KDD in Spotify
Final Project – CSC 466

Aditi Gajjar, Anagha Sikha, Nicholas Tan, Othilia Norell, Soren Paetau

[a] agajjar@calpoly.edu,
[b] asikha@calpoly.edu,
[c] nktan@calpoly.edu,
[d] onorell@calpoly.edu,
[e] spaetau@calpoly.edu,

Abstract

This study aimed to assess the efficacy of various Knowledge Discovery in Databases (KDD) methods in predicting music genres using real Spotify data, focusing on both music recommendation and genre classification/clustering. We employed techniques like Apriori Association Rules for pattern discovery, Induce C4.5 and Random Forest for genre prediction, and Collaborative Filtering using Adjusted Weighted KNN for personalized recommendations. An interesting trend emerged in user preferences: a shift towards Pop music from diverse genres like EDM, Dubstep, and Rock, highlighting Pop's universal appeal. Additionally, a mutual preference was observed between Rock and Pop, with Classic Rock showing versatility in listener appeal. In genre prediction, our Random Forest analysis identified tempo and energy as pivotal features for R&B, whereas liveness and acousticness were less informative, corroborating with Sci-Kit Learn's feature importance. Furthermore, when comparing decision trees (C4.5) to Random Forests, the latter demonstrated superior accuracy (50% vs. C4.5's 43%) and consistently higher precision and recall across genres. This suggests Random Forests' ensemble approach as more effective for genre prediction than single decision trees. Our clustering experiments revealed KMeans' effectiveness in grouping songs into specific genres (EDM, Rap, R&B) but its limitation in identifying others like Pop, Latin, and Rock. On the other hand, DBScan struggled to form meaningful clusters due to hyperparameter sensitivities, even after extensive tuning. Collaborative filtering gave results on which Genre's were more or less difficult to predict based on user's numeric rating of their genre affinity. Overall, this study highlights the potential of KDD methods in music genre prediction and clustering, emphasizing the importance of selecting appropriate techniques based on specific goals and dataset characteristics.

1. Introduction

Spotify is not only a music streaming company with over 550 million users worldwide, it is a tech company that wants to improve the listening experience for its users, which is why Spotify uses machine learning in almost every part of its business. This technology helps users find new songs and podcasts through recommendations and search, and creates personalized playlists. Machine learning is used to sort and organize music, show ads, to make their service work better. Spotify is committed to being at the forefront of machine learning technology.

In this study, we explore different KDD methods on Spotify datasets to explore different methods like Association Rules mining, Random Forest, KNN, and Collaborative Filtering and their potential usefulness to uncover patterns and associations between song features and listener preferences.

2. Background and Related Work

Machine learning has been increasingly used in music recommendation systems, a topic that has drawn a lot of attention. Previous studies have used supervised learning methods, such as decision trees, natural language processing models, and random forests, to classify music genres and create personalized music recommendation systems. There's also been research comparing these systems in different settings, like at Spotify and Netflix, to see how well they work with different types of media.

3. Datasets

Three datasets were used to explore our five research questions, the first one containing detailed information of songs on Spotify, the other on user listening preferences, and the last is a collected summary of music genre listening preferences.

3.1. 30000 Spotify Songs

The 30000 Spotify Songs dataset from Kaggle consists of 32,833 tracks sourced from Spotify. It has labeled essential characteristics of each song, including for instance its rhythm (tempo), the pitch scale it uses (key), and its length (duration). Additionally, it offers insights into listener preferences with information on each track's popularity and its musical style or category (genre).

3.2. Spotify Genres Preferences Dataset

The Spotify User Genres is a dataset that summarizes the genre preferences of 152 users on Spotify collected through a survey sent to Cal Poly Students. This dataset categorizes the main genres and their subsets to create a list of user files. The data is represented in a binary format, where each column represents a genre. Ones indicate that a user prefers a particular genre or subgenre, while zeroes represent the genres they did not prefer.

3.3. User Genre Ratings

This dataset was made similarly to above asking about the 1-10 rating the user gives to a standard list of Genres. The format of the responses are outlined in subsection 5.5 and contains 157 responses from Cal Poly Students.

4. Research Questions

Our research aims to explore how certain features of songs relate to their popularity, how people's music tastes change between genres, and if unsupervised learning can cluster songs in the same genre together. Thereby, five different research questions were asked to explore different areas of use for KDD at Spotify. The initial two questions concentrate on user experience, whereas the remaining three delve into more methodological aspects.

1. How well do different song features, such as *tempo* and *energy*, affect the prediction of a song being in the R&B genre?
2. Are there specific genres that users tend to switch between? For instance, do fans of a particular genre like classical music also tend to listen to jazz?
3. Does the data cluster by genre when applied to unsupervised learning algorithms?
4. How much does the accuracy improve at predicting genres when using decision trees versus random forests?
5. Given Spotify genre ratings by users, how does the accuracy of prediction vary between genres using an Adjusted Weighted K-nn approach?

5. Method

Multiple KDD methods like Association Rules mining, Random Forest, KNN and Collaborative Filtering were explored using one Kaggle datasets: *30000 Spotify Songs* and two surveys that we conducted ourselves.

5.1. *How well do different song features affect the prediction of a song being in the RnB genre?*

Understanding which song features most significantly impact various music genres is not only fascinating for the general public but also crucial for music producers and artists. This research question explores the importance of some features for the genre r&b, over others on accuracy metrics when classifying genre using Random Forest.

5.1.1. Preprocessing of Data

The study focuses on nine key attributes: 'danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', and 'tempo'. These attributes, all numerical in nature, were carefully selected from a much larger set of features available in the dataset. Since our dataset encompasses a range of genres beyond R&B, we transformed the genre class label into a binary format. To achieve a balanced representation, we undersampled the dataset, ensuring an equal number of R&B and non-R&B data points. This process yielded a dataset of 10,860 data points, evenly split between the two categories.

5.1.2. Random Forest

The classifier, applied to a subset of 30,000 songs from Spotify, generates numerous decision trees (based on the C4.5 algorithm) to analyze the data. Each tree is constructed from a randomly selected sample of attributes and data points. This process was implemented in 10-fold cross-validation, such that the parity of predictions for the training split was considered the final prediction for that observation. The heuristics and accuracy of these predictions were tracked and then reported.

For each iteration of the Random Forest analysis, a subset of four attributes was randomly chosen from the nine mentioned. The optimal setup for our model included a minimum of 50 trees, with each data point comprising no more than 10% of the dataset. Consequently, the model was trained on a sample of 1,000 data points, utilizing 50 trees.

