Three Feature Datasets extracted from Popular Brazilian Hit Songs and Non-Hit Songs from 2014 to 2019

André A. Bertoni[1], Rodrigo P. Lemos[2], André A. S. Coelho[3], Hugo V. L. e Silva[4]
[1 2 3] Universidade Federal de Goiás - UFG, Goiânia, Goiás, Brazil.
[4] Instituto Federal de Goiás - IFG, Anápolis, Brazil
Corresponding author: André A.Bertoni

ABSTRACT
The huge advance in Machine Learning and Deep Learning techniques encourages the confrontation of more challenging problems and produces an increasing demand for data sets that provide information appropriate to the complexity of the analysis. In fact, creating a dataset that is really functional and optimized for use in Artificial Intelligence systems can be as complex a task as its use. Particularly in the case of Hit Song Science, some companies or institutions specialized in the creation of this data, limit or greatly restrict its sharing. Sometimes the data sets include only a few specific artists, styles or contemporaneity, sometimes they bring a very limited set of characteristics or simply do not fit the scope of the research. Most databanks available in the Internet are very generic and compromise only songs in English, offering a limited number of features, and not caring about the release period of the works. The Brazilian music market is the 10th largest in the world and accounted for US$ 338 million in revenue in 2019, of which more than 90% was due to Brazilian songs. So, to provide suitable analysis of this market, this work deals with the creation and optimization of data sets using only Brazilian songs within a limited and contemporary timeframe. A large set of features extracted from a bank of 881 Brazilian popular songs of Success and Non-Success from January 2014 to May 2019. Three feature sets were created: the first with 3215 features; the second used filtering techniques to preserve only the most statistically relevant 74 characteristics among them; and the third bank is totally new, since it was formed from the Vocal Melody of each song (Predominant Melody), and there is no similar set available for research. All features were extracted using the Essentia package. The three feature sets were made available in a repository to help the development of new research in the future.

KEYWORDS: Feature Extraction, Hit Song Science, Music Information Retrieval (MIR), Audio Datasets, Dimensionality Reduction, Python, Pandas, Essentia Package, Predominant Melody, Brazilian Hit Songs.

Date of Submission: 16-02-2021 Date of acceptance: 02-03-2021

I. INTRODUCTION AND BACKGROUND

For several years, we have seen a profound change in the music entertainment industry. The disappearance of traditional formats, such as physical discs (Vinyl, CD and DVD) and the rise of new formats, consumed exclusively through Streaming, ended up forever changing socio-economic-cultural paradigms in the way we listen to music in our lives. Digital platforms like Deezer, Spotify and iTunes now provide thousands of songs in the palm of our hands, through smartphones. In order to continue to keep their subscriber base connected to their platforms for as long as possible, these companies have learned that they must have a deep understanding of their products (music) and their users. From this idea, research on musical features became more important. This practice, commonly known as MIR (Music Information Retrieval), comes down to extracting as much information from the songs by using statistical and semantic analysis algorithms. Thus, for each song, several musical information is generated and stored in a data bank and later used by Streaming companies in statistical analysis, thereby verifying the performances of the songs in order to promote greater engagement in their user base.

Some banks have already been made available to the scientific community, such as [1], which promotes a collection of audio and metadata resources available free of charge for one million contemporary popular international music tracks. Another example is [2], which provides a bank containing around 1000 songs, divided into ten musical genres, presenting approximately sixty audio characteristics for each song. Another widely used bank is [3], which proposes the extraction of about 20 characteristics of 175,000 songs between the years 1921 to 2020, using songs that occupied the best positions in the Billboard North American magazine. As in the first bank [1], these last two are also composed primarily by songs sung in the English language.

In parallel to this, there are researches involving the personal musical tastes and preferences of each individual, called by [4] Hit Song Science. This is another area of study that is also quite new - started in 2005. It also uses the same information for extracting musical characteristics for the analysis of song preferences, trying to predict popularity (commercially successful) or unpopularity (non-commercially successful) of songs.

The lack of datasets containing information on the musical characteristics of Brazilian songs, as well as the huge socio-cultural differences between the Brazilian market and other cultures, were decisive for the proposal of this work. Therefore, this paper deals with the creation of three databases containing characteristics extracted from contemporary Brazilian songs. The main characteristic of these banks is that they are temporally and culturally delimited. The analysis was made, therefore, in songs that were in vogue from January 2014 to May 2019. Altogether there are 441 hit songs and 441 non-hit songs.

