MEF UNIVERSITY

UNDERLYING THE BIAS FOR HUMAN MUSIC EVALUATION

Capstone Project

Burak YILDIRIM

İSTANBUL, 2018

MEF UNIVERSITY

UNDERLYING THE BIAS FOR HUMAN MUSIC EVALUATION

Capstone Project

Burak YILDIRIM

Advisor: Dr. Tuna ÇAKAR

İSTANBUL, 2018

ACKNOWLEDGEMENTS

I would like to present my deepest thanks to my capstone supervisor Dr. Tuna Çakar for his valuable guidance, motivation and support throughout this study.

I wish to express my sincere gratitude and appreciation to Prof. Dr. Özgür Özlük. You definitely provided the project that I needed to choose the right direction and successfully complete my capstone study.

EXECUTIVE SUMMARY

UNDERLYING THE BIAS FOR HUMAN MUSİC EVALUATION

Burak YILDIRIM

Advisor: Dr. Tuna ÇAKAR

JANUARY 2018, 15

Predictive analysis is the process of using data analytics to predict the future over historical data. Data analytics is the use of statistical modelling and / or machine learning methods to measure the future. In short, it is one of the data mining techniques for predictive analysis that focuses on creating a predictive model for the future by extracting relationships from the data.

Emotional experience and preference, the individual's multimedia content has an important place in the evaluation process. The brain carries emotional cues for emotion detection; but research to extraction these clues is unfortunately limited. Current emotion recognition techniques have been successful in associating emotional changes with EEG signals, and if appropriate stimuli are applied, they can be identified and classified from EEG signals.

In this study, a emotion recognition model with sound analysis data obtained from previous research is proposed. On this analysis data, firstly the dimensions were reduced by using PCA and imbalanced values of liking data were equalizing by using over-sampling model and trained by the classification algorithm.

The main purpose of this project is to extract a model based on a human's music evaluations and sound characteristics by mapping the relations between liking rate and arousal effects. Main method in this research proposes to define each audio by sound analysis and human music liking rates based on the research that was held in DEAP dataset before.

Key Words: DEAP, Emotion Analysis, PCA, Over-Sampling, Classification

ÖZET

İNSANLARIN MÜZİK DEĞERLENDİRME BASAMAKLARININ İLİŞKİLENDİRİLMESİ

Burak YILDIRIM

Tez Danışmanı: Dr. Tuna ÇAKAR

OCAK, 2018, 15

Tahmin analizi, geçmişe ait veri üzerinden geleceğe dair kestirim gerçekleştirmek için veri analitiği kullanma sürecidir. Veri analitiği terimi ise, geleceği ölçmek için istatistiksel modellemenin ve/veya makine öğrenmesi yöntemlerinin kullanılmasını ifade eder. Kısacası, tahmin analizi için, verilerden ilişkiler çıkararak gelecek için bir tahmin modeli oluşturmaya odaklanan veri madenciliği tekniklerinden biridir.

Duygusal deneyim ve tercih, bireyin multimedya içeriği değerlendirme sürecinde önemli bir yere sahiptir. Beyin, duygu tespiti için gerekli duygusal ipuçlarını taşır; ancak bu ipuçlarını çıkarmak için yapılan araştırmalar maalesef kısıtlıdır. Güncel duygu tanılama teknikleri duygusal değişiklikleri EEG sinyalleriyle ilişkilendirmede başarılı olmuştur ve bu nedenle uygun uyaranlar uygulanırsa EEG sinyallerinden tanımlanabilir ve sınıflandırılabilirler.

Bu çalışmada, önceki araştırmadan elde edilen ses analiz veriler ile çalışılan bir duygu tanıma modeli önerilmiştir. Bu analiz verileri üzerinde öncelikle PCA kullanılarak boyutları azaltması yapılmış, eşit sayıda olmayan beğeni verileri Over-Sampling modeli eşitlenerek, sınıflandırma algoritması ile eğitilmiştir.

Bu projenin temel amacı, DEAP veri seti üzerinde çalışarak, dinlenen müziğin işitsel özellikleri ile katılımcılarda oluşturduğu Liking ve Arousal etkileri arasındaki ilişkiyi belirlemektir. Aynı zamanda, her müzik için, daha önce yine DEAP veri seti ile gerçekleştirilen araştırılmalardaki, ses analizleri ile katılımcıların beyan ettiği Liking değerleri arasındaki sınıflandırma modeli ile tahmin etme gerçekleştirmektir.

Anahtar Kelimeler: DEAP, Duygu Analizi, PCA, Over-Sampling, Sınıflandırma

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Overview

Predictive analysis is the process of using data analytics to make prediction based on data for the future. For the Predictive Analytics term, we can say it is a combination of statistical and / or machine learning to quantify the future. Briefly, we can say that it is one of the data mining techniques that focuses on creating a prediction model for the future by extracting relations from the data.