5.1.3. Different Song Features' Importance

Using Sci-Kit Learn, we found the features' importance for the attributes:

Feature	Importance
Energy	0.13554
Tempo	0.11771
Loudness	0.10758
Duration_ms	0.10679
Acousticness	0.10016
Danceability	0.09469
Valence	0.08809
Liveness	0.08282
Instrumentalness	0.06698

Table 1: Result from running Sci-Kit Learn's Random Forest with Feature Importance

From this we learned that *Energy* and *Tempo* was the two most influential factors, and *Liveness* and *Instruementalness* were the least influential features, for the genre R&B.

5.1.4. Comparative Analysis Using Various Features

Firstly, we conducted runs using all nine attributes. Then, informed by the feature importance ranking, we specifically excluded *Energy* and *Tempo* – identified as the most significant attributes – to observe the variation in accuracy and precision

in predicting the R&B genre. This approach aimed to test the hypothesis that removing these two key features would lead to a notable decrease in the classifier's performance.

In a comparative approach, we also conducted runs where we removed the two attributes deemed least important for the R&B genre. This step was taken to evaluate whether omitting these less critical features could potentially enhance the performance metrics of our model.

5.2. *Are there specific genres that users tend to switch between?*

Understanding the dynamics of music preferences is a highly applicable insight with applications from personalized music recommendations to market segmentation in the music industry. One intriguing aspect is the tendency of listeners to switch between specific music genres. To find this information, we executed an Apriori algorithm to identify frequent itemsets and generate association rules. The Apriori algorithm was configured with a minimum support threshold of 0.1 and a minimum confidence threshold of 0.8 to extract significant and reliable patterns. Through performing hyperparameter tuning as showcased in Figure 3, we found the number of frequent itemsets generated in order to determine a cutoff.

Minimum Support (minSup)	Number of Frequent Itemsets
0.05	233
0.10	94
0.15	43
0.20	17

Table 2: Results of Apriori Algorithm with Varying minSup Values

Upon analyzing the information, we determined the minimum support of 0.1 to give us a reasonable number of itemsets (genre listening patterns) without being overwhelmed and too specific. This gave us a mix of common and somewhat less common (but interesting) patterns.

On the other hand, the minimum confidence threshold was also tuned over the possible values: 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99 as shown in Table 4 and selected at the value of 0.8.

Minimum Confidence (minConf)	Number of Association Rules
0.5	408
0.6	393
0.7	364
0.8	317
0.9	257
0.95	195
0.99	193

Table 3: Results of Association Rules Generation with Varying minConf Values

This gives strong confidence while still accounting for wanting to extract interesting patterns, we choose a minimum confidence of 0.8. This is also one of the biggest differences in the number of association rules generated as exemplified in Figure 1, and so can be noted as one of the best places to set a cutoff.

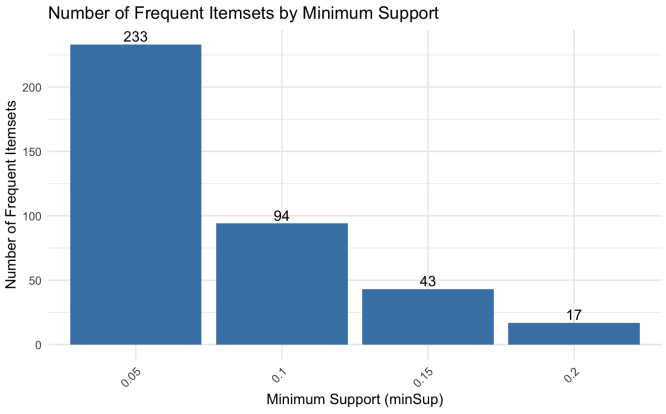

Figure 1: Hyperparameter tuning of minConf

After selecting the hyperparameters and training our data, we were able to extract frequent itemsets for listened-to genres, which gave us insights into listening trends similar to those of survey respondents.

5.3. *Does the data cluster itself by genre?*

Within Spotify, playlists are often created that group songs into certain genres that users may like. To explore potential improvements in the way Spotify generates these playlists, we aimed to leverage unsupervised clustering algorithms. Our goal was to utilize these algorithms to accurately group songs together by their genre based on their song characteristics. To assess whether the data clusters songs within the *30000 Spotify Songs* dataset by genre, we applied two unsupervised learning algorithms: KMeans and DBScan.

5.3.1. *Data Preprocessing*

Before feeding the data into the two clustering algorithms, we first isolated the numeric attributes of the dataset to collect the features of each song that relate to audio characteristics. We decided to omit some of the identifying characteristics, such as the track artist or album, as these may be more subjective and could vary across different sources. For example, the same artists could produce music in different genres, which may cause ambiguity within the clustering process.

5.3.2. *Clustering*

After parsing and cleaning the data, we ran both KMeans and DBScan on the data and performed parameter hypertuning to see if the models could generate accurate clusters. For KMeans, we chose to just utilize k = 6 since there were 6 overall genres within the data (pop, rnb, rock, latin, edm, rap), and see if it could generate 6 different clusters corresponding to each genre. We normalized the data since the columns were on different scales, and also used kmeans+ for centroid initialization, alongside euclidean distance as the distance metric. For DBScan, we hypertuned both the epsilon distance (ϵ) and the minNum parameter, which is the minimum number of neighbors that need to be within that epsilon distance. Below were a set of ϵ values and minNum values we ran DBScan with:

ϵ distance	minNum
250	5
500	5
750	5
250	10
500	10

Table 4: Set of ϵ and minNum parameter values that were used with DBScan

5.4. *How much does the accuracy improve at predicting genres when using decision trees versus random forests?*

Predicting music genres is a challenge we decided to explore in order to understand the relationship between various song features and perhaps improve the current way Spotify classifies their genres. In our exploration of Knowledge Discovery in Databases methods on the 30,000 Spotify Songs dataset from Kaggle, we focused on comparing the efficacy of decision trees, specifically the C4.5 algorithm, and Random Forests in genre prediction.

5.4.1. Preprocessing of Data

Similar to question 1, we focus on the same nine key (numerical) attributes: 'danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', and 'tempo', which were all the numerical song features from this dataset. We also created training and testing datasets in order to correctly test the accuracy of our C4.5 algorithm, where we deployed a 80-20 train-test split, resulting in 26,267 data points in the training data set and 6,568 data points in the test dataset.

5.4.2. C4.5

C4.5 algorithm is used to recursively split the training dataset based on song features, facilitating the identification of patterns that contribute to the accurate genre predictions. The implementation of our algorithm checks termination conditions such as songs with the same genres or no more song features to consider, leading to the creation of a leaf node and the construction of the decision tree. Following the execution of C4.5 algorithm on our training dataset, we evaluate the performance of the C4.5 decision tree on a separate test dataset by systematically classifying each data row based on the provided decision tree. Accuracy was computed by tracking the number of correct and incorrect classifications, and an overall confusion matrix was constructed to comprehensively assess the classifier's predictive performance on the test data.