The first bank is composed of 3215 distinct features, both qualitative and quantitative - with or without time dependence. As the large dimensionality of the first bank may require a prohibitive processing effort by artificial neural networks, statistical filtering techniques were used to generate a second bank with reduced dimensionality for only the 74 most statistically relevant features. The third bank contains the feature "Predominant Melody", which represents the melody of the main voice of the songs.

This paper describes the procedures for creating these three banks, and is organized as follows: section II describes the construction of the song bank and the formation of the first feature bank; section III presents the statistical filtering procedures to reduce the dimensionality and create the second bank; section IV deals with the extraction of the Predominant Melody and the formation of the third bank; and, finally, section V brings the conclusions of the work.

II. CREATION OF THE SONG BANK

First, we conduct a survey of the best-placed hit songs in an official ranking each year for a period of five years. However, there are several methods that can be used to measure the performance of songs, such as: a) performance on social media such as YouTube, Instagram or Facebook; b) Trending Topics (Twitter Trends is automatically generated by an algorithm that tries to identify the most relevant topics on the internet.); c) collection and distribution of copyright by ECAD (The Central Office of Collection and Distribution (ECAD) is a private Brazilian office responsible for collecting and distributing the copyright of songs to their authors, located in Rio de Janeiro); d) number of songs streamed from the main digital music platforms: Spotify, Deezer, iTunes; e) number of times that these songs are played on Brazilian radio within the analyzed period. The parameter number **of times played on Brazilian radios** was chosen, as this parameter would provide greater reliability in data acquisition.

2.1. Connect MIX

There are companies all over the world that specialize in monitoring the number of times a song is played on the radio. In Brazil, this service is also offered by the company ConnectMIX [5], which offers a homonymous monitoring tool in real time for auditing and managing Broadcasting for Radio and Television stations. The data obtained through ConnectMIX are displayed according to the example shown in Table 1, which are also available in the repository [6].

Table 1. Ranking100+ ConnectMIX - Year 2014

Year	Position	Artist	Title	Musical Genre	Times played on the Radio
2014	1º	Marcos & Belutti	Domingo de Manhã	Brazilian Music	384067
2014	2º	Zezé di C. & Luciano	Flores em Vida	B. Music	273933
2014	3º	Cristiano Araújo	Cê que Sabe	B. Music	246369
2014	4º	Eduardo Costa	Os 10 M. do Amor	B. Music	245669
2014	5º	Jorge & Mateus	Calma	B. Music	239337
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2014	100º	Fred & Gustavo	Tó, Sou Seu	B. Music	59060

A ranking was then created with 600 songs, from January 2014 to May 2019, taking the 100 most played songs on radio stations in Brazil each year. As the ConnectMIX banks do not provide us with very detailed statistical information about the songs, it was decided to classify all songs according to the percentage of times played on radio stations, for the general classification of the songs. This is so that the songs of the last year (2019) would not be so penalized in comparison to the other years in which the observation had already completed 12 months.

In order to bring more reliability, balance and robustness in the formation of the final Dataset of the songs, the "Non-Hit Songs" bank was built under the premise that they should belong to the same set of artists selected in the first class (Hit -Songs). However, they could not be listed in any position in the ConnectMIX Overall Ranking. In addition, the same number of songs was chosen for both banks, that is, if an artist was listed with 5 songs in the hit song bank, then 5 songs by the same artist were selected for the Non-Hit bank. Non-Hit Songs. This enabled the construction of a bank of songs quantitatively balanced between those of Hit Songs and Non-Hit Songs.

With the 600 most played songs on the radio in the years 2014 to 2019, it was necessary to submit them to some filters, in order to eliminate some redundancies. Are they:

1) songs that are repeated in more than a year of analysis;
2) different versions of the same songs (live or studio);
3) songs recorded by more than one artist in the period;
4) non-Brazilian songs, as they are not part of the scope of this study.

2.2. Extraction and Treatment of Features

After the filtering proposed above, there were 882 songs, 441 of which were labeled as Hit Song and 441 as Non-Hit Song.

