Hit Song Science: a genre based approach to Hit song prediction on acoustic features

1st Alex Santini
National College of Ireland Dublin, Dublin
Email: 1st x21211604@student.ncirl.ie,

2nd Enda Stafford
National College of Ireland Dublin, Dublin
Email: 2nd enda.stafford@ncirl.ie

Abstract — Thanks to recent developments; Music Information Retrieval (MIR) technologies are now starting to being used throughout the music industry with brilliant results especially in semantic tagging and recommendation approaches thanks to large amount of data that platforms such as Spotify, Bandcamp and iTunes can collect.

However MIR as a whole remains yet an immature field of research whit a lot to of question unanswered and a lot to discover; an important piece of this puzzle still remain unsolved: over the years many attempted to predict song popularity using ML models, a field sometime called "Hit-song science", yet still a comprehensive answer has not been offered with some researchers relegating the field to a dead-end while other being more possibilistic.

Different approaches have been tried to crack the code of what makes a song a true Hit such as: Text-analysis, Metadata prediction and Music features exploitation.

Aim of this project is to test and compare the classification accuracy of five different types of machine learning algorithms to answer the following question: To what extent the popularity of a song within a certain music genre can be predicted only by its acoustic features?

Thanks to the use of the recent MusicOset dataset (2019) this project tries to investigate a new genre-based approach hoping to successfully answer it.

The algorithms been used are Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Decision Tree Classifier (DTC).

Each model has been trained and tested upon three different genres: "album rock", "Dance pop", "contemporary country"; the results on those three different genre subsets have been compared to one with a similar sized dataset with multiple genres in it.

Based on the results, none of the five algorithms have been able to classify popularity better than on a standard dataset.

Overall, only considering acoustic features appears to not suffice in terms of hit song prediction, however the results shown scores next to 60% accuracy (reached on the standard dataset) which implies that acoustic features can be informative and may be used alongside other data points in hit song prediction.

I. RELATED WORK

A. Introduction

Every year Record Labels spend huge number of resources in developing new artists, songs and albums. According to [1], in 2021 the entire music industry recorded more than 25 billion dollars in revenue. Developing tools which will be able to predict and capitalize shares of this market is seen by many as a real golden goose.

B. Background Information and literature review

Music Information Retrieval (MIR) is a science dedicated to retrieving information from music where researchers have multivariate backgrounds from musicology to machine learning and computer science.

In recent years MIR has been growing exponentially thanks to lots of recent developments creating viable industry applications.

Techniques as user-centric recommendation tools and semantic auto-tagging and are now tools frequently used in day-to-day life in the media industry [1].

MIR systems [2] are now being used for music identification in application such as Shazam, for recommendation purpose on YouTube and for plagiarism and copyrights checks.

When interacting with music, developers and scientists are dealing with a constantly evolving creature which can be represented in many different ways and they have to overcome many challenges both technically and culturally [17].

Using the same techniques, a little branch of MIR is devoted to predicting and define the best ways to predict song popularity; a field sometime called "Hit song Science" (HSS).

As today many Researchers have been trying to produce effective models to separate hit songs from the rest, yet HSS remains an unexplored and immature field [2].

One important contribution in HSS came from F. Pachet where in [3][4] demonstrated how the claims from the MIR community regarding the ability of predicting popularity from acoustics and human features in a song dataset were impossible to achieve

through a multilabel classification between low, medium and high popularity.

Following the paper researchers have been split between two sides, one employing Pachet's point of view while other trying to prove him wrong: in [5] have been proved that algorithms were able to detect hit songs more accurately than a random oracle.

With a different approach, these researchers were able to differentiate the top 5 Hits between the rest of the Top 30-40 UK using a "Shifting preceptor algorithm" accounting for changes in taste and perception throughout the decades showing more optimistic result.

Another study conducted by Borg and Hokkanen [6] seem to agree with Pachet considering the acoustic features as not informative when predicting popularity, indicating how popularity might be driven mostly by social and cultural aspects.