Predictive analytics handles the prediction of future events depends on prior examined historical data by applying machine learning algorithms. The historical data is gathered and converted by using different techniques like filtering, associating the data, and etc. The major goal in data mining is to create and improve the certainty of predictive models, and a basic challenge lies in the discovery of new features, inputs or predictors.(Y.M. Şimşek,2018)

The aim of this research is to determine the relationship between the auditory characteristics of the listening music and the liking and arousal effects on the participants. In this framework, objective sound analysis is planned for each music sample. Then, the liking and arousal values declared by the participants were examined and trained by machine learning classification model.

1.2. Literature Review

Emotion plays a vital role in our daily life as it influences our intelligence, behavior and social communication. Knowledge of human emotions and its effects are crucial for the development in the field of affective computing that integrates emotions into human–computer interaction (HCI) (Picard, 2014). HCI needs to have emotional intelligence similar to human–human interaction. For this, HCI needs information regarding the human emotional experience and relation between emotional experience and the affective expression.(Daimi Syed Naser and Goutam Saha, 2014).

Music has become one of the main means of expressing emotions and communicating. One of the main problems in music studies is subjective because it is related to person's emotional processes. This subject may be restrictive when evaluating music. However, it is also possible to talk about the characteristics of the music that can be objectively assessed. More precisely, it is possible to find some conclusions about the anticipated music based on the acoustic characteristics of the music being played. One of these possible consequences is to be able to determine the psychological elevations of the listeners, starting with the acoustic properties of the listening music. (Tuna Çakar & Mevlut Serdar, 2017)

The aim of this study is to estimate the meaning relationship between brain stimuli and the subjective assessment performed by the individual on any phenomenon.

Experiments in practice will be performed via music phenomenon. According to many studies, music plays an important role in the relationship between emotional experience and emotional expression.

A common cause of listening to music is that your music is an effective tool to convey and remind the feelings. These emotions may be subjective, partly based on the cultural and musical background of the listener, but there are common points of emotionally perceived commonality among different listeners based on the characteristics of the music. Various works have been done to estimate the emotion while listening to music. Some studies have explored the relationship between the physiological activity of a listener and the perceived emotion.[1-2] Others have explored the relationship between perceived emotion and music / acoustics.[3-4] Although we accept that individual differences belong to the emotion of each piece of music, we believe that the modal evaluation is legitimate and that this assessment can be predicted from the characteristics of the music.(Naresh N. Vempala and Frank A. Russo, 2012)

Interfacing directly with the human brain is made possible through the use of sensors that can monitor some of the physical processes that occur within the brain that correspond with certain forms of thought. Researchers have used these technologies to build Brain–Computer Interfaces (BCIs), communication systems that do not depend on the brain's normal output pathways of peripheral nerves and muscles (Calvo & D'Mello, 2010).

In many studies in the literature, it is essential to extraction the features that brain activity represents.

There are few other studies concerned with non-stationary nature of EEG and use of time and frequency information as features. (Murugappan et al. 2008). Classified four emotions (disgust, happiness, surprise and fear) from 64-channel EEG signal recorded for six participants, where movie clips are used as stimuli.(Daimi Syed Naser and Goutam Saha, 2014)

2. ABOUT THE DATA

2.1 General Description of Data Set

This data set, taken from a research centre study of human emotions, includes Electroencephalography (EEG) and peripheral physiological signals of 32 participants. EEG is the process of recording the electrical activities of the brain and is the process of printing on digital media or on paper in accordance with internationally accepted mapping in certain standards. The brain is in constant electrical activity, and in certain periods of human life, this electroactive activity exhibits marked levels of development, and in certain phases of daily life (such as sleep and wakefulness) it maintains its electrical activity in certain standards.

It is aimed to evaluate each participant in terms of their level of liking / disliking and familiarity by watching certain sections of music videos of 40 minutes. In addition, 22 out of 32 participants were recorded with video.

A method that is often used in the literature for stimulating detection was chosen. An algorithm and an on-line evaluation tool have been developed to define the content that comes up in the video by compiling the collected last tags.

In order to use this data set which is open to the public, the user contract is signed first and the user name and password are used to download the data.

The DEAP dataset consists of two parts;

1. 120 videos and a minute of video were evaluated by 14-16 volunteers on the basis of arousal, valence and dominance.
2. 32 participants watched each of these 40 minute video footage, provided each participant watched 40 videos, and participant counts, physiological records, EEG signals of these participants were recorded. Of these participants, 22 faces were recorded as video.