5.4.3. Random Forest

The C4.5 single decision tree model was compared to the Random Forest model. Our Random Forest function is an ensemble learning method that builds a collection of decision trees using bagging or bootstrap aggregating, and consists of 3 main hyperparameters the number of attributes(m, selected without replacement), the number of data points selected randomly with replacement (k), and the number of decision trees to build (N). After a intensive hyper parameter tuning process, we configured the random forest with m=9, k=600, and N=5, which resulted in the highest accuracy, to assess the impact of these parameters on prediction accuracy.

5.5. *Given Spotify genre ratings by users, how does the accuracy of prediction vary between genres?*

It is no secret that people have varying different music tastes, spanning across multiple genres. In this question we hope to use collaborative filtering techniques, particularly a Weighted Adjusted sum with the 10 nearest neighbors, based on a ranking. Adjusted Weighted Sum fine-tunes the weights of users by considering factors such as user reliability and rating consistency. While in this particular case we had perfect data structure, such that their were no missing values, collaborative filtering's strength lies in it's ability to work with imperfect data. Therefore, we hope to extrapolate the results of which genre's are easy to predict throughout the answering of this question.

This approach serves as a robust tool for uncoverinng patterns in user behavior, ultimately contributing to a more tailored and enjoyable music recommendation system.

5.5.1. Collecting Data

To collect the data we sent out a poll to fellow students asking how they would rate their affinity/enjoyment towards the following genres

- Classical
- EDM
- Country
- Jazz
- Hip Hop
- Religious
- Pop
- Folk
- R&B
- Latin
- Film
- Rock
- Soul

These genres were used by on the Genre selection we had established throughout this entire report. The results from the poll were of the following format

userId	Country	Pop	$\cdots$	Rock	Jazz	R&B
0	8	3	$\cdots$	9	9	1
1	5	3	$\cdots$	7	6	5
2	5	4	$\cdots$	6	5	5

Table 5: Format of Data

such that the ratings varied on a scale from 1-10 for each of the primary genres of our study.

5.5.2. Collaborative filtering

Using `EvaluateCFList.py` a set of 50 random test 'users' was created and the given 13 genres were then each loaded into individual test case csv files. The same 50 user's were used across all genres, to control for a random chance from genre to genre influencing the results. The concept behind this methodology was that using a collaborative filtering method to predict across the same *n* user's based on their other music interests, would allow the model to find people who had 'similar' music

taste based on a Pearson similarity coefficient and then those 10 nearest neighbors would be averaged to predict what rating of genre y user x had. To quantify the metrics across each genre, we used the mean absolute error of the model's predicted rating from the actual rating. However, we would expect genres with less variance in their distribution to have predictions that were closer to the mean, so we divided the MAE of the prediction from the standard deviation of ratings for that genre, to develop a ratio heuristic. The statistical significance of this ratio can be mathematically shown, but the intuition behind the process gives interesting results.

6. Result

6.1. *How well do different song features affect the prediction of a song being in the R&B genre?*

The precision and recall in predicting the R&B genre improves when prioritizing more significant features and excluding the less crucial ones.

Configuration	Precision	Recall	Accuracy
All Features	67.0%	71.2%	68.0%
Without 2 Best Features	64.6%	68.3%	65.0%
Without 2 Worst Features	61.8%	72.2%	68.1%

Table 6: Results of 10-fold cross validation with 1000 data points and 50 trees

In this scenario, the highest recall for identifying R&B was achieved by excluding the two least important features. This means, that out of all R&B instances, 68.3% were correctly identified by the model. Furthermore, the model that excluded the two most important features exhibited the lowest recall.

6.2. *Are there specific genres that users tend to switch between?*

The analysis revealed intriguing insights into the genre-switching behavior of music listeners. Notably, several strong association rules were discovered that signify a tendency for listeners to switch between specific genres. While we were able to produce 317 rules, we limited our report analysis to only the interesting ones, where the tendencies were strictly across different genre types, and not within. With the strict cross-genre rules, with high confidence levels, there was a clear pattern of genre switching. Some of the noteworthy findings are outlined in Table 7.

Association Rule	Confidence
EDM, Dubstep → Pop	0.89
Hard Rock, Classic Rock, Rock → Pop	0.83
Techno, EDM, Rock → Pop	1.00
Soundtracks, Film, EDM → Pop	1.00
Techno, EDM, House → Pop	1.00
Instrumental, EDM, Classical → Pop	0.94
Classic Rock, Rock, Latin Music, Indie Rock → Pop	1.00
EDM, Pop, Alternative Pop, House → Rock	0.83
EDM, Pop, Indie Pop, House → Rock	0.80
Classic Rock, EDM, Rock, Indie Rock → Pop	0.94
Classic Rock, Rock, Musicals, Film → Pop	1.00
Classic Rock, Rock, Soundtracks, Film → Pop	0.94
Classic Rock, EDM, Rock, House → Pop	0.94
EDM, Rock, House, Indie Rock → Pop	0.94
Classical, Musicals, Soundtracks, Film → Pop	1.00
Classical, Musicals, Film, Instrumental → Pop	1.00
Classical, Pop, Indie Pop, Alternative Pop, Orchestrals → Rock	0.83
Classic Rock, Classical, Rock, Alternative Rock, Indie Rock → Pop	0.94
Classical, Rock, Film, Soundtracks, Instrumental → Pop	1.00
Classical, Film, Soundtracks, Instrumental, Orchestrals → Pop	1.00

Table 7: Strict Cross-Genre Association Rules and Their Confidence Levels

In summary, there was a significant trend that showed a switch to Pop music from various genres, indicating its wide appeal. Genres like EDM, Dubstep, Hard Rock, Classic Rock, and Techno frequently also led to a preference for Pop. Listeners of Soundtracks, Film music, and EDM showed a strong inclination towards Pop. Fans of Instrumental, Classical, and combinations of Rock genres (like Classic Rock and Indie Rock) also tended to prefer Pop. Combinations of significantly different genres, such as Classic Rock with Latin Music and Indie Rock, also demonstrated a transition towards Pop. Intricate genre combinations, like Classical, Musicals, Soundtracks, and Film, were also associated with a shift to Pop. Notably, there were also rules where Rock was the destination genre, indicating its appeal across different music tastes. For instance, combinations of EDM, Pop, and sub-genres of Pop (like Alternative Pop and Indie Pop) showed a transition towards Rock. This suggests a two-way interchangeability between Rock and other popular genres, particularly Pop and its sub-genres. Lastly, it is notable that classic Rock appeared in various combinations, both as a source and part of the destination genre, showcasing its versatility and wide-ranging appeal for listeners.

6.3. *Do unsupervised learning algorithms cluster the data by genre based on their song characteristics?*

Our analysis of clustering songs into genres using KMeans and DBScan revealed notable findings. KMeans outperformed DBScan, successfully grouping all songs into six distinct clusters. We examined the ground truth (genre column) within

these clusters and observed predominant genres such as EDM, rap, and R&B. Specifically, KMeans identified three clusters for EDM, two for rap, and one for R&B. See figure 8.1 within the Appendix for more details on the output. Intriguingly, popular genres like pop, Latin, and rock were absent in KMeans clusters.

