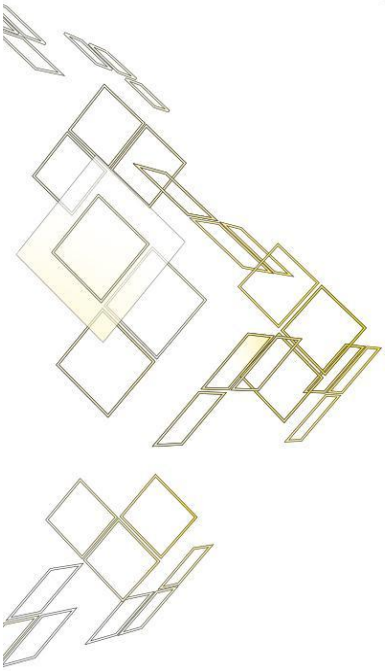
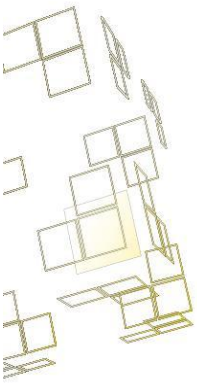
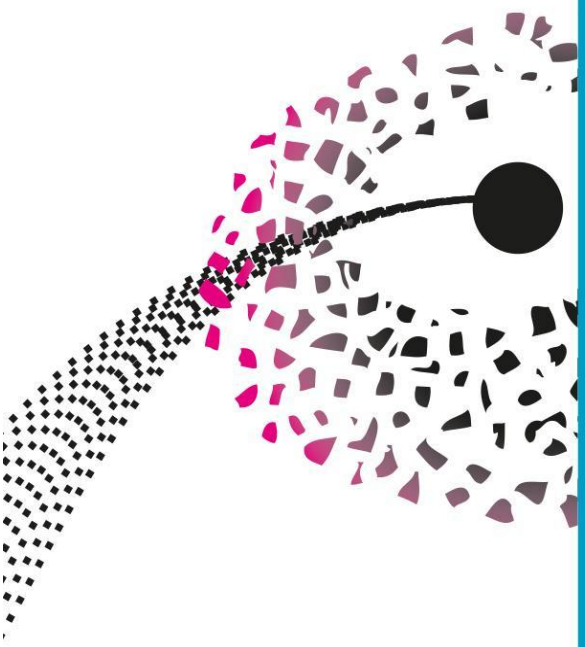






Music Genre Prediction Using Acoustic Features

Million Song & LastFM datasets

Group 17:

Dylan Pham - s2845016
Raef Kazi - s2733226
Silvi Fitria - s2800209
Kristen Phan - s2616874

Managing Big Data (201200044)
Teacher: dr. Doina Bucur

UNIVERSITY OF TWENTE.

1. Introduction

According to a recent listening attitudes study, there was an approximately 50% likelihood that respondents have heard music in the previous two hours at any randomly chosen time between 8 a.m. and 10 p.m (Greb et al., 2019). On the other hand, music had a 2% probability of being the major focus of their attention. Therefore, music plays an important role in society and daily life as a hobby, entertainment, and career.

Everyone has their personal preferences and interest about the music genre they want to listen to. Most of them moved to online music platforms to listen to their favorite music, such as Spotify, Soundcloud, Pandora, and Youtube Music. Given this freedom of choice, people tend to actively select and use music to accomplish specific goals in certain situations (Krause et al., 2015). These businesses must keep their members to generate additional income through subscriptions. These streaming applications have a unique feature called a playlist that listeners find it challenging to make a playlist from a large collection of songs.

In contrast to music listeners' broad adoption of new technical innovations, not all know about the mechanisms that underpin music choosing in everyday life. Scientific study on music listening in everyday life is still in its infancy (Barone, 2017). Thus, this research concerns the degree to which a person’s preferred musical genre's acoustic features impact their song or track selection. This question, while theoretically straightforward, covers active research areas in the disciplines of music cognition and music information retrieval (MIR). By analyzing the acoustic features of music, we hope to identify a preferred genre influencing song selection. As a result, this prediction may develop music recommendations in further research. Aside from that, this research would be beneficial for feature tracing to determine which traits influence certain genre predictions.

For those reasons, we conducted predictive analysis on the primary research topics that we intend to be addressed in this study:

- Can the genre of a song be predicted only by its acoustic features?