Work done by [7] in contrast, using SVM on a multivariate dataset of acoustic features and lyrics achieved better result showing a classifier with higher score than a random one.

Reinforcing a possibilistic view on HSS the authors in [8] took the first genre-based approach comparing different algorithms in just a specific genre showing how Logistic regression was performing better than a random classifier on a mixture of acoustic and temporal features.

Against the idea of acoustic features being informative, the conclusions drawn in [9] depict a probable reason for it in the variation in acoustic features being too high within a single song making difficult to extract metrics therefore modelling from them, however the paper reached good scores in classification using metadata features.

Another great result using a mix of metadata and intrinsic song features comes from [10] where authors have been able to predict the popularity of songs 2 months after the data collection with a score of more than 80% while in [18] authors demonstrated how charts metadata and music features can be synergic when predicting popularity.

Also [11] found how modelling classifier only using acoustic features did not achieved the accuracy values hoped, even though the author stated that the low values achieved might have been the result of lack of features selection and optimization modelling; however, suggested how more data would be beneficial especially in terms of genre label to mitigate the variation in acoustic features.

C. Motivation and Scope

As mentioned, the discussion around HSS as a viable path of discovery is still open and researchers have tried different pathways in predicting song's popularity.

It is hoped that with a more recent dataset [12] this project will have a final word relatively to the use of acoustic features in popularity prediction by comparing accuracy scores for different ML models on 3 different genre subsets "Album rock, Dance and Contemporary country"; a genre-based approach will reduce the variation in acoustic features within each dataset ultimately answering two fundamental questions:

1. Hit Song Science: what's the position of the author of this project on the ongoing discussion about the possibility of predicting song popularity?

2. Features: Are acoustic features informative on popularity prediction when taken by themselves or (some of them) might have a supplementary role but without explaining much when taken alone?

II. RESEARCH METHODOLOGY

A. Data Understanding

The data has been downloaded from the MusicOSet website [13]. The System used for this research project was as follow:

- Processor: 11th Gen Intel (R) Core (TM) i5-11400H @ 2.70GHz 2.69 GHz
- RAM 16.0 GB
- Edition: Windows 11 Home

The dataset was firstly loaded into MySQL environment for initial data exploration and understanding of the overall shape of the data.

The dataset consists of more than 20.000 songs and 11.000 artists with information regarding popularity from Hot 100 and Billboard charts; data regarding songs structured and acoustic features, lyrics and much more.

A SQL view has been created where the "main genre" attribute from the artist table has been linked to the song table; by doing this the genre of the artist will act as a proxy for categorizing songs into different bins and reducing the variation among the groups in which the ML modelling will take place.

The SQL view has been exported and loaded into python and shows the following shape: 11296 rows (not every song in the dataset had a main genre associated) and 22 features both categorical and continuous, data types and relative information from the MusicOSet notes [13] are listed here:

Categorical features:

- **Song type**: Solo or collaboration.
- **Main genre**: Main genre the artist is associated with.
- **Key**: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. If no key was detected, the value is -1.
- **Mode**: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- **Time signature**: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- **Explicit**: if the song contains explicit lyrics

Continuous features:

- **Duration**: The duration of the track in milliseconds.
- **Acousticness**: how much a track is acoustic. 1.0 represents high confidence the track is acoustic.
- **Danceability**: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **Energy**: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- **Instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- **Liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- **Loudness**: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
- **Speechiness**: detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- **Valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- **Tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **Popularity:** The popularity of the track. The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in

the past. Artist and album popularity is derived mathematically from track popularity.

It is clear how the strength of this project relays mostly on the value that the enhanced MusicOset dataset brings.

B. Data Preparation

Fortunately, the data quality of the MusicOSet was excellent and 0 missing values were found, however the distribution of the various features was in some cases profoundly skewed.