As the selected songs last between 150 and 240 seconds on average, in which the melody and other characteristics are repeated more than once, it is desirable to restrict the records to short passages to reduce the effort of extracting and generating the features that describe them. This was necessary due to the fact that some analyzes depend primarily on temporal variables, which would cause an imbalance in the number of features analyzed for each song. In order to define the duration of this time interval, he used the author's experience over three decades of experience in the musical business market, but a study was also carried out on the approximate time of the musical structures of the selected songs. Since all songs are designed for radio performance, they usually share a similar musical structure, consisting of: Intro, Verse A / Semi-Chorus, Chorus A, Solo, Verse B / Semi-Chorus, Final Chorus and Final Solo. Figure 1 illustrates this process by showing the durations of the structural portions of ten songs, with the first five being Hit Songs and the next five being Non-Hit Songs.

To better understand the pattern of Pop Songs that normally play on the radio, Figure 1 was constructed. In this chart, the distribution of events or similar musical structures for all songs is made.

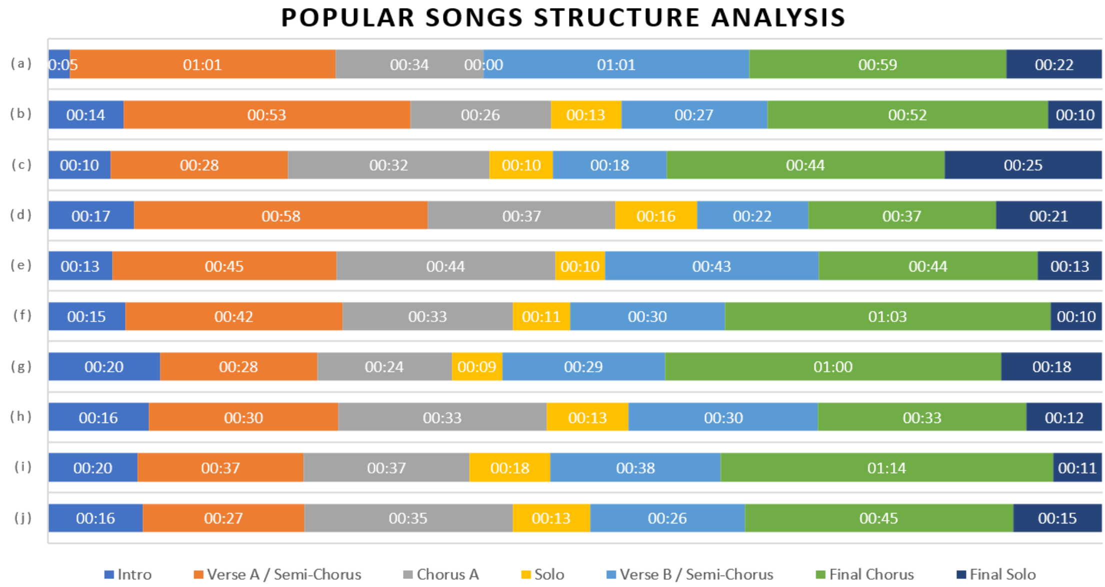


Figure 1. Arrangement and duration of the structural elements of Hit Songs (a, b, c, d, e) and Non-Hit Songs (f, g, h, i, j)

The analysis of the complete set of selected songs revealed that, from the end of the First Chorus, the song returns to the Introduction, Verse / Semi-Chorus A and Chorus. Most of the time, the same arrangement created for the Introduction is repeated on Solo. The Verse B generally has the same melody and metric structure as Verse A. After Verse B, the Chorus is typically repeated twice, thrice or even four times until the song ends. The Final Solo is optional and can either have the same structure as the Beginning or Middle Solos, or even it can be a simple improvisation on the theme of the song's Final Chorus melody, which is also quite common.

Thus, the songs can only be considered technically new until the end of the first chorus. As can be seen in Table 2, the initial 90 seconds bring the main characteristics of a song.