To answer the main research question, we compare various machine learning algorithms: Logistic Regression, Linear Support Vector Classification, Naive Bayes Classifier, Decision Tree Classification, and Gradient Boosting Classification. We trained those algorithms on UT's cluster using the One Million Song Dataset and LastFM.

This research offers the following utilities:

- Enables auto-tagging of genres for platforms dealing with big data, for instance, streaming services and video sharing platforms.
- Provides a baseline work for constructing an assembled model that combines multiple machine learning models to improve accuracy

2. Related Work

The study of music genre classification has been ongoing for almost two decades, with various variations in the methodology and approach to these problems. These attempts, most notable, have been undertaken with small data sizes (relative to big data) to test certain models and hypotheses, the reproducibility of which is a limiting factor on a larger scale.

The first attempt into music genre classification was made by Tzanetakis and Cook (2002). They worked with 30 feature vectors extracted from audio signals, using a dataset comprising only 1000 songs spread evenly across ten genres. This dataset, referred to as the GTZAN dataset, became the most popular dataset for music genre classification

problems over the years. Although this was a seminal work in the MGC domain, the size of the dataset is nowhere close to the degree of which we have today.

Most existing works on the One Million Dataset use only a small subset of the dataset to train machine learning models in a non-distributed manner. Jiang et al. (2017) studied a recommendation system using Recurrent Neural Network (RNN) to predict the user's next most possible song by similarity. They made evaluations using the Million Song Dataset and demonstrated how it outperformed the conventional approaches. Audio and lyrics data have been used for experiments, and they are also combined with the LastFM dataset. To process 34,412 songs, they queried using API and paid close attention to how they distributed it.

Another study using this data was conducted by Nysater & Reinhammar (2013) to investigate the possibility of automatically classifying the similarity of song pairs. They utilized K-Nearest Neighbors combined with bootstrap aggregating and an attribute selection classifier to classify the music genre. Several acoustic features were utilized but only implemented to 100 songs dataset.

Liang et al. (2011) also used 300GB audio features and metadata Million Song Dataset to classify music genres. They proposed a cross-modal retrieval framework of model blending, which combines features from audio and lyrics. The results showed that the blending features perform better than any individual one and are suitable for careful testing and analysis of what different submodels contribute. All of the mentioned research did not explain how to distribute the big data and only focused on the algorithms.

3. Methodology

3.1 Dataset Descriptions

The data used comes from 2 datasets: 1) The Million Song Dataset, which contains audio features from 1 million songs, which are used as the input feature vectors to make the prediction, and 2) the Last.fm dataset, which contains the genre data for these songs, which are used as the output vector to train the data. Both these datasets share the same “track_id” attribute for each song.

The Million Song Dataset was extracted from the audio provided by The Echo Nest. Million Song Dataset was released by Bertin-Mahieux et al. (2011), which consists of more than a million audio features, metadata, and unique track IDs. This dataset has a significant benefit because all sources have unique linkages to various additional data sources, including LastFM. It is stored on the UT cluster in 26 CSV files totaling about 676 GB in size. The dataset contains 44 acoustic features and metadata about each song, without the actual audio of the song present in it. A well-known issue about the Million Song Dataset is the duplicate entries. There are, in total, 78,190 duplicate songs, which result in a total dataset of 921,810 songs.

The LastFM dataset was taken from the website, containing around 1,2 million JSON files that are 2.34 GB in total size. From this, we use two attributes, including track_id and tags.

3.1.1 About the Genres

The LastFM dataset contains **522,366** unique tags, many of which are sub-genres of a particular genre (e.g., indie rock, pop-rock, folk-rock, etc.), and many of which are location-specific (Latin, Canadian, german) or niche-specific, such that we cannot accurately extrapolate them for future songs. It is, thus, unfeasible to try and predict every single genre. For this purpose, we decided to choose the top 10 genres present in the dataset, namely: rock, pop, alternative, indie, electronic, soul, dance, metal, jazz, instrumental.