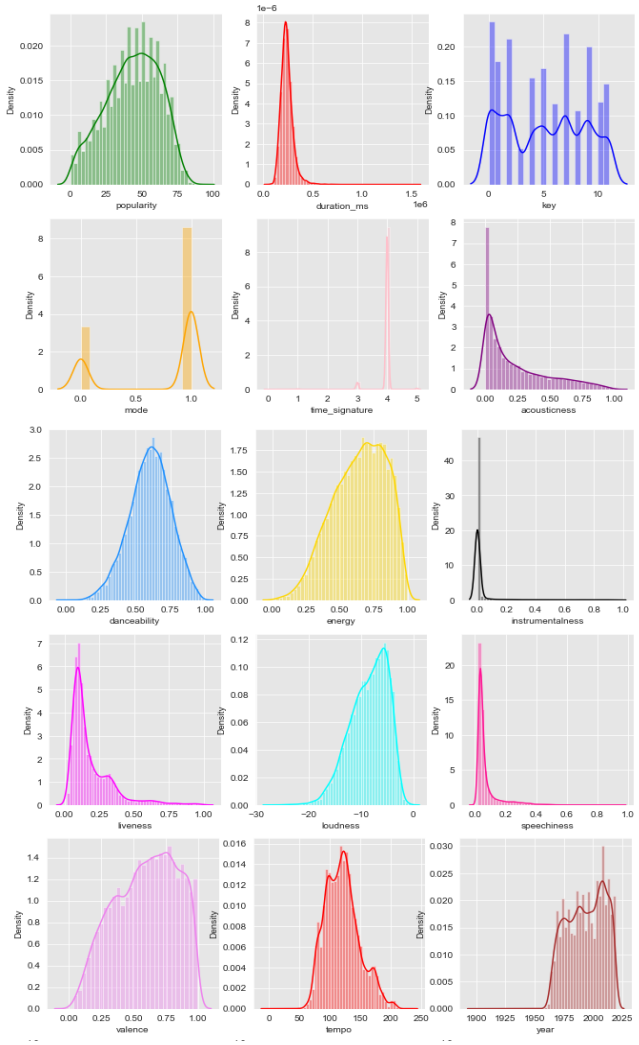

Figure 1. Features frequency distribution

Subsets of genres have been created for the groups with the higher number of entries.

main_genre	count(song_id)
album rock	1596
dance pop	1219
contemporary country	965
adult standards	904
classic soul	454
-	451
brill building pop	320
alternative metal	271
disco	255
atl hip hop	227
funk	183
dance rock	151
bubblegum pop	150

Figure 2. Count of songs by genre

For features with higher skewness: within each genre group outliers have been detected and removed by using the following formula:

$$\text{Low} = Q1 - 1.5*IQR$$
$$\text{High} = Q3 + 1.5*IQR$$

The reason for dropping the missing values were because of the low number of outliers, although it could have been replaced with the mode or median the data was large enough for the analysis.

The definition of "Hit song" has been decided on the 75th quantile in popularity score per each genre; alongside with previous research definitions [11].

Class imbalance between Hit and non-Hit was quite strong, under-sampling and SMOTE techniques have been adopted on a case-by-case basis upon creating the classifiers.

The same data transformation approach has been adopted for the larger comparison dataset.

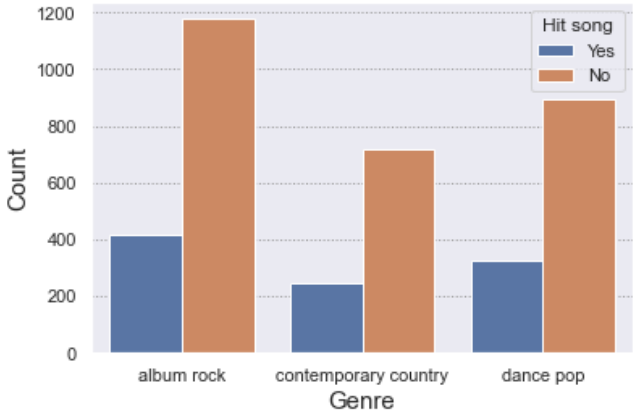

Figure 3. Class imbalance between Hit and non-Hit songs on genre based datasets

C. Statistical and Trend analysis

The following statistical analysis were visualized, figure 4 shows how the popularity is distributed in years per genre with the "album rock" label being more prominent until the '90 and the other two genres ruling in recent decades.