The priority of this study is to examine the relationship between participants' valence and arousal levels calculated via MATLAB.

Findings obtained from studies related to the subject indicate that the level of arousal can be predicted. In other words, sound waves are transformed into different properties that are linked to subjective measures. An example of the sound wave obtained from a 48-second music piece is given below. (Tuna Çakar, Mevlüt Serdar, 2017)

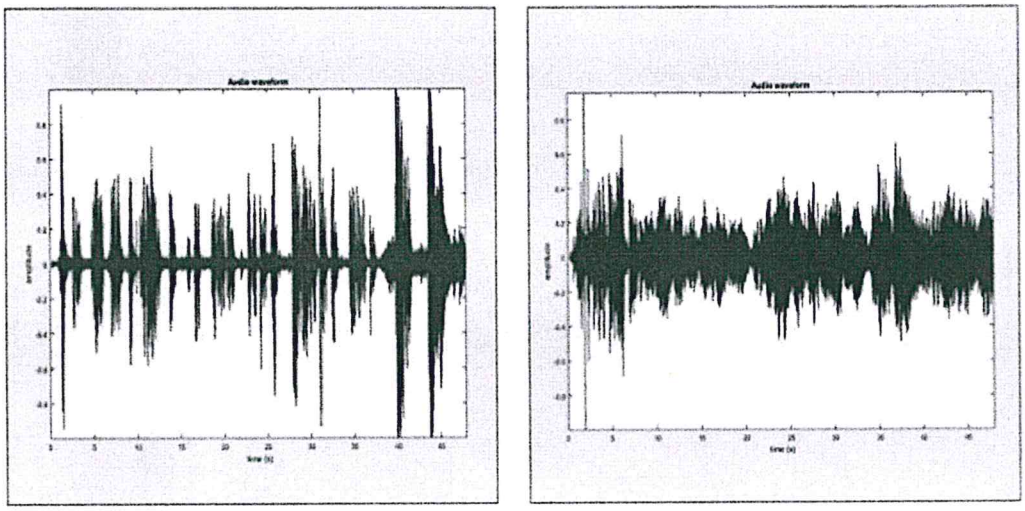


Figure 1: Sound Wave

From the given dataset columns that detailed below:

i. Liking : User likes score

ii. x1 – x160 : Sound Waves

2.2 Data Pre-Processing

In this phase, pandas library was used to make more effective use of python. First, 'read_csv()' function in the pandas library was used to read the data, then all the blank lines in the dataset were dropped against the possibility of any blank lines in the data. The Zscore values for Arousal, Valance and Liking values were calculated and these values were converted to numpy data type using NumPy, another python library, to use as Array. The codes written are as follows.

```
music_eeg = pandas.read_csv("music_eeg.csv")
music_eeg = music_eeg.dropna()


x1 = numpy.array(scale(music_eeg['Arousal']))
x2 = numpy.array(scale(music_eeg['Valence']))
x3 = numpy.array(scale(music_eeg['Liking']))
```

Figure 2: Data Reading & Scaling

For Arousal, Liking and Valence values, values such as averages, standard deviations, min and max values are shown as follows. In addition, quartile values are given.

	Arousal	Liking	Valence
count	1.160000e+03	1.160000e+03	1.160000e+03
mean	2.143879e-16	4.287758e-17	1.362894e-16
std	1.000431e+00	1.000431e+00	1.000431e+00
min	-2.170778e+00	-2.056043e+00	-2.006012e+00
25%	-7.012712e-01	-6.922299e-01	-6.996991e-01
50%	1.465210e-01	2.416237e-01	-7.815663e-02
75%	8.864123e-01	7.102566e-01	8.762272e-01
max	1.939730e+00	1.583825e+00	1.811523e+00

Figure 3: Data Summary

In the dataset, data has approximately 1.1K rows and 161 columns.

```
Int64Index: 1160 entries, 0 to 1279
Columns: 161 entries, x1 to Liking
dtypes: float64(161)
memory usage: 1.4 MB
```

Figure 4: Data Inspection

The inspection output tells:

i. It's an instance of a Data Frame.
ii. Each row was assign an index of 0 to 1279
iii. 1160 rows.
iv. Our dataset has 161 columns.
v. An Approximate amount of RAM used to hold the Data Frame: 1.4mb

Before going through the analysis, the continuous variable was need to be converted Liking feature as categorical variable, so that the classification algorithm could be used. There are a few different steps / methods that could be used for this. However, the method was chosen to find out how many levels of Liking value should be sorted / separated. For this, first it was need to to be found the Euclidean distances for the Liking and Arousal features using the KMeans clustering algorithm, and the elbow method used to determine how many clusters are used for KMeans.

