paradise  ,Target: 1

4. Tricks on Me  ,Target: 1

5. I Wanna Love You - Album Version (Edited)  ,Target: 0
```

Content Based Recommendation

Model 1: Naive Bayes

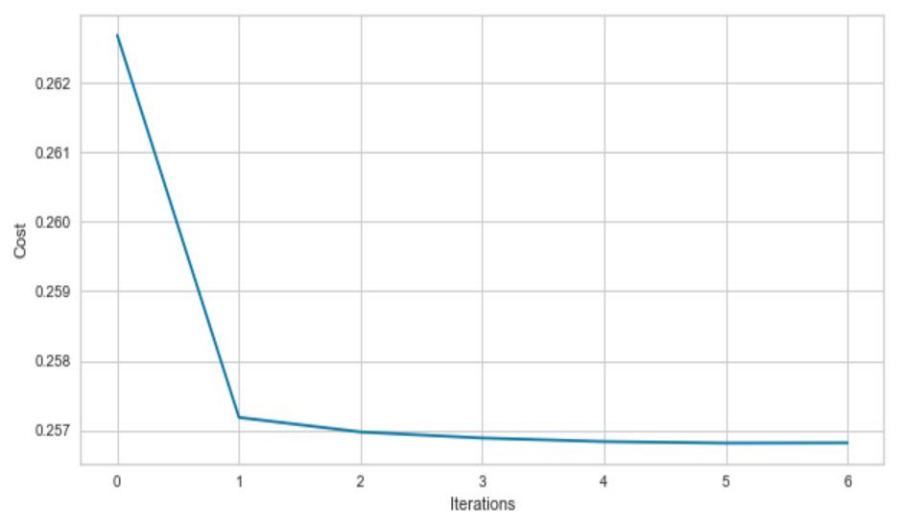
$$\textbf{Naive Bayes classifier: } v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

- It is not only a simple approach but also a fast and accurate method for prediction.
- Naive Bayes has very low computation cost.
- It can efficiently work on a large dataset.
- No other learning model outperforms this model on average.
- Assuming attribute values are conditionally independent given target value.

Model 2: Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a technique that is used to find the minima of a function.

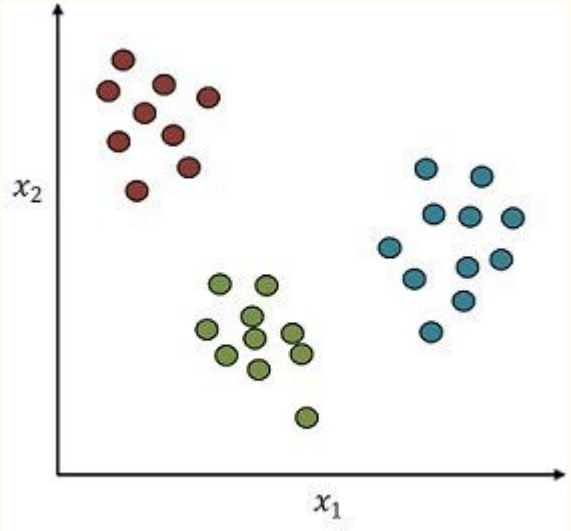
SGDClassifier a linear classifier (by default in SKlearn it is a linear SVM) that uses SGD for training.