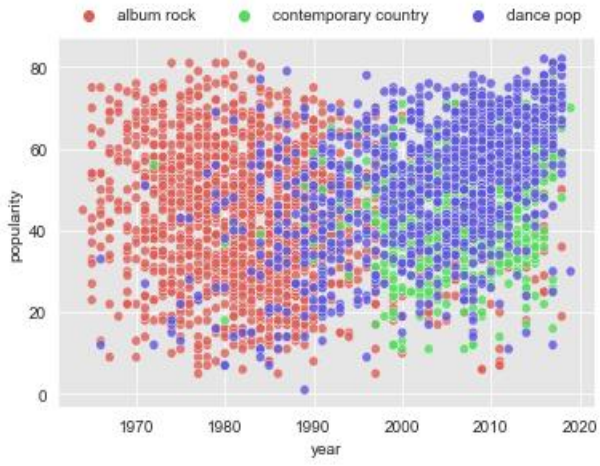

Figure 4. Popularity in years group by genre

A sequence of trend analysis has been drawn.

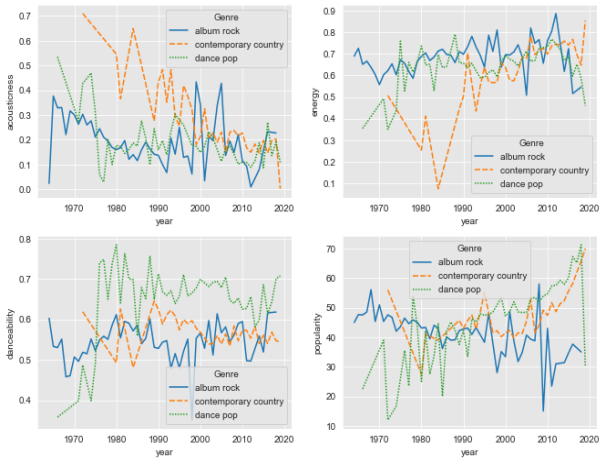

Figure 5. Trend analysis of features over years

Figure 5 shows the variation over time of some features per each genre, it's apparent how trends over time are shared by different genres in certain cases while not in others.

For some feature multi-dimension plotting has been helpful for understanding and better modelling the tools.

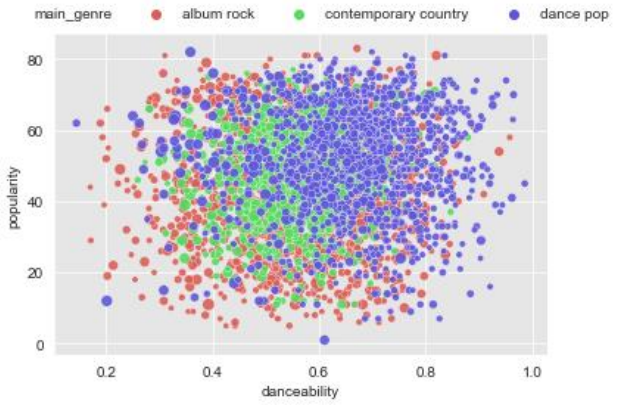

Figure 6. Danceability and popularity sized by tempo

In Figure 7 a 3D plot has been created summarizing four of the total features present in the dataset in relation to the Hit song label.