In contrast, DBScan posed challenges in hyperparameter tuning, requiring substantial time to build clusters on the entire dataset. To mitigate this, we sampled 1000 songs for analysis. Establishing baseline parameter values for DBScan proved crucial, as certain combinations often labeled all points as noise. After experimentation, a minimum number of neighbors (minNum) set to 5 yielded optimal results, effectively generating meaningful clusters. Attempts to increase this value to 10 led to a reduction in the number of clustered points.

Exploring epsilon distances for DBScan, we tested values of 250, 500, and 750. Unfortunately, none of these configurations produced satisfactory genre-based clusters. Instead, we consistently obtained over 10 clusters with a substantial number of noise points for some parameter combinations, indicating the sensitivity of DBScan to these parameters. See figure 8.2 in the Appendix for an example DBScan output.

In summary, KMeans demonstrated superior performance in clustering songs by genres, effectively capturing distinct music categories. Conversely, DBScan posed challenges in hyperparameter tuning and consistently struggled to produce meaningful clusters with the given dataset. The absence of certain genres in KMeans clusters, along with the difficulties encountered in optimizing DBScan, emphasizes the importance of carefully selecting and fine-tuning clustering algorithms for music genre analysis.

6.4. ***How much does the accuracy improve at predicting genres when using decision trees versus random forests?***

In our exploration of genre prediction using Spotify datasets, we observed distinct performance outcomes between the Random Forests and the C4.5 decision tree induction algorithm.

Test	Accuracy
Induce C4.5	0.431247
Random Forests	0.501934

Table 8: The overall accuracies for C.45 and RF on the test fold of the dataset

6.4.1. Random Forest

The Random Forest classifier obtained an accuracy across all genres of slightly over 50%.

To gain deeper insights into this result, the accuracy, precision and recall were calculated for every genre:

	Accuracy	Precision	Recall
EDM	86.81%	63.43%	66.99%
Latin	82.04%	41.70%	36.18%
Pop	78.90%	34.46%	28.66%
R&B	80.60%	40.80%	38.30%
Rap	83.66%	52.82%	61.99%
Rock	88.38%	60.23%	67.60%

Table 9: The accuracy, precision and recall for all genres using the Random Forest classifier

The accuracies are high across all genres, which can be explained because the dataset is very unbalanced when looking at one genre at a time, since the number of true negatives is very high the accuracy will consequently also be very high. Thereby, precision and recall become more interesting to look at. The precision and recall are the highest for EDM, Rock and Rap. On the other hand, the Pop genre had the lowest precision and recall.

6.4.2. C4.5

In contrast, when employing the C4.5 algorithm on the test data, the overall accuracy was 43%, while the average accuracy was higher 56%.

To gain deeper insights into this result, the accuracy, precision and recall were calculated for every genre:

	Accuracy	Precision	Recall
edm	82.91%	56.81%	53.73%
latin	79.55%	33.98%	34.68%
pop	75.06%	26.55%	25.38%
r&b	79.40%	36.28%	34.95%
rap	82.02%	45.55%	49.40%
rock	87.32%	57.53%	60.26%

Table 10: The accuracy, precision and recall for all genres using one C4.5 decision tree

The precision and recall are both again the highest for EDM, Rock and Rap. Furthermore, the Pop genre has also again the lowest precision and recall.

6.5. ***Given Spotify genre ratings by users, how does the accuracy of prediction vary between genres?***

The results from fitting the adjusted weighted sum with the nearest 10 neighbors, returns a MAE of prediction from the actual value. Additionally, the standard deviation in that column was found in order to use a

Genre	MAE of Prediction	SD of Ratings in Genre
Classical	1.765	2.166
Country	2.112	1.931
EDM	2.49	1.955
Film	2.436	3.039
Folk	1.138	2.022
HipHop	5.059	1.684
Jazz	2.266	1.83
Latin	1.922	1.393
Pop	3.389	1.644
R&B	2.311	1.993
Religious	1.789	2.084
Rock	1.6	2.065
Soul	2.761	1.297

Table 11: **Results from Model**

Additionally the ratio between the MAE and SD was plotted relative to genre such

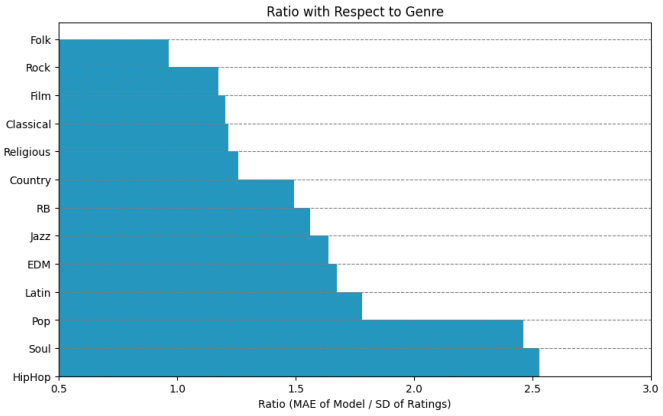

Figure 2: Preidction MAE to SD by Genre

7. Discussion

The findings of this data will be discussed in the discussion section of this report.

7.1. *How well do different song features affect the prediction of a song being in the RnB genre?*

Our results from using running three Random Forest models with various features validated our hypothesis and approved the benchmark set by Sci-Kit learns feature importance. The models demonstrated a clear ability to surpass the 50% accuracy threshold of random guessing, indicating that our chosen features effectively identify R&B music.An important finding from our research is that energy and tempo are crucial in identifying R&B songs. These parts really capture the rhythm and soul of R&B music. Therefore, it can be viewed logical that they had an impact on the accuracy, precision and recall.

The enhanced performance after removing the two least significant features suggests that these features introduced noise rather than clarity. By eliminating them, the model could focus on the more impactful attributes, thereby improving the predictive accuracy.

Given that recall measures the model's success in correctly identifying R&B songs, a higher recall rate means fewer R&B tracks are missed. This is particularly crucial for applications like music recommendation, where ensuring a comprehensive selection of true R&B songs enhances user satisfaction. Precision matters when we need to be really sure a song is R&B before we say it is. On the other hand, recall is important when we want to find as many R&B songs as possible. The choice depends on what's more important for the task at hand. Also, since many songs today mix elements from different genres, it might not be considered a too big of a problem if a song gets placed in a playlist of a similar genre.

7.2. *Are there specific genres that users tend to switch between?*

The analysis of association rules derived from listeners' music preferences has unveiled a series of interesting patterns and tendencies in genre switching. These findings shed light on the complex interplay between various music genres and how listeners' preferences evolve from one genre to another.