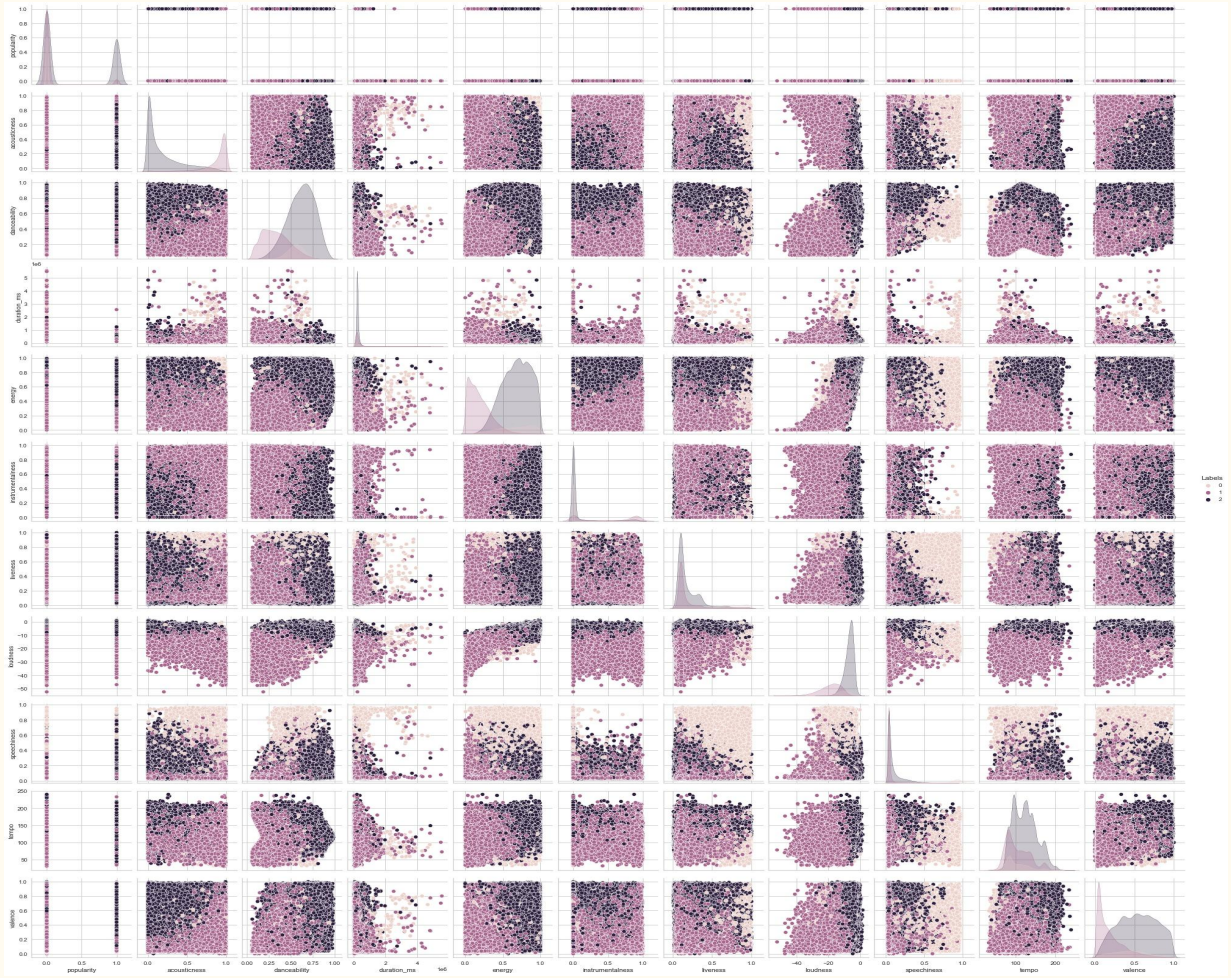


Collaborative Based Recommendation

K Means Clustering algorithm

- K-Means is an unsupervised machine learning algorithm which can be used to categorize data into different group
- Choosing value of k by **hyperparameter tuning**.
- Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
- Assign each data point to the closest cluster (centroid)
- Keep iterating until there is no change to the centroids. I.e assignment of data points to clusters isn't changing.
- Used **silhouette_score**




popularity
acousticness
danceability
duration_ms
energy
instrumentalness
liveness
loudness
speechiness
tempo
valence
Labels
0
1
2

Conclusion :

	Model	Accuracy	F1_Score
5	SGDClassifier	0.886953	0.779912
0	RandomForestClassifier	0.756132	0.406650
3	XGBClassifier	0.751487	0.324076
4	NaiveBayes	0.715059	0.626741
1	KNeighborsClassifier	0.683659	0.280199
2	DecisionTreeClassifier	0.682814	0.423579

Future Works :

In this project we have implemented a Recommendation system for recommending songs using spotify's dataset.

For future we could use this information to suggest the artist what features of the songs makes for a popular track.

We can also use Hybrid methods to improve recommendations of new songs to user.

Zoom Link :

https://stevens.zoom.us/rec/play/MiAJmTU6xdjKYmCGyQ6dLKV3P3O1Aff9os7OmMvU79aMHHTvuBEYz07G0MxObBsrdC5eVYcQAMf_f1no.ty74mQTWiZ-MJZJI?continueMode=true&_x_zm_rtaid=hZA83YI-RA-BUFev07yxPA.1638968792723.d81a5995d4a30d0bef7e74a77041739b&_x_zm_rhtaid=432&startTime=1638967626000

Thank You