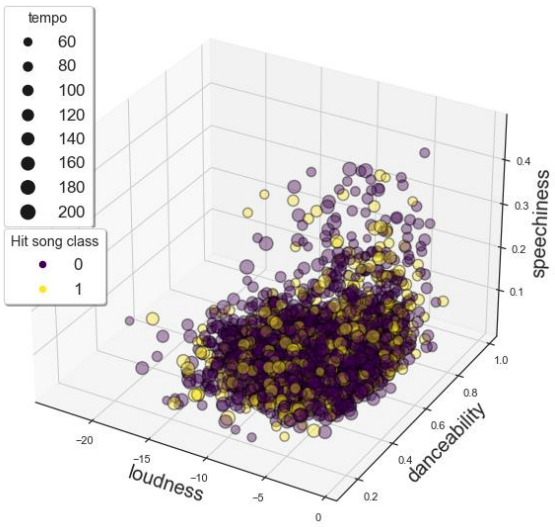

Figure 7. Loudness, Speechiness, Danceability and tempo in relation to Hit Song label

D. Feature Selection

After extracting categorical variables the dataset had a high number of dimensions, which may cause issues; the number of features has been reduced to achieve better accuracy and computational efficiency.

Feature selection is also import in order to reduce multicollinearity which has been detected by drawing a feature correlation matrix.

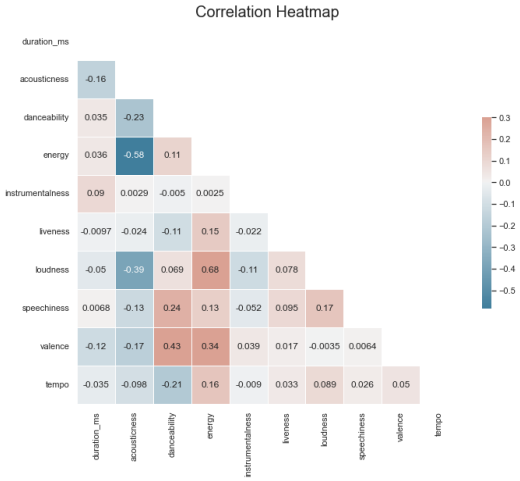

Figure 8. Features correlation heatmap

Feature extraction has been done by employing PCA on each genre subset and then selecting the number of components needed to explain at least 85% of the variation.

	no. components	*Variation explained*
album rock	9	86%
dance pop	9	85%
contemporary country	8	86%

Figure 9. number of Components per genre

III. MODELLING

A. General approach and evaluation metrics

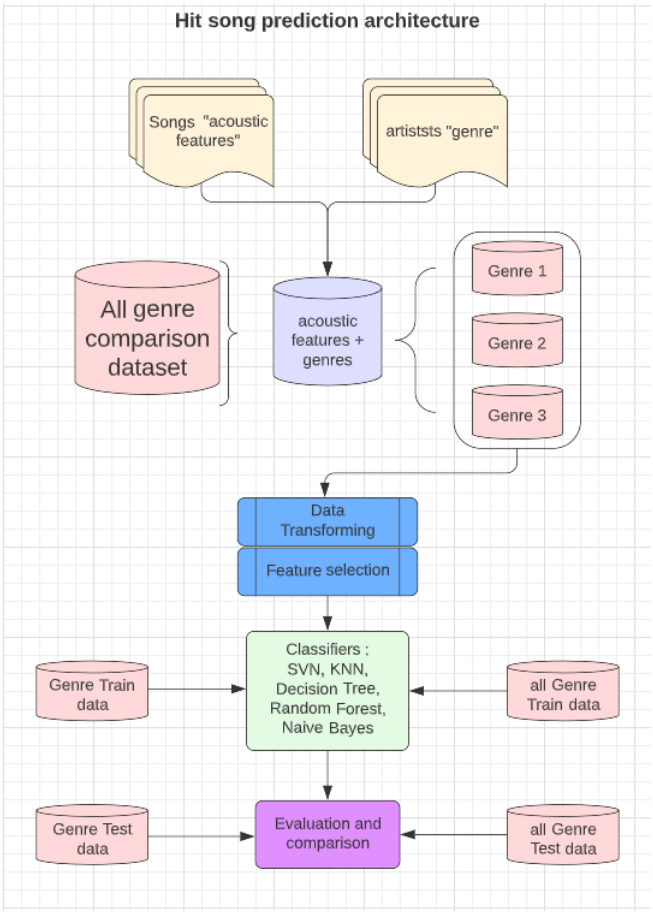

Figure 10. Architecture Design

Five algorithms (SVM, NB, Decision Tree, Random Forest, KNN) have been used on the processed data; each with a 5-fold cross validation method.