3.2 Data Pre-processing

Figure 1 illustrates the general data pre-processing pipeline, starting from left to right

- Processing Million Song Dataset (MSD) stored on UT's cluster: we remove duplicates, drop non-feature columns
- Reading Last.fm dataset from source http://www.millionsongdataset.com/lastfm/ into UT's cluster
- Processing Last.fm dataset: each song in this dataset comes with a set of genres and genre confidence. For example, the song "Let It Be" by the Beatles having the genre "rock" with a confidence level of 80% means that the algorithm used to classify the genres of this song yields an 80% confidence that the song should be classified as rock. We keep only genres with at least 20% confidence and subsequently choose the top 10 genres, which we will attempt to predict using machine learning models.
- Merging MSD and Last.fm: the result dataset combines the song features with genres.
- Processing features: We process three types of feature data in our dataset differently: one hot encoding categorical data, normalizing numerical data, and bucketing to timestamps time-series data.
    - Examples of categorical data: key - what key the song is in, e.g., E major.
    - Examples of numerical data: tempo - how many beats in a minute the song has
    - Examples of time-series data: segment pitches - what is the pitch of a song segment (e.g., a segment can last a few seconds)

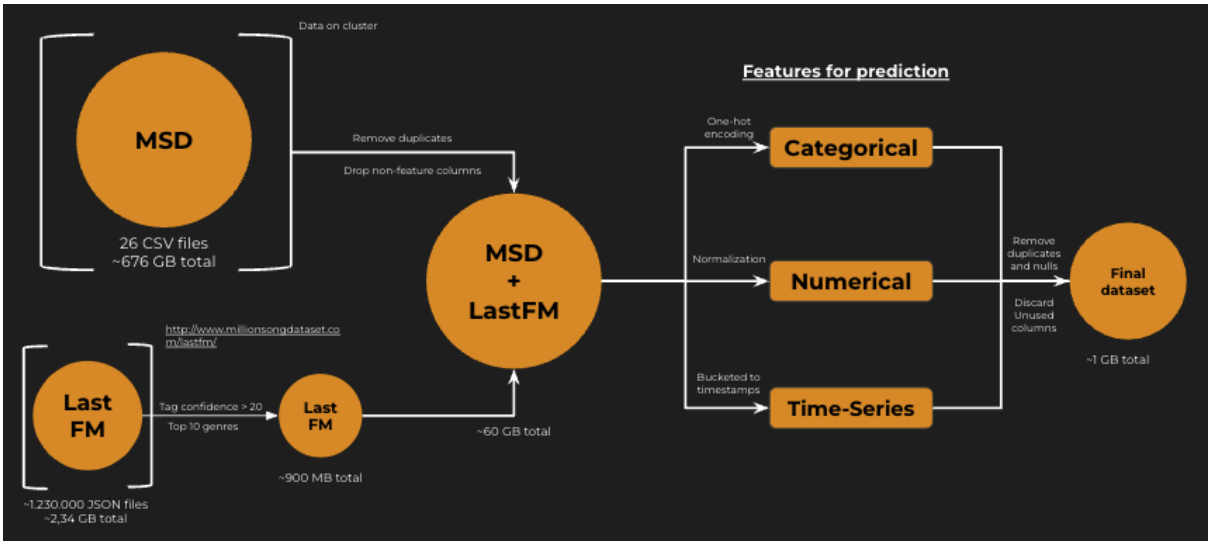


Figure 1. Flowchart of Data Processing

3.3 Machine Learning Models

Using the features from the final dataset from the pre-processing pipeline, we trained five different machine learning models:

- Logistic regression
- Linear SVC
- Naive Bayes classifier
- Decision tree classifier

- Gradient boosting classifier

We will predict each genre separately (binary classification). Since the problem is binary classification, any model that raises prediction accuracy less than 50% is considered unusable. We accept the model with at least 66% accuracy, which is two correct answers out of three guesses. Each model is trained on 70% of the dataset and tested on 20%. Model performance is evaluated using accuracy level.

4. Results

In this section, we depict the key results of our study objectives. By running five machine learning algorithms, we compare evaluation models shown in Table 1.

Genre	Logistic Regression	Linear SVC	Naive Bayes Classifier	Decision Tree Classifier	Gradient Boosting Classification
Alternative	84.06%	84.16%	36.34%	77.30%	83.56%
Dance	93.18%	93.48%	79.21%	90.56%	93.18%
Electronic	88.80%	88.05%	76.22%	86.29%	88.74%
Indie	84.98%	84.99%	38.74%	78.68%	84.31%
Instrumental	96.59%	96.59%	76.96%	94.17%	96.18%
Jazz	94.47%	94.50%	74.45%	90.70%	94.00%
Metal	94.08%	94.07%	70.15%	90.76%	93.66%
Pop	82.61%	82.61%	35.85%	75.39%	82.24%
Rock	67.94%	63.77%	47.07%	62.41%	67.78%
Soul	93.32%	93.32%	58.77%	87.25%	93.04%
Average	**88.00%**	**87.55%**	**59.38%**	**83.35%**	**87.67%**

Table 1. Comparison of Model Accuracy (Non-weighted version)

Figure 2 and Table 2 visualize the performance of different models. When averaging the non-weighted accuracy level of different models, logistic regression comes out first with the highest accuracy level of 88%, while Naive Bayes comes out last.