Firstly, there was a pronounced trend indicating a significant switch to Pop music from a variety of other genres. This suggests the extensive appeal of Pop music across diverse listener groups. Genres such as EDM, Dubstep, Hard Rock, Classic Rock, and Techno were frequently observed to lead to a preference for Pop. This inclination towards Pop music might be attributed to its broad spectrum and its ability to incorporate and blend elements from various genres, making it universally appealing to a wide audience.

Furthermore, listeners of Soundtracks, Film music, and EDM exhibited a strong preference for transitioning towards Pop music. This pattern was also evident among fans of Instrumental and Classical music, as well as those of Rock sub-genres like Classic Rock and Indie Rock. These tendencies highlight the cross-genre appeal of Pop music, transcending traditional musical boundaries.

Interestingly, Rock emerged not only as a source genre but also as a destination. This was evident in instances where combinations of EDM and Pop, along with their sub-genres such as Alternative Pop and Indie Pop, showed a tendency to switch to Rock. This pattern suggests a two-way interchangeability between Rock and other popular genres, notably Pop and its various offshoots. Classic Rock, in particular, was notable for its appearance in a range of combinations, serving both as a source and a destination genre. This underscores the versatility and broad appeal of Rock music in the contemporary music preference landscape.

7.3. *Do unsupervised learning algorithms cluster the data by genre based on their song characteristics?*

The findings from our analysis of clustering songs into genres using KMeans and DBScan shed light on the strengths and limitations of these unsupervised learning algorithms in the context of music categorization.

7.3.1. DBScan

DBScan's challenges in generating meaningful clusters can be attributed to its sensitivity to hyperparameters, particularly minNum and epsilon distances. Hyperparameter tuning for DBScan proved to be a non-trivial task, and the algorithm's performance heavily depended on finding an optimal combination of these parameters. The sensitivity to noise points and the difficulty in selecting appropriate values for minNum and epsilon distances hindered the algorithm's ability to generate cohesive clusters. Additionally, the computational demands of DBScan on the entire dataset necessitated sampling, introducing the challenge of ensuring that the sampled subset was representative of the broader dataset.

7.3.2. KMeans

While KMeans outperformed DBScan by successfully clustering all songs, it exhibited limitations in capturing the diversity of music genres. Notably, KMeans predominantly clustered songs into EDM, rap, and R&B, neglecting popular genres such as pop, Latin, and rock. This limitation may be attributed to the algorithm's sensitivity to centroid initialization, especially with the KMeans++ variant. The initialization method can impact the convergence of the algorithm and influence the resulting clusters. Further exploration into centroid initialization strategies and their impact on genre clustering is a potential avenue for future research.

7.3.3. Challenges

The challenges encountered with both algorithms have implications for playlist generation on music streaming platforms like Spotify. The absence of certain genres in KMeans clusters suggests the need for a more nuanced approach in playlist curation, incorporating additional features or refining clustering algorithms to better capture the diversity of musical genres. The difficulties faced in optimizing DBScan underscore the importance of considering algorithmic complexity and efficiency in the context of large-scale datasets. Thus, while both KMeans and DBScan offer insights into the clustering of songs by genre, our study emphasizes the need for a careful consideration of algorithmic intricacies and parameter tuning strategies.

7.4. *How much does the accuracy improve at predicting genres when using decision trees versus random forests?*

Comparing the performance metrics of C4.5 and Random Forests in predicting music genres reveals distinct characteristics in accuracy, precision, and recall.

In terms of overall accuracy, Random Forests outperform C4.5, with an accuracy of 50.19% compared to C4.5's 43.19%. This indicates that, on average, Random Forests provide a higher percentage of correct genre predictions. Analyzing precision, which measures the accuracy of positive predictions, Random Forests consistently exhibit higher precision across all genres compared to C4.5. Precision is notably higher for EDM, Latin, Pop, R&B, Rap, and Rock in the Random Forest model. This suggests that when Random Forests predict a particular genre, they are more likely to be correct than C4.5. Examining recall, which measures the ability to capture all instances of a given genre, Random Forests also generally outperform C4.5. The Random Forest model demonstrates higher recall for EDM, Latin, Pop, R&B, Rap, and Rock. This implies that Random Forests are more effective at identifying instances of each genre, minimizing false negatives compared to C4.5.

Delving into the genre-specific performance, both the C4.5 decision tree and the Random Forest models show commendable precision and recall in predicting genres like EDM, Rock, and Rap. These genres seem to be more easily and accurately classified than others, particularly Pop, which scored the lowest in terms of precision and recall. The broad and inclusive nature of the Pop genre might contribute to its reduced predictability. This contrasts with genres such as Rap, which might be identifiable by distinct characteristics like higher speechiness, or EDM, often associated with a quicker tempo and likely a higher danceability. Such specific attributes render certain genres more readily distinguishable, thereby facilitating more accurate predictions.

Comparing to the random guessing baseline of approximately 16.67% with the achieved accuracies of Random Forests (50.19%) and C4.5 (43.19%), it is evident that both algorithms significantly outperform random chance. The algorithms demonstrate a capacity to discern patterns and relationships within the data, leading to predictions that are substantially more accurate than what would be achieved by random guessing alone. The higher accuracies, along with improved precision and recall metrics, further emphasize the effectiveness of these machine learning algorithms in predicting music genres based on the provided datasets.

Considering these metrics collectively, Random Forests emerge as the better-performing algorithm in predicting music genres. This method consistently achieves a higher overall accuracy, precision, and recall across all genres compared to C4.5. This suggests that the ensemble approach of Random Forests, leveraging multiple decision trees, contributes to a more robust and accurate prediction of music genres in the context of the provided datasets.

7.5. *Given Spotify genre ratings by users, how does the accuracy of prediction vary between genres?*

Comparing genre performance in collaborative filtering methods reveals many interesting results. We see that Folk music was the easiest to predict, even relative to the deviation present in the data. In other words, Folk music was ratings were the most distinct from other's given ratings wo were similar to them.

On the other side the spectrum, our model performed quite disappointingly on Hip-Hop with a mean absolute error of over 55, considering 2nd worst was around 3.3, this is very surprising given it's consistent rating, with a standard deviation of only 1.684. The ratio of Hip-Hop was also the highest at above 2.5.

So while collaborative filtering did have a perfect dataset for this method, there is hope that these results can be pushed beyond with imperfect data with notable effectiveness at least with Folk, Rock and Film genre predictions. Further investigations include gathering more data from a more diverse group of people.

8. Conclusion

For future directions in terms of clustering the data by genre, refining hyperparameter tuning strategies for DBScan and exploring alternative clustering algorithms could enhance the accuracy and efficiency of genre-based song clustering. Additionally, investigating the impact of diverse feature sets on clustering performance may provide a more comprehensive understanding of music genre analysis. Future research should aim to address the identified limitations and contribute to the development of more robust and accurate clustering techniques for music categorization.