Table 2. Minimum, Maximum and Average - Musical Structure

Duration	**Intro**	**Verse A/Semi-Chorus**	**Chorus A**	**Solo**	**Verse B/Semi-Chorus**	**Final Chorus**	**Final Solo**
Min	00:05	00:27	00:24	00:00	00:18	00:33	00:10
Max	00:20	01:01	00:44	00:18	01:01	01:14	00:25
Average	00:15	00:41	00:34	00:11	00:32	00:51	00:16

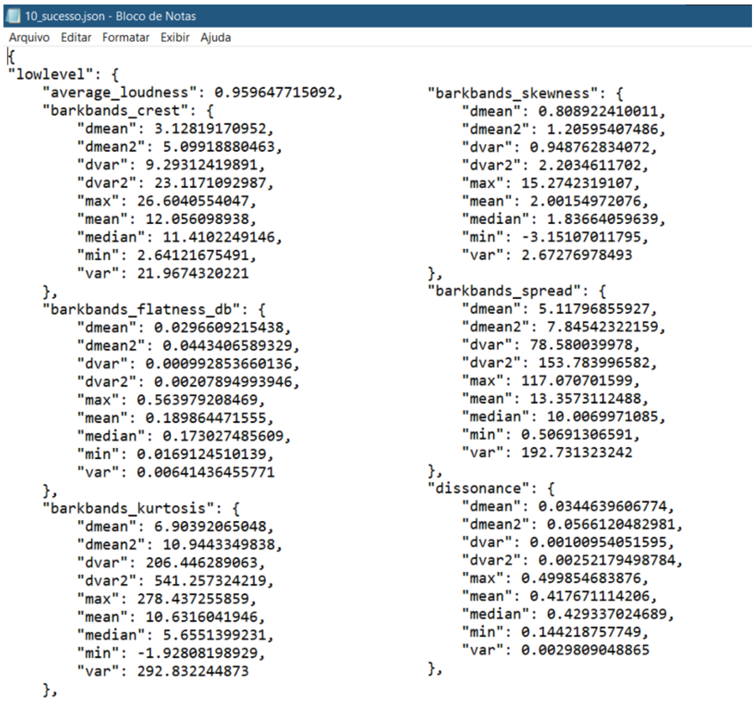
Average Song Duration until the end of the Chorus A	**01:29**

2.3. Brazilian Song Feature Bank

After standardizing the duration of the songs, we performed the feature extraction procedure. For this, the Streaming Extractor Music application from the Essentia package [7] was used. The entire extraction operation was performed on the Linux Operating System (p. 18.04.4 LTS), as the Essentia Package could only be fully installed in a Linux or MacOS environment, according to the developers.

As the extraction application outputs a single "json" extension file (JavaScript Object Notation [8] - JSON is a data exchange extension. It is based on a subset of the JavaScript programming language).

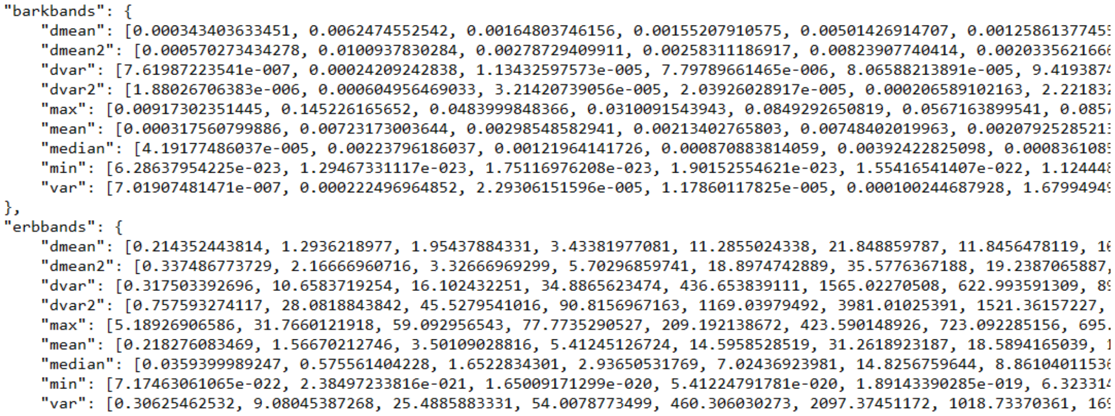
Figure 2 represents a part of the structure of the json file, created from the features extractor, for song number 10 of the hit song bank. It is important to note that this figure represents only a small part of all the features contained in the original file.