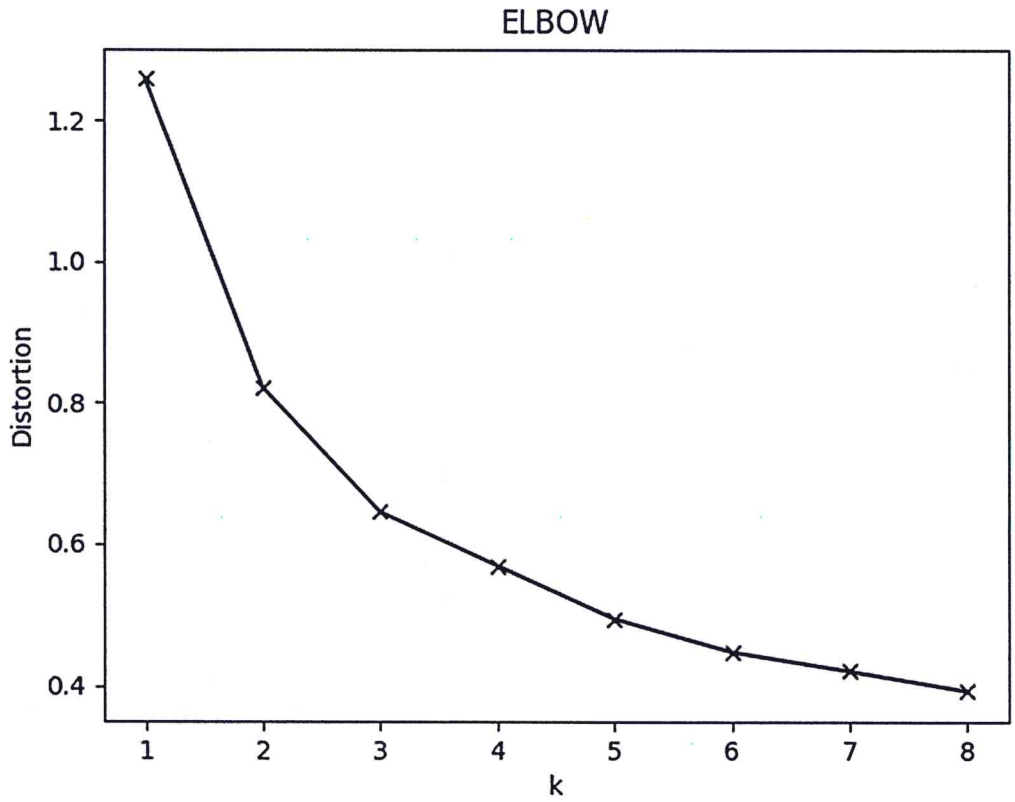


Figure 5: Elbow Method

Above, as can be seen in the graph in Figure 4, there are elbows clearly at points where the number k is 2 and 3. In this case, the graph shows that the number of clusters which could be chosen is 2 or 3.

```
y = pandas.cut(y, bins=3, labels=[0, 1, 2])
X = music_eeg.iloc[:, 14:174]
```

Figure 6: Data Categorization

As shown in Figure 5, Liking values are divided into 3 categories and named 0, 1, and 2.

3. PROJECT DEFINITION

3.1 Problem Statement

In the analysis, 12 audio attributes were calculated and expressed as numerical values for each of the 48 audio tracks, which were the technical outputs obtained using the

MATLAB and MIR library. These features; rms value, low energy evendentity, tempo, intelligence, zeroacity, center, spread, rolloff, brightness, irregularity and mod.

3.2 Problem Objectives

The main purpose of this study is to show whether there is any relationship between objective and subjective measures of the music we listen in our life. In addition, the 9-point Likert scale was used in this study. For each music video from 32 participants, the individual scores on the valence and Arousal dominance axes were collected. Then the average valence arousal and Dominance values for each video were calculated.

4. METHODOLOGY

4.1 Exploratory Data Analysis (EDA)

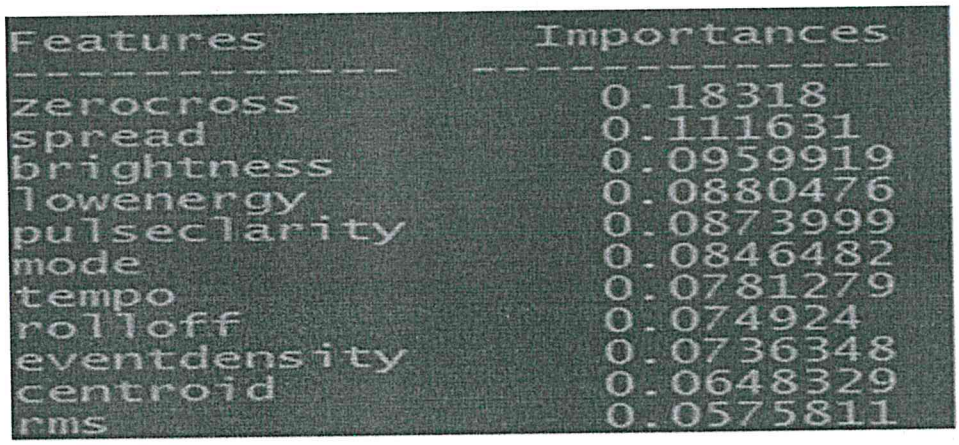
An approach that uses data analysis to maximize insight into a data set, finds important variables, determine uncertainty and anomalies. The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

i. Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.
ii. Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
iii. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

4.2 Descriptive Statistics

Above it has been mentioned 12 audio attributes for each of the 48 audio tracks. During EDA, it was tried to find the numerical values of the information/importance they contain in these 12 features. The values importance of these features was calculated by using a random forest algorithm which has a machine learning algorithm and finds feature importance of the features when applying this algorithm. And feature importance values respectively;



Features	Importances
zerocross	0.18318
spread	0.111631
brightness	0.0959919
lowenergy	0.0880476
pulseclarity	0.0873999
mode	0.0846482
tempo	0.0781279
rolloff	0.074924
eventdensity	0.0736348
centroid	0.0648329
rms	0.0575811

Figure 7: Feature Importance

As it is shown in Figure 7, it could be said that the zerocross feature is the most important variable. And the distribution of mean values for 40 music videos can be seen in the following graphs;

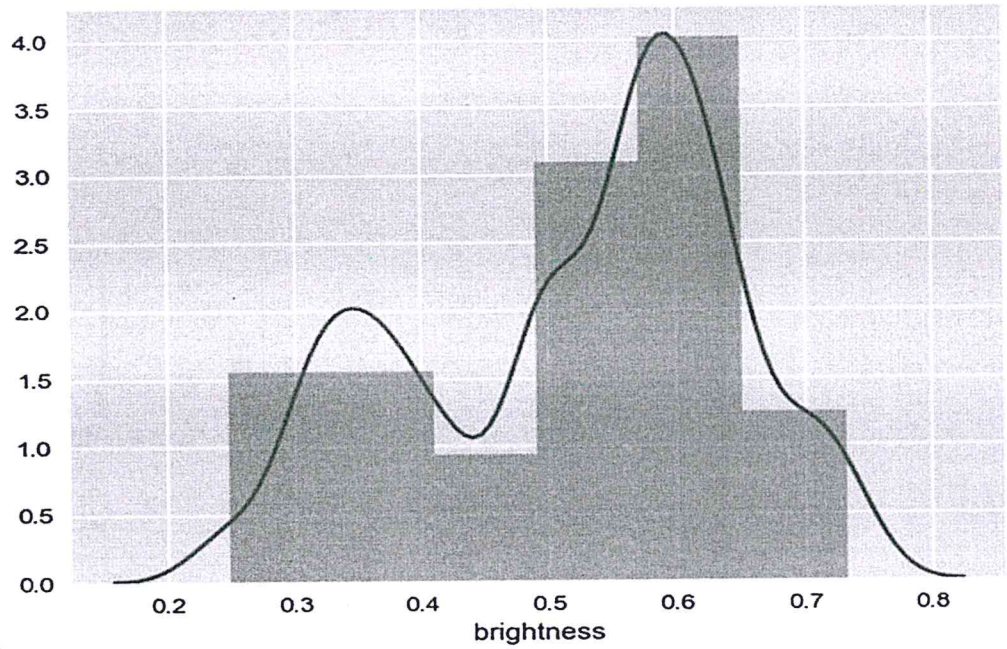


Figure 8: Frequency histogram of brightness for 40 videos

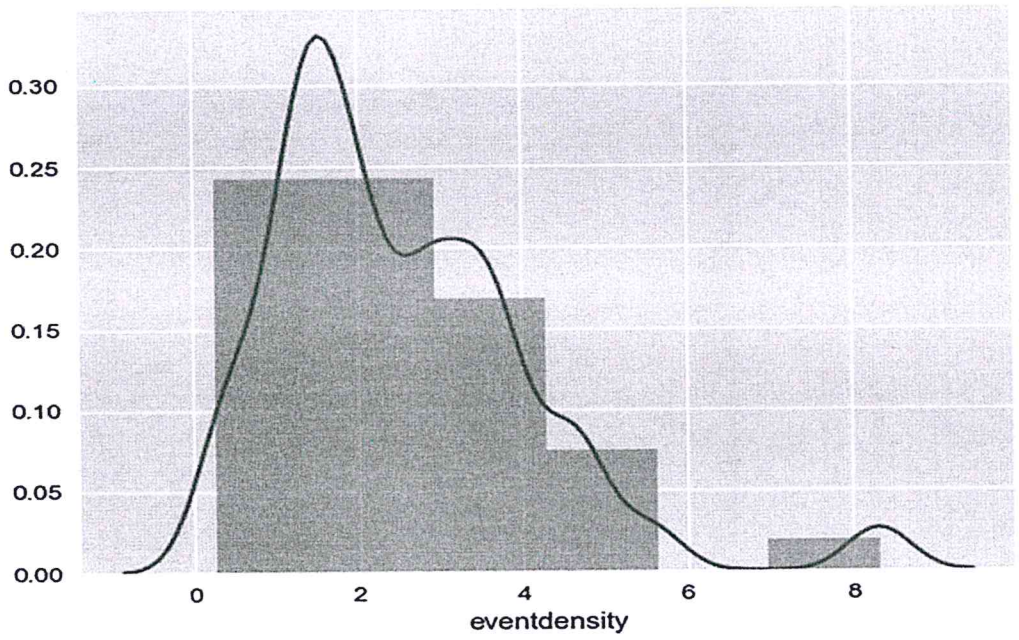


Figure 9: Frequency histogram of eventdensity for 40 videos

And This table shows the correlation coefficient between measures;

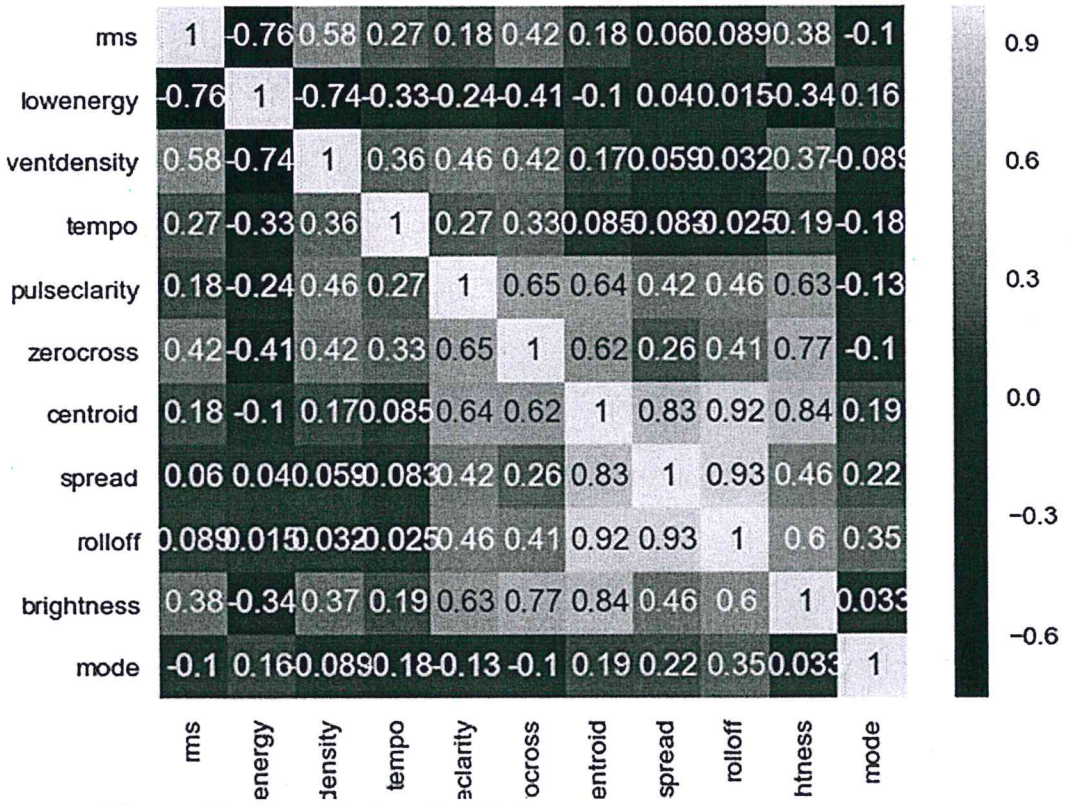


Figure 10: Correlation Coefficient

4.3 Imbalanced Data Set

If you have spent some time in machine learning and data science, you would have definitely come across imbalanced class distribution. This is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes.

This problem is predominant in scenarios where anomaly detection is crucial like electricity pilferage, fraudulent transactions in banks, identification of rare diseases, etc. In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate.

