RAVDESS Analysis

Data Mining Project

Roberto Caldari, 649249
Sandro Cantasano Martino,660974
Alessandro Mastrorilli,657939

Contents

1 Introduction

"The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English" contains video and music files about a neurological scientific study performed in North America, starting from the assumption that facial emotions can help to understand neurological disorders and possible rehabilitation treatments.

Briefly, twenty-four different actors make an experiment where they have to pronounce two different statements, varying the emotion while speaking or singing the statement previously mentioned. The goal of the analysis is to gather information useful to the cause.

2 Data Understanding

The dataset contains 2456 rows and 38 columns: the rows include the audiovisual clips in the different modes performed in the experiment, between the columns are inserted the attributes that are displayed in the table below, differentiated so that these are numerical or categorical:

Numerical Attribute Name	min Value	Max Value	Explaination
Frame rate	48000	48000	Frequency per frame
Intensity	-63864	-16354	Loudness in dBFS
Zero crossing sum	4721	30153	Times the audio function reach zero
Mean	-0.00094	0.0012	Mean value of audio function
Std	0.00065	1.1521	Standard deviation of audio function
Min	.0.998	-0.0061	Min value of the audio function
Max	0.0047	0.999	Max value of the audio function
Kur	1.758	59.086	Kurtosi index
Skew	-2.356	1.800	Simmetry index
Mfcc mean	43.83	-15.49	Middle value of MFC
Mfcc std	83.62	195.94	Standard deviation of MFCC
Mfcc min	-1085.48	-461.49	Min value of MFC
Mfcc max	126.25	280.17	Max value of MFC
Sc mean	2360.88	7655.33	Middle value of spectral centroid
Sc std	1491.34	4819.78	Standard deviation of spectral centroid
Sc min	0	2121.42	Min value of spectral centroid
Sc max	11516.03	17477.54	Max value of spectral centroid
Sc kur	-1.80	3.66	Kurtosi index of spectral centroid
Sc skew	-0.51	1.83	Simmetry index of spectral centroid
Stft mean	0.21	0.72	Middle value of STFT
Stft std	0.21	0.39	Standard deviation of the STFT
Stft min	0.000	0.039	Min value of the STFT
Stft max	1.0	1.0	Max value of the STFT
Stft kur	-1.700	0.795	Kurtosi index of the STFT
Stft skew	-0.993	1.466	Simmetry index of STFT

Figure 1: Numerical attribute

Thirty-one variables are numeric, the first seven are summarizable in variables related to general audio technical notions because *the audio is understood analytically as a sound wave that for appropriate technical transformations receives a sectioning first in* ***sound samples*** *that are divided into frames that is small parts of the sample itself*; the next six variables are related to pure statistical variables that refer to the original audio, after that the last variables are instead mathematical sound analysis variables like the Short-time Fourier Transform, Mel-Frequency Cepstrum Coefficient and spectral centroid.
The **Short-time Fourier Transform** refers to the intuition that a function can be compared to a succession of sine and cosine, so the sound wave in question can be rewritten as a sine wave frequency. In sound processing, the *mel-frequency Cepstrum (MFC)* is the representation of the short-term power spectrum of a sound, as according to Fourier analysis any signal can be broken down into a number of discrete frequencies. **MFCC** is the coefficient in relation to this representation. In other words, with the Fourier transformation we get a sinusoidal function, with the MFC we get a power spectrum that informs us about the frequencies that this function has and at what time cadence. **The spectral centroid** is a measure used in the process of analyzing digital signals that allow to characterize the spectrum, it has a very intense connection with the value of sound clarity. It's calculated as the weighted average of the frequencies in the signal, using the Fourier transformation.