The building of the SVM model took the longest time to train with 2 minutes along with the Random Forest while other classifiers took no longer than some seconds.

The programming language used to model train and test the various classification algorithms was Python.

To manage the class imbalance within the classes under sampling and oversampling methods have been implemented on a case-by-case basis.

The generated confusion matrixes were used in determining the accuracy, recall and precision and F1 score of the models; these are the four metrics used to compare the various algorithms.

B. Genre classification

As validation and testing for the following genre-based approach in Hit song prediction, a series of algorithms has been run to check if ML techniques can differentiate between genres using acoustic features; this will act as proof that the three subsets: 'Album rock', 'Dance pop' and 'Contemporary country' have indeed different profiles in terms of acoustic features and those features are informative in genre classification.

Genre classification based on audio signals in MIR has been already proven worth pursuing by [14] and [15] in early 2000 and confirmed in recent years by [6] and [16] achieving accuracy scores around 85%.

Among the five classifiers used (Support vector machine, Random Forest, Naïve Bayes, K-Nearest Neighbors, Decision Tree) Random Forest and SVM performed the best.

	KNN	SVM	NB	Decision Tree	Random Forest
Precision	0.70	**0.76**	0.60	0.65	0.74
Recall	0.70	**0.76**	0.50	0.64	0.72
Accuracy	0.71	**0.76**	0.44	0.65	0.74
F1 Score	0.69	**0.75**	0.42	0.64	0.73

Figure 11. Genre classification scores (0 to 1)

SVM	*album rock*	*dance pop*	*contemporary country*
album rock	*357*	*35*	*53*
dance pop	*52*	*263*	*40*
contemporary country	*40*	*36*	*217*

Figure 12. SVM Genre classification Confusion Matrix

As shown algorithms perform well in terms of genre differentiation which led to think that acoustic features by themselves can be informative under the right circumstances.

C. Hit song prediction and evaluation of the models

The data in each genre subset has been splitted with a 70-30 ratio between train and test data.

As explained previously, models have been created and trained with a binary approach: each song had a label of "0" if below the 75th percentile, "1" if above.

Data has been gathered and the following results were shown in Table.

		Album Rock	*Dance Pop*	*Contemporary Country*
KNN	*Precision*	48.55%	51.65%	59.09%
	Recall	48.35%	51.98%	61.30%
	Accuracy	61.08%	56.41%	63.55%
	F1 score	22.62%	34.62%	44.78%
SVM	*Precision*	55.78%	55.56%	58.23%
	Recall	57.23%	56.62%	59.92%
	Accuracy	59.58%	56.41%	61.58%
	F1 score	40.53%	43.96%	**45.07%**
Naïve Bayes	*Precision*	50.13%	52.02%	54.53%
	Recall	50.16%	52.47%	55.41%
	Accuracy	50.00%	52.56%	46.80%
	F1 score	36.50%	38.67%	41.30%
Decision Tree	*Precision*	56.24%	50.91%	**59.47%**
	Recall	56.80%	51.16%	**61.61%**
	Accuracy	**68.26%**	51.71%	46.31%
	F1 score	33.75%	35.43%	41.71%
Random Forest	*Precision*	54.56%	54.81%	56.34%
	Recall	54.93%	55.83%	58.23%
	Accuracy	65.87%	61.11%	62.56%
	F1 score	32.14%	37.24%	38.71%

Figure 13. Models' evaluation (highest scores per metric in bold)

The results of the five classifiers were evaluated using confusion matrix, Accuracy, F1, recall and precision score.