Our experiments with Random Forest models not only affirm the effectiveness of our feature selection for R&B genre prediction but also highlight the strategic balance required in model tuning and metric prioritization for optimal performance. Additionally, we found that Random Forests were more effective at predicting genres compared to a single decision tree. Future research should focus on fine-tuning the parameters more which may lead to improved model performance and considering the development of hybrid models that combine the strengths of different machine learning algorithms.

The metrics returned from our collaborative filtering runs show's that all genre's are not equally as predictable, even if they vary similarly in the data. Understanding user preferences and predicting how much a user enjoys a certain genre come's with a level of complexity that cannot be inferred from Pearson similarity to other user's. We hope to further investigate these findings to develop more accurate models.

9. Appendix

9.1. Question 1

4 attributes, 1000 datapoints, 50 trees using 10-fold cross val:

Actual \Predicted	0	1
0	3518	1543
1	1911	3887

Table 12: Confusion matrix using all attributes

Actual \Predicted	0	1
0	3390	2039
1	1718	3712

Table 13: Confusion matrix excluding the two best features (tempo and energy)

Actual \Predicted	0	1
0	3482	1947
1	1511	3919

Table 14: Confusion matrix excluding the two worst features (liveness and instrumentalness)

9.2. Question 3

9.2.1. KMeans with k = 6, kmeans++, normalization, and euclidean distance

```
Cluster 0:
Center [0.309365  ,0.65372578,0.71623427,0.61134838,0.80594238,0.31853382,
 0.08469541,0.16837314,0.78632249,0.2066107 ,0.35336441,0.51441684,
 0.4571419 ]
Max Distance to Center: 1.4656467849295458
Min Distance to Center: 0.4026763632301592
Average Distance to Center: 0.795266259517764
SSE from Center: 1290.2614978862564
1937 Points:
Top genres: edm
-------------------------------------------------------------

Cluster 1:
Center [0.39325795,0.69733234,0.72030691,0.57343134,0.83681021,0.88897214,
 0.12111328,0.13022509,0.07373644,0.1827015 ,0.58422955,0.50032292,
 0.42939641]
Max Distance to Center: 1.4969033767216846
Min Distance to Center: 0.20010128727750204
Average Distance to Center: 0.6590096459625864
SSE from Center: 7049.183933195851
14681 Points:
Top genres: rap
-------------------------------------------------------------

Cluster 2:
Center [0.3219847 ,0.64420725,0.77103713,0.15132767,0.84321627,0.99251058,
 0.10827468,0.0715422 ,0.13520224,0.20701951,0.4896623 ,0.5152023 ,
 0.44402188]
Max Distance to Center: 1.6585489165458573
Min Distance to Center: 0.12549525980906834
Average Distance to Center: 0.5623453955871398
SSE from Center: 2167.7545824967133
6142 Points:
Top genres: edm
-------------------------------------------------------------

Cluster 3:
Center [0.46548156,0.68833353,0.63949999,0.27306111,0.82030264,0.01652479,
 0.12520715,0.25629493,0.11866836,0.16878269,0.52853945,0.49673827,
 0.41485462]
Max Distance to Center: 1.7978127446050396
Min Distance to Center: 0.1920491366534746
Average Distance to Center: 0.6258335477595464
SSE from Center: 2647.3940902714257
5991 Points:
Top genres: rap
-------------------------------------------------------------

Cluster 4:
Center [0.39594036,0.66093448,0.73508173,0.65491091,0.84257363,0.00792677,
 0.12277802,0.12538182,0.05089835,0.21081956,0.50887772,0.50854611,
 0.44052605]
Max Distance to Center: 1.6487069128153022
Min Distance to Center: 0.18725723075132708
Average Distance to Center: 0.5860227010697145
SSE from Center: 3946.7374108337413
10597 Points:
Top genres: edm
-------------------------------------------------------------

Cluster 5:
Center [0.57032441,0.64393328,0.57082025,0.23272727,0.8062551 ,0.92913386,
 0.10648493,0.3425684 ,0.05980628,0.17138205,0.43756028,0.50013713,
 0.43347555]
Max Distance to Center: 1.5407285748458566
Min Distance to Center: 0.1701556009132886
Average Distance to Center: 0.6448333206177727
SSE from Center: 2973.709025857518
6350 Points:
Top genres: r&b
-------------------------------------------------------------
```