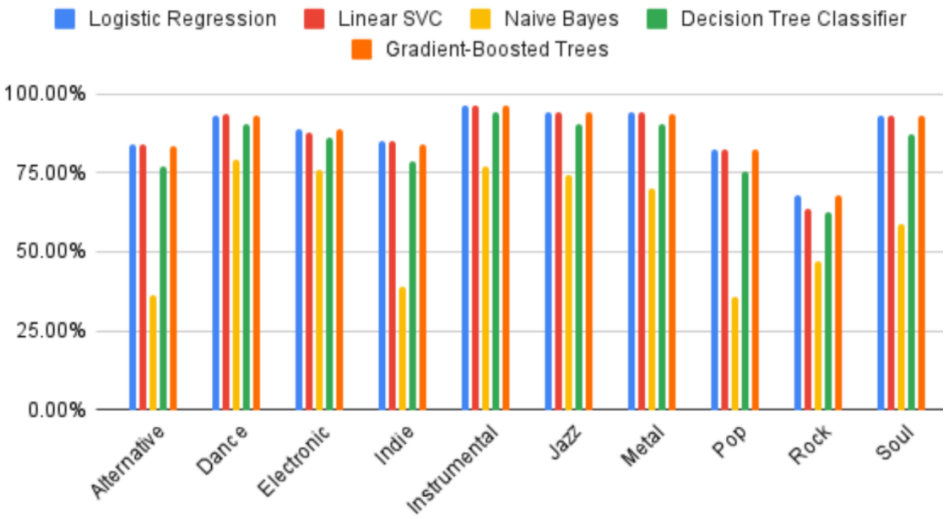


Figure 2. Genre prediction accuracy by model

Algorithm	Accuracy Score
Logistic Regression	82.02%
Linear SVC	80.77%
Naive Bayes	51.38%
Decision Tree Classifier	76.68%
Gradient-Boosted Tree	81.70%

Table 2. Average Accuracy Score (Weighted by the number of songs per genre)

Figure 3 provides another interesting insight. Despite accounting for the largest portion of the dataset, the rock genre cannot be predicted using machine learning models and other genres despite accounting for the largest portion of the dataset. This goes against the intuition that the more data we can feed into machine learning models, the better the models should perform. We speculate that this low performance is since we have to aggregate multiple rock subgenres (e.g., metal rock, soft rock) into a genre category of “rock.” The machine learning models cannot predict the rock genre well because of the diffusion of distinctive features of different rock subgenres.

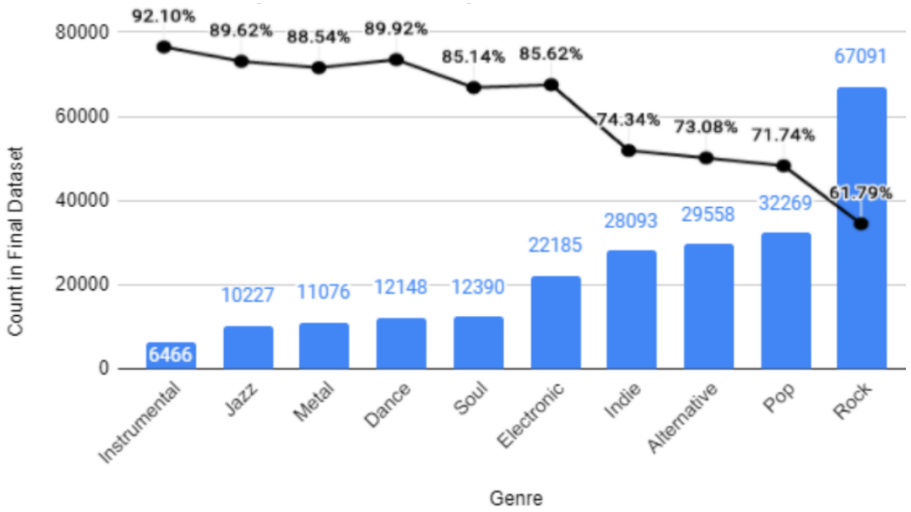


Figure 3. Average prediction accuracy level by genre.