Decision Tree seems to have the highest results for some metrics among the models while SVM agglomerates the best scores overall; however even in the best performing instances (mostly on Contemporary Country subset) the results were not high enough to justify the response of the algorithms as successful.

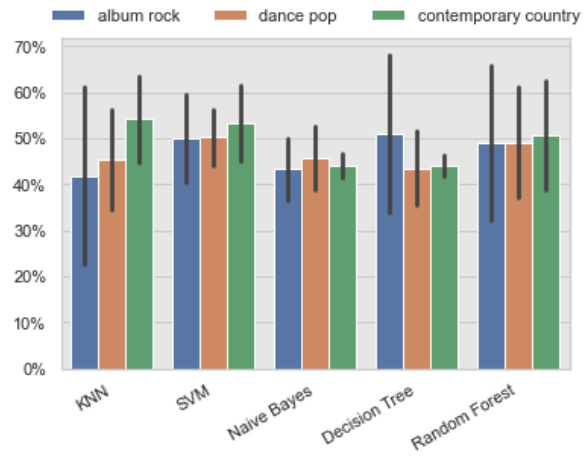

Figure 14. Average score per genre and classifier

	Positive (1)	*Negative (0)*
Positive (1)	97	54
Negative (0)	19	33

Figure 15. SVM (Contemporary country) test confusion matrix

Some models did perform quite good on training dataset as shown in figure but then accuracy scores dropped once presented with testing dataset.

Figure 16. KNN Training and testing accuracy (contemporary country)

To understand if the new genre-based approach is worth pursuing: the metrics have been averaged and compared with the results from the same five classification models on a genre-mixed dataset.

In fact, the best results were obtained using a mixed dataset with 6000 songs which indicate how size played a role in the genre-based approach.

However, the genre-based approach did not suffice and perform slightly worse than a standard dataset with a similar size.

Figure 17. Genre-based against standard dataset with similar size comparison

IV. CONCLUSIONS

A. Conclusions

The results provided were not impressive and the overall scores on standard dataset (up to 60%) were in line with previous literature [7][11].

The genre-based approach resulted in slightly worse popularity prediction accuracy, this suggests that the approach is not worth pursuing when only taking in consideration acoustic features; however, in the future can be tested on different feature datasets such as lyrics and metadata.

A possible reason for the lack of accuracy might be found in the temporal aspect of the dataset: because of size issues the datasets used were generated without differentiating between years, as result modern and old songs were clustered together without reflecting the musical and cultural changes of the public throughout the years.

Other reasons are probably to be found in the features being used, which might not have captured all the different instances in the music: songs might have similar acoustic profiles and yet being a completely different musical experience; the suggestion would be to implement a more in dept dataset around acoustic features with information borrowed from the musicology field and with a higher granularity such as rhythm and shuffle type for example.

Overall, the size of the genre subsets was small (around 1000 rows) which might have implied a worst accuracy in training data than expected, therefore being outperformed by the bigger standard dataset.

In summary, this project has highlighted the problems behind acoustic features in popularity prediction; a genre-based approach seems to not be beneficial when only taking in consideration such features.

However, the position of the author of this project on HSS as field remains positive as literature and recent developments prove how hit-song prediction is an open question worth answering.

B. Future work

According to relevant literature [19] lyrics can outperform acoustic features in song prediction success; applying the genre-based method to a mix of musical and lyrics features might be beneficial for future discoveries.

Another application of a genre-based approach might be to help develop a in-dept genre classification algorithm, such algorithms are already discussed and used in literature [20][21]and the mentioned approach of this paper might help to better grasp accuracy within a certain music genre for better user-recommendations, genre-recognition has already been proven successful in this and other papers [14][15][22].

On the matter of acoustic features: this and other papers have proven how acoustic features can be partially informative regarding popularity prediction, however, for future work: a more comprehensive approach including lyrics, metadata and more granularity on music features must be considered.

Lastly, developing new datasets with smaller granularity and additional qualitative data, maybe borrowed from the musicology field, might bring ulterior insights and discoveries.