In our work, we have had to work with such a set because the groups of Liking features that we have transformed into categorical units during pre-process are becoming imbalanced. Even though we obtained high accuracy during our initial analyzes, it was because we were imbalanced to make a correct prediction even if we predicted it without learning. We realized that the data should be equalized because of this situation. Many models were available.

4.3.1 Challenges with standard Machine learning techniques

The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets.

Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.

Evaluation of a classification algorithm performance is measured by the Confusion Matrix which contains information about the actual and the predicted class.

Actual	Predicted	
	Positive Class	Negative Class
Positive Class	True Positive(TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Figure 11: Confusion Matrix

Accuracy of a model = (TP+TN) / (TP+FN+FP+TP)

4.3.1.1 Random Under-Sampling

Random Sub-Sampling aims to balance the class distribution by removing randomly the class instances that are more than the others. This is done until the majority and minority class samples are equilibrated.

4.3.1.2 Random Over-Sampling

Over-Sampling increases the number of samples in the minority class by duplicating them to provide a higher representation of the class than the others in the sample.

4.4 Using Over-Sampling

We used the oversampling method. In fact, the most important reason for us to use it was that we did not reduce our already small data count further, so we decided to use the oversampling method.

In our project, we used the RandomOverSampler function of the "iMBLEARN" library, which is a python library. Value counts after Random Over-Sampling;

```
# Liking Value Counts After using OverSampling Method
2     540
1     540
0     540
dtype: int64
```

Figure 12: Liking Value Counts

4.5. Principal Component Analysis

Principal Component Analysis (PCA) is used to reduce the size of the dataset during data analysis. For example, by converting a dataset with n features to a k (n) dimensional dataset, some operations can be completed faster (eg training of the classification algorithm). Of course, some of the features of the data will be lost during these operations, but the main goal here is to work at a minimum loss while keeping the variance high.

PCA was made to reduce the number of features that were 160. First, it was need to be decided how many dimensions to yield, and here it was tried to find out how many variable could explain the cumulative sum of variance explanatory values.

```
[0.7311178  0.84601332 0.89891268 0.93107392 0.95098693 0.96412819
 0.97356609 0.97939367 0.98442022 0.98868014 0.99194846 0.99377978
 0.99509606 0.99600674 0.99686122 0.99761968 0.9981461  0.9985352
 0.99884674 0.99906293 0.99925845 0.99938963 0.99950032 0.99958365
 0.99965523 0.99971504 0.99976884 0.99981626 0.99985351 0.99988367
 0.99990653 0.99992666 0.99994506 0.99996217 0.99997322 0.99998121
 0.99998544 0.99998812 0.99999013 0.99999209 0.99999369 0.99999508
 0.9999961  0.99999704 0.99999775 0.99999822 0.99999864 0.99999891
 0.99999911 0.99999928 0.99999942 0.99999953 0.99999962 0.99999969
 0.99999976 0.99999981 0.99999985 0.99999988 0.99999992 0.99999994
 0.99999996 0.99999998 0.99999999 1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.        ]
```