In the following part(Figure 2) we summarize the categorical attributes, with their domain and explanation:

Categorical Attribute Name	Domain	Explaination
Actor	{1;24}	Identification number
Repetition	{1st repetion , 2th repetition}	Number of repetition of the experiment
Modality	{Audio Only}	Format of the file
Vocal Channel	{Speech , Song}	Modality of the performance
Emotional Intensity	{Normal,Strong}	Different intensity of the emotion
Emotion	{Neutral , Calm , Happy , Angry , Fearful , Disgust , Surprised}	Different kind of emotions
Statement	{Dogs are sitting by the door ; Kids are talking by the door}	Different statements used in the performance
Sex	{Male , Female}	Actors'gender
Channel	{1-mono , 2- stereo}	Mode of listening
Sample width	{1-8 bit , 2-16 bit}	Number of bit for sampling
Frame width	{2,4}	Number of bytes for each frame. One frame contains a sample for each channel.

Figure 2: Categorical Attribute

2.0.1 Constant and missing value

Analysis of the dataset showed that several variables have constant values such as modality, which refers just to "audio only". The same category includes sample width, frame rate, and stft_max(frame rate has just a value in the number of 48000, sample width with a value equal to 2, and stft max with a value equal to 1). In addition, the analysis of RAVDESS shows that three variables contained missing values within them, these are vocal channel, actor and intensity respectively with 196, 1126 and 816 missing values.

2.0.2 Correlation matrix

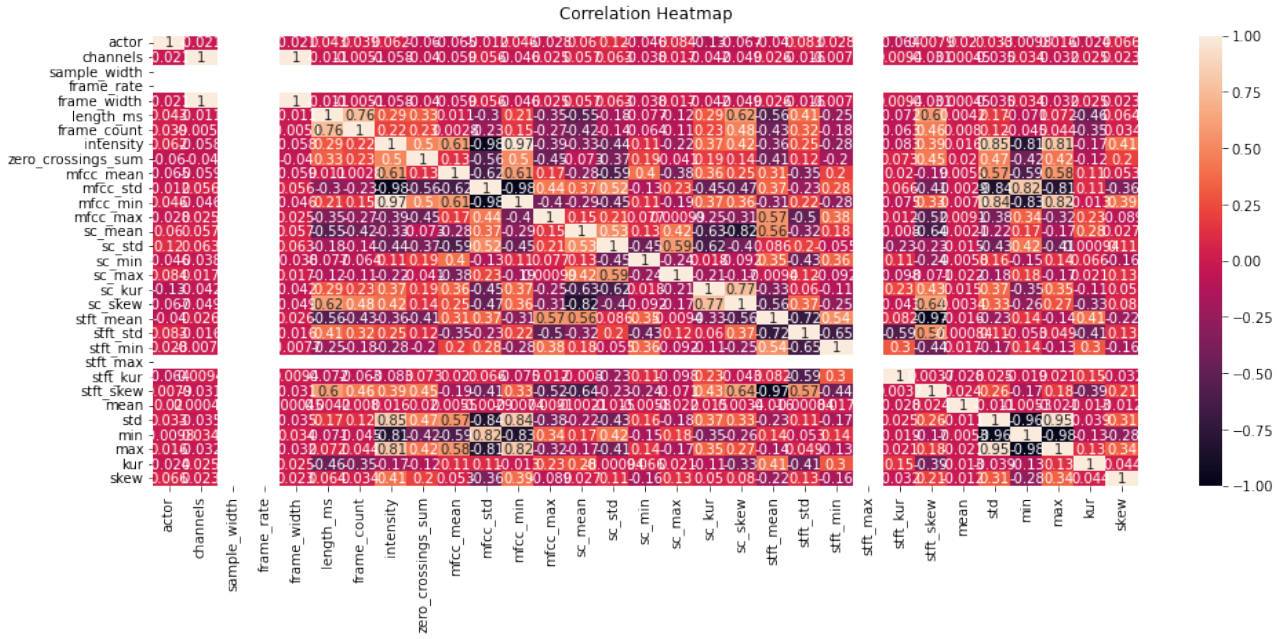

Figure 3: Correlation Matrix

As we may deduce from (Figure 3), high levels of correlation have emerged for the variables "lenght_ms" and "vocal_channel": the length of the audio file appears to have components close to the way of the experiment, both in the form of speech and singing, thus suggesting a relationship of absolute linearity between the two variables. Important levels of positive correlation have been deduced from variables such as the "intensity" and the "min" value of MFCC(0.97) as well as between "max" and the "standard deviation"(0.95). A positive correlation that is considered interesting is between "frame_count" and the "length_ms" of the audio files with a value that amounts to 0.76. The remaining positive correlations were found to relate in large numbers to the comparison between statistical variables of MFCC(such as "standard deviation", and "min" of the MFCC value) with statistical variables such as "min", "max", standard deviation all in a range from 0.822 to 0.844.