ACKNOWLEDGMENTS

This research project was carried out as final award of degree of Higher Diploma in Data Analytics.

The author would like to thank the entire teaching staff at National College of Ireland for the enormous amount of knowledge shared and Chiara Pierini for encouragement and support shown during these months.

REFERENCES

[1] M. Schedl, E. Gómez and J. Urbano (2014), "Music Information Retrieval: Recent Developments and Applications", Foundations and Trends® in Information Retrieval: Vol. 8: No. 2-3, pp 127-261

[2] R. Typke, F. Wiering, R. Veltkamp, J.D. Reiss, G.A Wiggins, 2005, September. “A survey of music information retrieval systems. “ In Proc. 6th international conference on music information retrieval (pp. 153-160). Queen Mary, University of London.D. Byrd, T. Crawford, “Problems of music information retrieval in the real world” in, Information Processing & Management, Volume 38, Issue 2, 2002, Pages 249-272,

[3] T. Li, M. Ogihara, & G. Tzanetakis, (Eds.). (2011). “Music Data Mining” (1st ed.). CRC Press..

[4] F. Pachet, P. Roy “Hit Song Science Is Not Yet a Science.“, Conference: ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008

[5] Y. Ni, R. Santos-Rodriguez, M. Mcvicar, T. De Bie “Hit Song Science Once Again a Science?” 2011

[6] N. Borg and G. Hokkanen. “WHAT MAKES FOR A HIT POP SONG ? WHAT MAKES FOR A POP SONG ?” 2011.

[7] R., Dhanaraj, B. Logan, “Automatic Prediction of Hit Songs”. ISMIR, 2005.

[8] D. Herremans, D. Martens, K. Sörensen., “Dance Hit Song Science”, International Workshop on Music and Machine Learning, 2013

[9] J. Pham, E. Kyauk, E. Park, “Predicting Song Popularity”, Department of Computer Science Stanford University, Stanford Univ., Stanford, CA, USA, Tech. Rep 26 (2016).

[10] C. V. Soares Araujo, M. A. Pinheiro de Cristo and R. Giusti, "Predicting Music Popularity Using Music Charts," *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 859-864

[11] M. Reiman & P. Örnell, (2018). Predicting Hit Songs with Machine Learning (Dissertation)

[12] O. Silva La´ıs, M. Rocha Mirella, M. Moro, 2019, “MusicOSet: An Enhanced Open Dataset for Music Data Mining” Mariana Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

[13] https://marianaossilva.github.io/DSW2019/

[14] N. Scaringella, G. Zoia and D. Mlynek, "Automatic genre classification of music content: a survey," in *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133-141, March 2006

[15] 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

[16] N. Pelchat and C. M. Gelowitz, "Neural Network Music Genre Classification," in *Canadian Journal of Electrical and Computer Engineering*, vol. 43, no. 3, pp. 170-173, Summer 2020

[17] J.S. Downie., “Music information retrieval. Annual review of information science and technology”, 37(1), pp.295-340, 2003.

[18] Lee, Junghyuk and Jong-Seok Lee. “Predicting Music Popularity Patterns based on Musical Complexity and Early Stage Popularity.” *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia* (2015)

[19] *A. Singhi and D. G. Brown, “Can Song Lyrics Predict Hits?” University of Waterloo, Cheriton School of Computer Science*

[20] *R. Basili, A. Serafini, A. Stellato, “Classification of musical genre: a machine learning approach”. In ISMIR. October 2004*

[21] *A. Elbir, H. Bilal Çam, M. Emre Iyican, B. Öztürk and N. Aydin, "Music Genre Classification and Recommendation by Using Machine Learning Techniques,"* 2018 Innovations in Intelligent Systems and Applications *Conference (ASYU), 2018, pp. 1-5*

[22] *N. Pelchat and C. M. Gelowitz, "Neural Network Music Genre Classification," in* Canadian Journal of Electrical and Computer Engineering, *vol. 43, no. 3, pp. 170-173, Summer 2020.*