Figure 13: Explained variance ratio

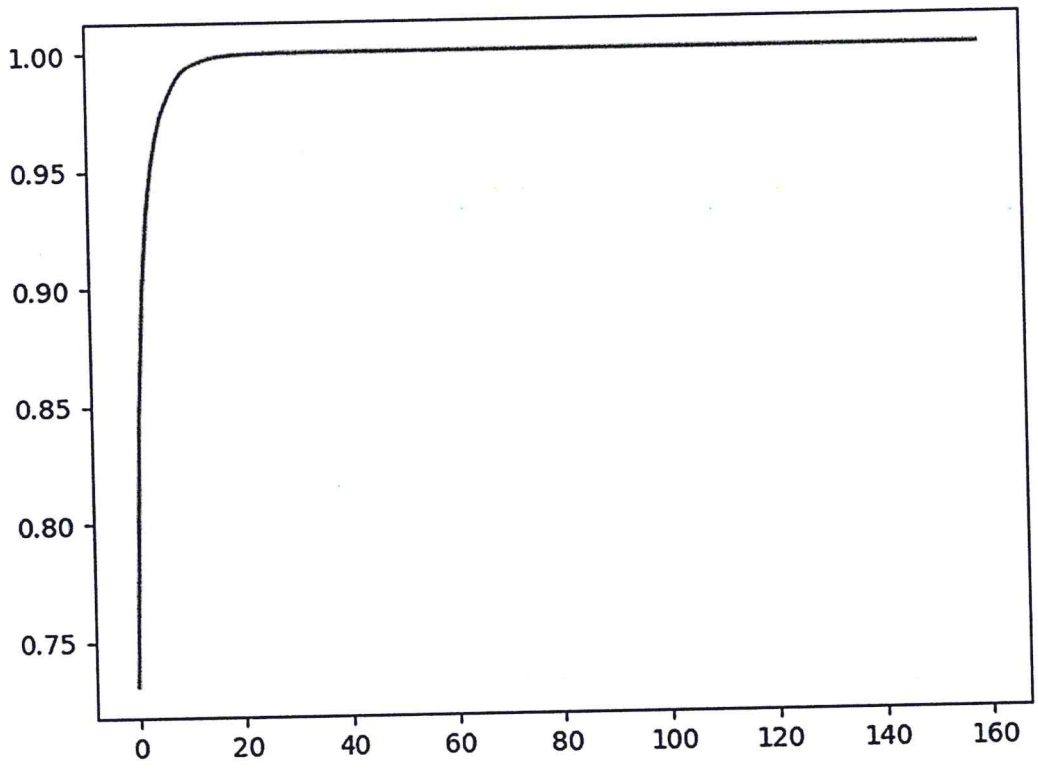


Figure 14: PCA Graph

As shown in Figure 14 above, the explanatory ratios of the variables are given. The study was continued with 18 variables explaining the rate of 9985352%.

4.6 Splitting Data Two part. Train & Test

A common way to get data for classifiers is to split the available data into two sets, a training set and a test set. The classification model is built on the training set and is applied to the test set. The test set has never been seen by the model so the resulting performance will be a good guide to what will be seen when the model is applied to unseen data.

Very often, the proportion chosen is 70% for the training set and 30% for the test. The idea is that more training data is a good thing because it makes the classification model better whilst more test data makes the error estimate more accurate.

In this study, the data was splitted in two parts %70 data for training and %30 data for test and the training was prepared for the Classification algorithm.

5. RESULTS

The decision tree learning method is one of machine learning topics. There are applications in the literature such as a classification tree or a regression tree that can be considered as sub-methods of decision tree learning.

In decision tree learning, a tree structure is created, class labels at the level of the leaves of the tree, and the processes on the properties with the leaves that go to these leaves and from the beginning.

Decision tree learning algorithms;

Some of these algorithms can be used during the execution of these trees

- Random Forest: During the classification process, it is aimed to increase the classification value by using more than one decision tree.
- Boosted Trees: An algorithm that can be used for both classification and regression problems.
- Rotation Forest: Randomly uses more than one tree in a similar manner, but each tree is first trained using Pricipal Component Analysis (PCA). For this training a randomly selected subset of the data set is used (by cistern method).

In addition, the following algorithms are used in decision tree learning:

- ID3 algorithm
- C4.5 algorithm

5.1 Decision Tree Algorithm

Decision Tree algorithm is used in model. Our model was trained using 70% of the data we splited and tested with the data we reserved for testing. The result of this process is as follows.

```
Decision Tree Predictions Score:  0.7333333333333333
```

Figure 15: Decision Tree Score

As you can see in the chart above, our model's score is around 73%.

5.2. Random Forest Algorithm

Unlike the decision tree algorithm, there are many classification trees in the random forest algorithm. To classify new input, each tree gives a class and The forest chooses the classification having the most assigned class.

As you can see below, the prediction score of the randomforest algorithm is :

```
RandomForest Predictions Score:
```

Figure 16: Random Forest Score

5.3 Choosing The Algorithm

When choosing between different models, selection is usually made according to the prediction scores. For this reason, among the algorithms tested in this study, a model generated by Random Forest algorithm with a higher prediction score was chosen. The reasoning behind this selection is that the Random Forest algorithm is consisted of multiple decision trees, hence usually yielding better results compared to the decision tree algorithm.

REFERENCES

1. Naresh N. Vempala and Frank A. Russo. (2012). Predicting Emotion from Music Audio Features Using Neural Networks. 9th International Symposium on Computer Music Modelling and Retrieval.

2. Tuna Cakar and Mevlut Serdar (2017). Evaluation of Music Excerpts via Subjective and Objective Measures. Arge Dergisi

3. Picard, R. Affective computing. Cambridge, MA: MIT Press.

4. Daimi Syed Naser and Goutam Saha, Classification of emotions induced by music videos and correlation with participants' rating, Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, Kharagpur 721

5. https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

6. http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

7. https://shiring.github.io/machine_learning/2017/04/02/unbalanced

8. https://www.jair.org/media/953/live-953-2037-jair.pdf

9. https://link.springer.com/chapter/10.1007%2F11538059_91?LI=true

10. http://information-gain.blogspot.com.tr/2012/07/why-split-data-in-ratio-7030.html

11. http://bilgisayarkavramlari.sadievrenseker.com/2012/04/11/karar-agaci-ogrenmesi-decision-tree-learning/