9.2.2. DBScan with random sample of 1000, ϵ = 500 and minNum = 5

```
Cluster 0:
Center [ 4.29166667e+01, 6.39083333e-01, 7.22083333e-01, 3.00000000e+00,
 -6.22250000e+00, 5.83333333e-01, 7.53500000e-02, 2.03879167e-01,
  2.07069025e-01, 1.22858333e-01, 4.92750000e-01, 1.27373833e+02,
  2.47825333e+05]
Max Distance to Center: 811.7073346959912
Min Distance to Center: 110.4006351873913
Average Distance to Center: 351.1950527369948
SSE from Center: 2001930.670158211
12 Points:
Top genres: 0
-------------------------------------------------------------

Cluster 1:
Center [ 4.44857143e+01, 6.37971429e-01, 7.40200000e-01, 5.80000000e+00,
 -5.38600000e+00, 6.28571429e-01, 8.60314286e-02, 1.09531262e-01,
  8.82018314e-03, 1.59602857e-01, 4.90331429e-01, 1.12979886e+02,
  2.17972029e+05]
Max Distance to Center: 1652.2834642658186
Min Distance to Center: 48.31502253886101
Average Distance to Center: 800.5351016163339
SSE from Center: 28441680.23258435
35 Points:
Top genres: 1
-------------------------------------------------------------

Cluster 2:
Center [ 4.0700000e+01, 6.9450000e-01, 6.9838500e-01, 4.7000000e+00,
```

```
 -6.7348500e+00, 7.0000000e-01, 1.0684000e-01, 1.6876390e-01,
  1.3067854e-01, 1.8337500e-01, 4.8945000e-01, 1.2142745e+02,
  2.2499860e+05]
Max Distance to Center: 1386.4187961822824
Min Distance to Center: 54.10130779156566
Average Distance to Center: 707.8269872027827
SSE from Center: 12492926.649477389
20 Points:
Top genres: 4
------------------------------------------------------------

Cluster 3:
Center [ 4.85833333e+01, 6.32233333e-01, 6.96933333e-01, 5.15000000e+00,
 -6.69606667e+00, 5.33333333e-01, 1.06606667e-01, 1.73202218e-01,
  7.83610903e-02, 1.92498333e-01, 5.15005000e-01, 1.25073567e+02,
  2.36007817e+05]
Max Distance to Center: 4207.954188152001
Min Distance to Center: 154.86612007604248
Average Distance to Center: 1973.9242080717315
SSE from Center: 312951917.1420733
60 Points:
Top genres: 14
------------------------------------------------------------

Cluster 4:
Center [ 4.58260870e+01, 6.47862319e-01, 7.16717391e-01, 5.30434783e+00,
 -6.16478986e+00, 6.81159420e-01, 1.05574638e-01, 1.55734075e-01,
  3.54731320e-02, 1.85726087e-01, 4.90772464e-01, 1.20364826e+02,
  2.07749413e+05]
Max Distance to Center: 7179.734362263989
Min Distance to Center: 115.63676951313141
Average Distance to Center: 3530.058184188848
SSE from Center: 2219909485.3729253
138 Points:
Top genres: 6
------------------------------------------------------------

Cluster 5:
Center [ 4.60857143e+01, 6.51057143e-01, 6.31314286e-01, 4.88571429e+00,
 -6.59328571e+00, 5.42857143e-01, 1.10605714e-01, 2.19675714e-01,
  3.62202663e-02, 1.69620000e-01, 4.89742857e-01, 1.11966114e+02,
  1.98132257e+05]
Max Distance to Center: 2360.837401985804
Min Distance to Center: 52.49551638107818
Average Distance to Center: 1193.361816389417
SSE from Center: 67521426.3596123
35 Points:
Top genres: 26
------------------------------------------------------------

Cluster 6:
Center [ 4.38888889e+01, 6.91000000e-01, 6.74555556e-01, 5.33333333e+00,
 -6.87000000e+00, 5.55555556e-01, 1.09227778e-01, 1.75725556e-01,
  4.37338322e-02, 1.59327778e-01, 4.98000000e-01, 1.06624833e+02,
  1.81068944e+05]
Max Distance to Center: 1040.6823379211673
Min Distance to Center: 83.8050932760727
Average Distance to Center: 536.4817010899865
SSE from Center: 6746206.041881243
18 Points:
Top genres: 29
------------------------------------------------------------

Cluster 7:
Center [ 4.42777778e+01, 6.72833333e-01, 6.72055556e-01, 4.50000000e+00,
 -6.77533333e+00, 5.00000000e-01, 1.30866667e-01, 1.15677222e-01,
  6.79183033e-02, 2.62733333e-01, 4.76827778e-01, 1.22077500e+02,
  1.91505444e+05]
Max Distance to Center: 936.5795917309588
Min Distance to Center: 102.88681702880154
Average Distance to Center: 484.58522925921795
SSE from Center: 5195757.807254611
18 Points:
Top genres: 32
------------------------------------------------------------

Cluster 8:
Center [ 5.06666667e+01, 7.07266667e-01, 6.96933333e-01, 4.93333333e+00,
 -6.34726667e+00, 3.33333333e-01, 1.15713333e-01, 1.91642600e-01,
  4.35517113e-02, 1.44020000e-01, 5.06933333e-01, 1.24186800e+02,
  1.77496067e+05]
Max Distance to Center: 1118.1499775558214
Min Distance to Center: 32.33385614755494
Average Distance to Center: 511.28810754152306
SSE from Center: 5664258.271854355
15 Points:
Top genres: 38
------------------------------------------------------------

Cluster 9:
Center [ 4.02000000e+01, 6.23800000e-01, 7.25320000e-01, 5.52000000e+00,
 -6.18200000e+00, 6.40000000e-01, 7.86320000e-02, 1.34034700e-01,
  5.01982044e-02, 2.46228000e-01, 4.90480000e-01, 1.30457000e+02,
  2.21526040e+05]
Max Distance to Center: 1686.552247695736
Min Distance to Center: 41.06045414916837
Average Distance to Center: 813.1495134651316
SSE from Center: 22899610.397211622
25 Points:
Top genres: 19
------------------------------------------------------------

Cluster 10:
Center [ 4.20750000e+01, 6.84075000e-01, 7.79350000e-01, 5.87500000e+00,
 -5.29855000e+00, 5.75000000e-01, 1.23155000e-01, 1.25176550e-01,
  8.18758965e-02, 2.10610000e-01, 5.26930000e-01, 1.24382875e+02,
  1.71466125e+05]
Max Distance to Center: 2479.754139476689
Min Distance to Center: 124.64402672004616
Average Distance to Center: 1345.7663461467334
SSE from Center: 93985036.53599837
40 Points:
Top genres: 30
------------------------------------------------------------

Cluster 11:
Center [ 4.66666667e+01, 7.30166667e-01, 7.90666667e-01, 3.83333333e+00,
 -5.04750000e+00, 8.33333333e-01, 8.17000000e-02, 1.77001667e-01,
  3.54638333e-01, 1.71150000e-01, 5.50666667e-01, 1.20681333e+02,
  1.31610667e+05]
Max Distance to Center: 390.11552972687304
Min Distance to Center: 94.0827679349849
Average Distance to Center: 272.28007248096236
SSE from Center: 532629.5102388598
6 Points:
Top genres: 65
------------------------------------------------------------

Cluster 12:
Center [ 4.80000000e+01, 6.63647059e-01, 7.95411765e-01, 4.35294118e+00,
 -5.27417647e+00, 5.88235294e-01, 6.92000000e-02, 8.35946765e-02,
  1.27462926e-01, 2.92364706e-01, 5.00529412e-01, 1.22369118e+02,
  1.56792529e+05]
Max Distance to Center: 1073.0242652388881
Min Distance to Center: 29.129913755942432
Average Distance to Center: 577.9899711657368
SSE from Center: 7234600.409912385
17 Points:
Top genres: 70
------------------------------------------------------------

Cluster 13:
Center [ 4.25882353e+01, 7.06352941e-01, 7.08647059e-01, 6.00000000e+00,