10_sucesso.json - Bloco de Notas
Arquivo Editar Formatar Exibir Ajuda

```
{
"lowlevel": {
    "average_loudness": 0.959647715092,
    "barkbands_crest": {
        "dmean": 3.12819170952,
        "dmean2": 5.09918880463,
        "dvar": 9.29312419891,
        "dvar2": 23.1171092987,
        "max": 26.6040554047,
        "mean": 12.056098938,
        "median": 11.4102249146,
        "min": 2.64121675491,
        "var": 21.9674320221
    },
    "barkbands_flatness_db": {
        "dmean": 0.0296609215438,
        "dmean2": 0.0443406589329,
        "dvar": 0.000992853660136,
        "dvar2": 0.00207894993946,
        "max": 0.563979208469,
        "mean": 0.189864471555,
        "median": 0.173027485609,
        "min": 0.0169124510139,
        "var": 0.00641436455771
    },
    "barkbands_kurtosis": {
        "dmean": 6.90392065048,
        "dmean2": 10.9443349838,
        "dvar": 206.446289063,
        "dvar2": 541.257324219,
        "max": 278.437255859,
        "mean": 10.6316041946,
        "median": 5.6551399231,
        "min": -1.92808198929,
        "var": 292.832244873
    },
    "barkbands_skewness": {
        "dmean": 0.808922410011,
        "dmean2": 1.20595407486,
        "dvar": 0.948762834072,
        "dvar2": 2.2034611702,
        "max": 15.2742319107,
        "mean": 2.00154972076,
        "median": 1.83664059639,
        "min": -3.15107011795,
        "var": 2.67276978493
    },
    "barkbands_spread": {
        "dmean": 5.11796855927,
        "dmean2": 7.84542322159,
        "dvar": 78.580039978,
        "dvar2": 153.783996582,
        "max": 117.070701599,
        "mean": 13.3573112488,
        "median": 10.0069971085,
        "min": 0.50691306591,
        "var": 192.731323242
    },
    "dissonance": {
        "dmean": 0.0344639606774,
        "dmean2": 0.0566120482981,
        "dvar": 0.00100954051595,
        "dvar2": 0.00252179498784,
        "max": 0.499854683876,
        "mean": 0.417671114206,
        "median": 0.429337024689,
        "min": 0.144218757749,
        "var": 0.0029809048865
    },
```

Figure 2. Example of an excerpt from the json file generated by the Essentia Feature Extraction tool, in which some features and the respective values of their statistical descriptors are shown.