However, when we further study the algorithm, we notice an imbalance between positive data (1's labeled records) and negative data (0's labeled records). Since the imbalanced training data can affect the prediction result, we record the percentage of 0 and 1 labels of testing data for both cases, ground truth, and prediction for each algorithm. Table 4 indicates the result. The three most accurate models (Linear Regression, Linear SVC, Gradient-Boosted Classifier) all suffer from imbalance prediction, where most of the prediction results are 0. Since the label 0 has the majority in the testing dataset, the model can raise high accuracy by predicting every record as 0. Still, the actual application of the model will not be accurate. Surprisingly, the most well-rounded model is the Decision Tree Classifier, which has the weighted accuracy of 76.68% but has the most balanced prediction compared to ground truth.

Table 4. Percentage of actual and predicted labels

	Actual No of 0s	Actual No of 1s	Predicted No of 0s	Predicted No of 1s
Linear Regression	87.56%	12.44%	97.35%	2.65%
Linear SVC	87.56%	12.44%	100.00%	0.00%
Naive Bayes	87.56%	12.44%	55.02%	44.98%
Decision Tree Classifier	87.56%	12.44%	90.14%	9.86%
Gradient-Boosted Classifier	87.56%	12.44%	95.76%	4.24%

Additionally, to address the data imbalance problem, during the model training session, we balance the number of 0 and 1 labeled records for each genre by removing random records of the larger side. Since we want to improve the prediction accuracy of the three most accurate algorithms, we only test with Logistic Regression, Linear SVC, and Gradient-Boosted Classification. Table 5 indicates the testing result of the balanced training data (testing data is kept the same). The result shows that balancing the training data reduces the prediction accuracy significantly. We assume that the remaining data after balancing does not contain enough information to train the model, thus reducing the accuracy. Moreover, these three algorithms tend to balance percentage positive and negative prediction results; thus, the prediction will be half positive half negative, which is wrong since Table 4 indicates the other insight.

Table 5. Comparison of Model Accuracy with balance data (Non-weighted version)

Genre	Logistic Regression	Linear SVC	Gradient Boosting Classification
Alternative	62.71%	55.36%	60.58%
Dance	78.60%	59.04%	77.05%
Electronic	75.60%	71.53%	76.66%
Indie	58.48%	51.12%	57.39%
Instrumental	65.64%	73.79%	66.37%
Jazz	69.90%	84.46%	66.79%
Metal	73.86%	80.50%	71.84%
Pop	54.55%	32.21%	55.53%
Rock	65.52%	57.10%	63.61%
Soul	62.14%	52.07%	59.57%
Average	**66.70%**	**61.72%**	**65.54%**

5. Conclusion and Discussion

In this report, we have shown that acoustic features can roughly predict the genres of a song. The Million Song Dataset was well structured and able to be analyzed. It links to the LastFM dataset to combine them to get the best prediction. Using these two datasets, we managed to get the prediction accuracy of 76.68% for genre binary classification using the Decision Tree Classifier. However, this report suffers from some limitations that need to be addressed in the future to acquire better accuracy.

Firstly, due to the subjective nature of genre classification [9], we cannot expect perfect results from automatic predictions or human annotations. Given this caveat, our model can only prove the existence of relations between acoustic features and the related genres.

Secondly, to simplify and filter out redundant genres, we selected the top ten genres out of 522,366 genres represented in the LastFM dataset. Although this provides practical advantages, it does mean that some genres, despite having a lower frequency in the dataset, are not correctly represented in the selected ten genres.

Thirdly, running more complex machine learning algorithms on the big data cluster was time-consuming. The cluster is not optimized for deep learning, and moving to a dedicated deep learning cluster was expensive for the given timeframe and scope. Due to this limitation, we could not test deep learning models on the dataset.

Fourthly, the accuracy metric is not a suitable measurement for imbalanced data. We need to evaluate the importance of False Positive and False Negative prediction for this type of data, then decide to use Precision or Recall, respectively. However, since our project evaluates False Positive and False Negative predictions to have the same significance, we choose to keep accuracy as the primary measurement. Since this issue solely depends on applications, future works need to address this.