 -7.37564706e+00, 4.70588235e-01, 1.25611765e-01, 1.85163765e-01,
  5.47712471e-02, 1.86605882e-01, 5.71294118e-01, 1.13173176e+02,
  2.52018471e+05]
Max Distance to Center: 741.6266542600681
Min Distance to Center: 45.59232497107436
Average Distance to Center: 366.27296402156423
SSE from Center: 3380788.2902094815
17 Points:
Top genres: 44
------------------------------------------------------------

Cluster 14:
Center [ 4.05000000e+01, 6.60406250e-01, 7.46187500e-01, 5.53125000e+00,
 -6.34931250e+00, 6.25000000e-01, 1.29159375e-01, 1.43149562e-01,
  4.48716338e-02, 1.88937500e-01, 5.43968750e-01, 1.26194625e+02,
  2.28735312e+05]
Max Distance to Center: 2078.082685525748
Min Distance to Center: 50.59523728006986
Average Distance to Center: 1037.2106971192318
SSE from Center: 45468433.65962976
32 Points:
Top genres: 15
------------------------------------------------------------

Cluster 15:
Center [ 3.45555556e+01, 6.71333333e-01, 6.76888889e-01, 4.33333333e+00,
 -9.34233333e+00, 6.66666667e-01, 9.33333333e-02, 7.81447022e-02,
  1.91167458e-01, 1.71844444e-01, 5.31777778e-01, 1.29078444e+02,
  2.94906667e+05]
Max Distance to Center: 746.6408425444006
Min Distance to Center: 74.5311721298192
Average Distance to Center: 421.8456226560724
SSE from Center: 1986074.2686155085
9 Points:
Top genres: 105
------------------------------------------------------------

Cluster 16:
Center [ 3.89193548e+01, 6.61725806e-01, 7.23354839e-01, 4.96774194e+00,
 -5.73458065e+00, 6.12903226e-01, 1.15135484e-01, 1.47470790e-01,
  5.76828873e-02, 2.24575806e-01, 5.19783871e-01, 1.20457968e+02,
  1.86419274e+05]
Max Distance to Center: 3739.3553165430053
Min Distance to Center: 44.98123802800114
Average Distance to Center: 1984.4246599392038
SSE from Center: 308391475.9854835
62 Points:
Top genres: 47
------------------------------------------------------------

Cluster 17:
Center [ 4.21000000e+01, 6.89400000e-01, 7.03900000e-01, 4.90000000e+00,
 -6.57390000e+00, 5.00000000e-01, 1.09860000e-01, 3.00093000e-01,
  1.20627802e-01, 2.02890000e-01, 6.01800000e-01, 1.26669500e+02,
  2.45211100e+05]
Max Distance to Center: 592.74098774554
Min Distance to Center: 44.278228732602074
Average Distance to Center: 262.0803382351134
SSE from Center: 945631.985324186
10 Points:
Top genres: 143
------------------------------------------------------------

Cluster 18:
Center [ 5.1400000e+01, 6.2760000e-01, 7.3256000e-01, 6.8800000e+00,
 -5.4024400e+00, 6.4000000e-01, 6.8932000e-02, 8.7797080e-02,
  6.6454120e-04, 2.3888800e-01, 4.6868000e-01, 1.2953544e+02,
  1.9424372e+05]
Max Distance to Center: 1517.5152910476859
Min Distance to Center: 39.96009712151787
Average Distance to Center: 749.9233145616377
SSE from Center: 18057467.307230897
25 Points:
Top genres: 5
------------------------------------------------------------

Cluster 19:
Center [ 4.05625000e+01, 7.02062500e-01, 6.26125000e-01, 5.56250000e+00,
 -7.83593750e+00, 5.62500000e-01, 1.75318750e-01, 2.59330625e-01,
  4.19354219e-02, 2.27281250e-01, 5.06918750e-01, 1.30819312e+02,
  2.59533688e+05]
Max Distance to Center: 654.813943590256
Min Distance to Center: 57.32994361666628
```

```
Average Distance to Center: 309.41171846436725
SSE from Center: 1998615.1954590324
16 Points:
Top genres: 17
------------------------------------------------------------

Cluster 20:
Center [ 4.21666667e+01, 6.92833333e-01, 7.08166667e-01, 5.16666667e+00,
 -5.30783333e+00, 6.66666667e-01, 6.07166667e-02, 1.17116667e-01,
  2.20833333e-04, 1.56233333e-01, 5.66166667e-01, 1.21982333e+02,
  1.67948333e+05]
Max Distance to Center: 449.22506843453476
Min Distance to Center: 44.67907813228849
Average Distance to Center: 190.17670794169462
SSE from Center: 349486.229176838
6 Points:
Top genres: 28
------------------------------------------------------------

Cluster 21:
Center [ 4.11600000e+01, 6.53440000e-01, 6.52040000e-01, 4.64000000e+00,
 -7.64964000e+00, 5.20000000e-01, 5.44040000e-02, 1.36793744e-01,
  3.32412116e-02, 2.40816000e-01, 4.76960000e-01, 1.22212640e+02,
  2.42909920e+05]
Max Distance to Center: 1430.6211345007046
Min Distance to Center: 87.43364850483826
Average Distance to Center: 634.4404989678472
SSE from Center: 14335302.89571816
25 Points:
Top genres: 37
------------------------------------------------------------

Cluster 22:
Center [ 2.27333333e+01, 6.41733333e-01, 6.50466667e-01, 5.60000000e+00,
 -8.93933333e+00, 7.33333333e-01, 9.87666667e-02, 9.34816667e-02,
  2.80703033e-02, 2.21966667e-01, 5.22333333e-01, 1.10151533e+02,
  2.84151267e+05]
Max Distance to Center: 1001.385320704614
Min Distance to Center: 88.03659952699628
Average Distance to Center: 558.405740010435
SSE from Center: 5875408.219604032
15 Points:
Top genres: 27
------------------------------------------------------------

Cluster 23:
Center [ 5.6125000e+01, 6.0475000e-01, 7.0775000e-01, 5.7500000e+00,
 -7.2381250e+00, 7.5000000e-01, 9.0037500e-02, 3.0210375e-01,
  8.1723375e-03, 1.4436250e-01, 5.7150000e-01, 1.1028500e+02,
  3.0724750e+05]
Max Distance to Center: 665.4366606480149
Min Distance to Center: 93.08211758697212
Average Distance to Center: 280.46687634122264
SSE from Center: 886318.2568174861
8 Points:
Top genres: 236
------------------------------------------------------------

Cluster 24:
Center [ 4.2250000e+01, 7.1900000e-01, 7.2750000e-01, 6.5000000e+00,
 -6.6035000e+00, 5.0000000e-01, 2.4841250e-01, 1.1873175e-01,
  3.6591250e-05, 1.0401250e-01, 5.2350000e-01, 1.2859450e+02,
  1.6670000e+05]
Max Distance to Center: 527.2428466123799
Min Distance to Center: 34.086933809932496
Average Distance to Center: 254.18838938322017
SSE from Center: 739401.9037027456
8 Points:
Top genres: 411
------------------------------------------------------------

epsilon:  500.0  and minpts:  5
percent of noise points:  32.03125 %
percent of border points:  8.7890625 %
percent of clustered points:  59.1796875 %
```

9.3. Question 4

9.3.1. Random Forest

	TP	FP	FN	TN
EDM	4048	2334	1995	24455
Latin	1865	2607	3290	25070
Pop	1578	3001	3928	24325
R&B	2080	3018	3351	24383
Rap	3562	3182	2184	23904
Rock	3347	2210	1604	25671

Table 15: The confusion matrix from running Random Forest on all nine attributes

9.3.2. C4.5

	TP	FP	FN	TN
EDM	692	526	596	2140
Latin	351	682	661	2481
Pop	287	794	844	2545
R&B	374	657	696	2458
Rap	532	636	545	2300
Rock	596	440	393	2236

Table 16: The confusion matrix from running Random Forest on all nine attributes