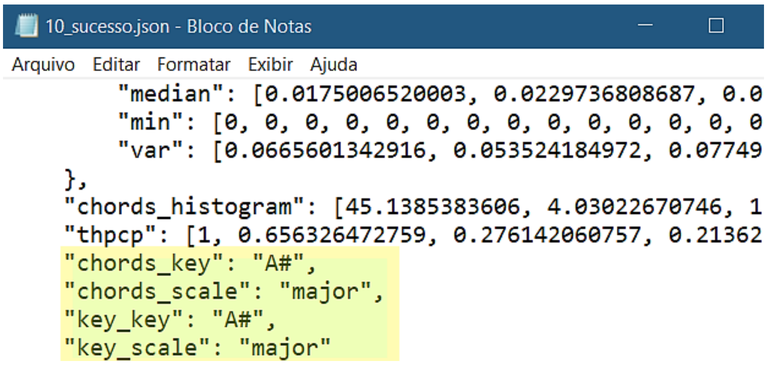
Figure 3 shows some time-dependent features. Here it is used a time windowing technique for extraction, hence the importance of limiting the songs within a standard time interval.



```
"barkbands": {
    "dmean": [0.000343403633451, 0.0062474552542, 0.00164803746156, 0.00155207910575, 0.00501426914707, 0.0012586137745
    "dmean2": [0.000570273434278, 0.0100937830284, 0.00278729409911, 0.00258311186917, 0.00823907740414, 0.002033562166
    "dvar": [7.61987223541e-007, 0.00024209242838, 1.13432597573e-005, 7.79789661465e-006, 8.06588213891e-005, 9.419387
    "dvar2": [1.88026706383e-006, 0.000604956469033, 3.21420739056e-005, 2.03926028917e-005, 0.000206589102163, 2.22183
    "max": [0.00917302351445, 0.145226165652, 0.0483999848366, 0.0310091543943, 0.0849292650819, 0.0567163899541, 0.085
    "mean": [0.000317560799886, 0.00723173003644, 0.00298548582941, 0.00213402765803, 0.00748402019963, 0.0020792528521
    "median": [4.19177486037e-005, 0.00223796186037, 0.00121964141726, 0.000870883814059, 0.00392422825098, 0.000836108
    "min": [6.28637954225e-023, 1.29467331117e-023, 1.75116976208e-023, 1.90152554621e-023, 1.55416541407e-022, 1.12444
    "var": [7.01907481471e-007, 0.000222496964852, 2.29306151596e-005, 1.17860117825e-005, 0.000100244687928, 1.679949
},
"erbbands": {
    "dmean": [0.214352443814, 1.2936218977, 1.95437884331, 3.43381977081, 11.2855024338, 21.848859787, 11.8456478119, 1
    "dmean2": [0.337486773729, 2.16666960716, 3.32666969299, 5.70296859741, 18.8974742889, 35.5776367188, 19.2387065887,
    "dvar": [0.317503392696, 10.6583719254, 16.102432251, 34.8865623474, 436.653839111, 1565.02270508, 622.993591309, 8
    "dvar2": [0.757593274117, 28.0818843842, 45.5279541016, 90.8156967163, 1169.03979492, 3981.01025391, 1521.36157227,
    "max": [5.18926906586, 31.7660121918, 59.092956543, 77.7735290527, 209.192138672, 423.590148926, 723.092285156, 695.
    "mean": [0.218276083469, 1.56670212746, 3.50109028816, 5.41245126724, 14.5958528519, 31.2618923187, 18.5894165039, 1
    "median": [0.0359399989247, 0.575561404228, 1.6522834301, 2.93650531769, 7.02436923981, 14.8256759644, 8.8610401153
    "min": [7.17463061065e-022, 2.38497233816e-021, 1.65009171299e-020, 5.41224791781e-020, 1.89143390285e-019, 6.32331
    "var": [0.30625462532, 9.08045387268, 25.4885883331, 54.0078773499, 460.306030273, 2097.37451172, 1018.73370361, 16
```

Figure 3. Example snippet generated by the feature extractor tool Essentia, which include certain features and values of their respective statistical descriptors extracted by Temporal Windowing.

Finally, the code used should also extract a few categorical features. These features cannot be described numerically and needed to be converted to binary variables during the process of building the final Dataset. One Hot Encoder techniques were used for this task. Figure 4 shows us the final part of the json file, containing categorical features (chords_Key, chords_scale, key_key and key_scale).



10_sucesso.json - Bloco de Notas
Arquivo Editar Formatar Exibir Ajuda

```
        "median": [0.0175006520003, 0.0229736808687, 0.0
        "min": [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
        "var": [0.0665601342916, 0.053524184972, 0.07749
    },
    "chords_histogram": [45.1385383606, 4.03022670746, 1
    "thpcp": [1, 0.656326472759, 0.276142060757, 0.21362
    "chords_key": "A#",
    "chords_scale": "major",
    "key_key": "A#",
    "key_scale": "major"
```

Figure 4. Example of an excerpt from the json file in which the categorical variables and their respective estimated values were highlighted in light yellow, extracted using the Essentia Package Tool.

2.3. Parsing – The Construction of the Dataset:

After extracting all the features, it was necessary to perform a process known as Parsing. This process consists of extracting and reorganizing the data contained in the json files, converting them into a **CSV** extension table. CSV is a format used to store data and can be imported and exported in programs such as Microsoft Excel, Google Sheets, Apple Numbers, OpenOffice Calc and other applications. By definition, CSV is a file format that mean comma-separated-values. This means that all data contained in this type of file is delimited by a comma. The Parsing process is, therefore, very important, since only from this organization of the data will it be possible to properly manipulate the data obtained, enabling the treatment of possible inconsistencies in the Dataset.