6. Appendix

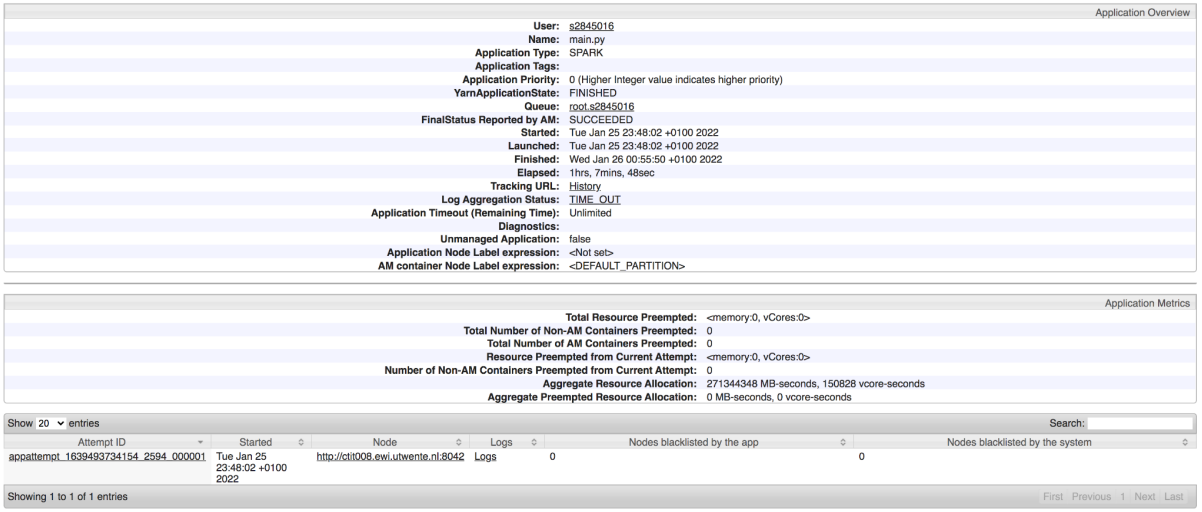
[Github Repository]

The procedure for running the code is contained in the “Final” folder. Please consult the ReadMe.md for more information.

[Million Songs Dataset]

[Last.fm Dataset]

Snapshot of running session on Spark:



Application Overview

User:	s2845016
Name:	main.py
Application Type:	SPARK
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.s2845016
FinalStatus Reported by AM:	SUCCEEDED
Started:	Tue Jan 25 23:48:02 +0100 2022
Launched:	Tue Jan 25 23:48:02 +0100 2022
Finished:	Wed Jan 26 00:55:50 +0100 2022
Elapsed:	1hrs, 7mins, 48sec
Tracking URL:	History
Log Aggregation Status:	TIME_OUT
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	271344348 MB-seconds, 150828 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1639493734154_2594_000001	Tue Jan 25 23:48:02 +0100 2022	http://ctit008.ewi.utwente.nl:8042	Logs	0	0

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Image 1. Multiple models (5) prediction running status with clean data (1GB)

7. References

Barone, M. D., Bansal, J., & Woolhouse, M. H. (2017). Acoustic Features Influence Musical Choices Across Multiple Genres. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00931

Bertin-Mahieux, T., Ellis, D. P. W., Wandamere P. (2011). The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference.*

G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," in *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, July 2002, doi: 10.1109/TSA.2002.800560.

Greb, F., Steffens, J., & Schlotz, W. (2019). Modeling Music-Selection Behavior in Everyday Life: A Multilevel Statistical Learning Approach and Mediation Analysis of Experience Sampling Data. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00390

Jiang, M. Yang, Z., & Zhao, C. (2017). What to play next? A RNN-based music recommendation system. *51st Asilomar Conference on Signals, Systems, and Computers*, pp. 356-358, doi: 10.1109/ACSSC.2017.8335200.

Krause, A. E., North, A. C., and Hewitt, L. Y. (2015). Music-listening in everyday life: devices and choice. *Psychol. Music* 43, 155–170. doi: 10.1177/0305735613496860

Liang, D., Gu, H., & O’Connor, B. (2011). Music Genre Classification with the Million Song Dataset. Carnegie Mellon University.

Lippens, Stefaan & Martens, jean-pierre & De Mulder, Tom & Tzanetakis, George. (2004). A Comparison of Human and Automatic Musical Genre Classification. 4.

Nysater, R. & Reinhammar, T. (2013). Song Similarity Classification Using music Information Retrieval on the Million Song Dataset. Royal Institute of Technology.