2.4. Treatment of the Dataset with Pandas:

After the Parsing step and the creation of the Dataset in CSV format, it is possible to load it in Spyder [6] using the Pandas library (Pandas is a software library created for the Python language, used in Data Science. Offers structures and operations to manipulate numerical tables and time series). Within Pandas it will be possible to start handling the likely inconsistencies of the Dataset. At first, we will perform a visual search, trying to find the most common problems.

Inconsistencies can be: unwanted features, missing data (NaN), null, divergent, duplicates, outliers and categorical variables (before being treated with One Hot Encoder). After all treatments, which are always necessary when building new Datasets, it was finally possible to format a more consistent Dataset, containing the predictive data and the vector that represents the classes (Hit Song or Non-Hit Song).

After the treatments mentioned above, the final Dataset was as follows:

1) 882 Lines (One line for each Song)
2) 3215 Columns (Features)
3) 1 Column (Classes - Hit Song / Non-Hit Song)

III. REDUCED DIMENSIONALITY DATASET

When you have a large number of features in a Dataset, and you want to use it to train Artificial Neural Network algorithms, it is very common to use computational tools that can reduce its dimensionality.

Machine Learning algorithms tend to have significantly improved performance when it is possible to select features that are statistically more efficient. This usually brings more control and reliability to the models, in addition to significantly reducing the computational time for their training and parameter adjustment. In our case, the RFE (Recursive Feature Elimination) technique was used, as it was the one that was most efficient for this application.

The most common methods, according to [9] are:

1) Filter Methods: This method is based on statistical calculations, assigning a score to each feature. Usually, univariate tests are used that consider the independence of a given feature with the target variable.

2) Wrapper Methods: this method selects and prepares several sets of features, evaluating and comparing them. A predictive model is used to evaluate the combination of characteristics, also assigning a score based on the accuracy of the model. The most common algorithm is RFE (Recursive Feature Elimination).

3) Embedded Methods: These methods learn which features best contribute to the accuracy of the model at the time of construction. Example: penalty methods, Lasso, Elastic Net and Ridge Regression algorithms.

IV. DATASET OF PREDOMINANT VOICE MELODY

According to Jason Blume [10], the melody is the main key if making a Hit Song. Therefore, based on this thesis, which is shared by several musical producers and composers all over the world, and also by the proponent of this work who has been a music producer for almost thirty years in the Brazilian music market, it is proposed to build of a feature bank with information about the Predominant Melody of the songs' voices (this feature was extracted using the Predominant Melody tool, which is also part of the Essentia package).

In Figure 5, we have the example of the initial eight seconds of the main melody for the song number “01” of the Hit Song Data Bank. Note that the graphic represents the main melody of the singing voice of the song.

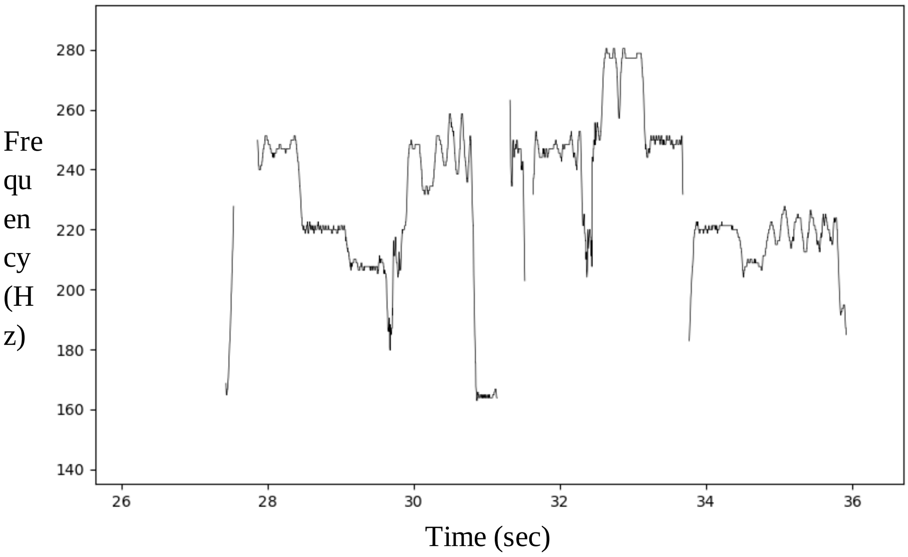


Figure 5. Frequency range of the vocal melody in time - for Song 01 (Hit Song Bank).

The 3rd bank, therefore, contains the feature of the Predominant Voice melody for each of the 882 songs in the study, conditioned in a CSV file, ready for any use in Data Science.

V. CONCLUSIONS

Data Science is a field of Computing that has been growing a lot in recent years. The lack of new databases for studies has always been a challenge for researchers, especially for studies involving Music. The objective of this work is to collaborate with the development and creation of new research and Work Groups that are interested in helping the development of themes related to music. Certainly, the offer of new Datasets would be of great help for the development of new research.

Our proposal differs greatly from other banks available on the internet, as it specifically focuses on contemporary Brazilian songs, selected in a short period of time. In addition, Dataset reflects a vision of the musical preferences of the largest country in Latin America, which is Brazil - a country with more than 200 million people, according to [11]. In addition, the other available Datasets focus mainly on the North American market, containing hundreds of thousands of songs that only cover the English language, that is, they are Datasets with few features extracted from a very large number of songs, limited to a single language.

Another important factor to consider is that our Datasets bring a large number of features (numeric and categorical), being possible to filter them from any needs, allowing a much more comprehensive use.

Unlike other Datasets built with other tools, our banks add more than three thousand and two hundred different features. And this number is much higher than the banks offered in [3] and [12], which contain only nineteen columns of predictors (features), or even [2], which offers 60 features. Our first bank contains all possible features offered by the Essentia package extractor [7]. The second bank, on the other hand, includes the main statistically most relevant features, with 74 features, which were carefully chosen using the RFE feature selection tool. Finally, the third bank offers the features of the predominant melody of the voice for the 881 songs analyzed in the period, which was never proposed in any other Dataset available for free on the internet for studies.

In future works, it is also intended to improve the information contained in this last Dataset of vocal melodies, using filtering techniques so that they can also be used in several other researches, such as those that deal with themes related to music, helping the development of studies on music theory, composition, arrangements or any other related subjects.

REFERENCES

[1]. Bertin-Mahieux, Thierry and Ellis, Daniel PW and Whitman, Brian and Lamere, Paul. The Million Song Dataset. ISMIR, 2011.

[2]. GTZAN. GTZAN Dataset - Music Genre Classification. https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification, Feb. 2021.

[3]. Spotify. Spotify Dataset 1921-2020, 160k+Tracks”, https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks, Feb. 2021.

[4]. Dhanaraj, Ruth and Logan, Beth. Automatic Prediction of Hit Songs. HP Laboratories. Cambridge. HPL-2005-149, Aug. 2005.

[5]. ConnectMIX. Real-time audio monitoring, auditing and management on radios and TVs. https://www.connectmix.com, May 2019.

[6]. Bertoni, André. Three features Datasets of Brazilian Pop Songs (Hit Songs and Non-Hit Songs) from 2014 to 2019. URL= https://github.com/tocaestudio/Music_3DataSets, Jan 2021.

[7]. Dmitry Bogdanov and Nicolas Wack and Emilia Gomez and Sankalp Gulati and Perfecto Herrera and Mayor, O. and Gerard Roma and Justin Salamon and Zapata, J. R. and Xavier Serra. ESSENTIA: An Audio Analysis Library for Music Information Retrieval. International Society for Music Information Retrieval Conference (ISMIR), Nov. 2013.

[8]. json.org. Introduction to json. URL=https://www.json.org/json-pt.html, Aug 2020.

[9]. Dimensionality Reduction of Datasets. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[10]. Jason Blume. What Makes a Song a Hit? BMI. https://www.bmi.com/news/entry/what-makes-a-song-a-hit, Aug. 2019.

[11]. IBGE. Population of Brazil. https://www.ibge.gov.br, Jan. 2021.

[12]. Billboard-Spotify. The Spotify Hit Predictor Dataset (1960-2019). https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset, Aug. 2020.

André A.Bertoni, et. al. "Three Feature Datasets extracted from Popular Brazilian Hit Songs and Non-Hit Songs from 2014 to 2019." *International Journal of Computational Engineering Research (IJCER*), vol. 11, no.1, 2021, pp 01-08.