An interesting positive correlation also emerged between the comparison between "intensity" and statistical variables such as standard deviation and "max", respectively with values of 0.808 and 0.845.

Moving on to the side of negative correlations, there are high correlations, especially with the "intensity" values that are compared with MFCC associated the standard deviation value , calculated in the value of -0.98. The MFCC values associated with standard deviation and "min" find also important negative correlation values with statistical variables such as standard deviation, "max" and "min" with values ranging from 0.82 and 0.84. In the analysis of the correlation matrix, analysts find remarkable variables that, when correlated, find "nan values". The conclusion is that the variables in question, summarized in the variables sample width, frame rate and STFT "max" don't have a correlation compared to each other because they assume constant values as seen in the previous paragraph.

2.0.3 Categorical attribute's analysis through histogram

After the analysis of the correlation matrix, we have represented in some graphs the distribution of the categorical attributes.

At first, we verified the number of "components" of each variable within our dataset, this verification was carried out both numerically and visually thanks to the use of some graphs. Example: We have counted and represented the number of men (M) and women (F) that can be found within our dataset. This process, as already mentioned, was then developed for each variable. The second part related to the study of the representations of variables through graphs was focused on the study of categorical variables. The subdivision of every single variable in this way has allowed us to verify every single behavior relative to the composition of 2 variables inside the dataset. An interesting observation is made by focusing on the emotion chart shown in (figure 4); we can see that the distribution of emotions is balanced for the emotions of fearful, happy, calm, sad and angry. However, the remaining emotion(surprise, neutral and disgust) have a much lower frequency than the previous one mentioned(but they are still balanced).

Figure 4: Emotion Histogram

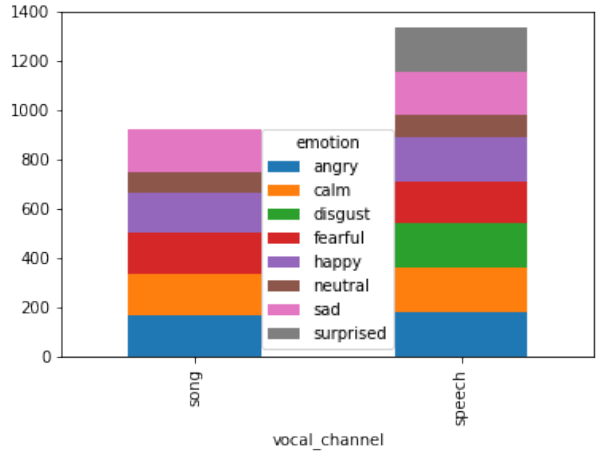

Figure 5: Emotion compared to vocal channel histogram

In the second graph (Figure 5) we can find how the emotions are not frequent in the equal distribution in the song, concerning speech. The two graphs are therefore connected: from the first we assume a different distribution of emotions, and in the second graph, we notice the lack of the emotions such as "surprised" and "disgust" with the respect to the song channel(instead, in the speech channel we found all emotions).

From (Figure 5) we also found a difference between speech and song, speech measures a number of recordings equal to 1335 while the song to 921, a substantially different value.

Focusing on the variable of sex, it has been found that generally, women make a performance in speech, instead, a great number of man performs a song in the experiment.

Interesting data can be deduced from the index represented for each actor: the actors with odd index are men while those with even index are women. Refers to the other categorical variables("sex", "repetition", "statement" and "actor") these are all balanced one to another, except for "channel" and "frame_width"(2446 records concentrated in one class, both).

2.0.4 Continues attribute's analysis through box plots

The analysis of our dataset is then continued thanks to the representation of the data through the box plots. The most interesting representation was given by the box plot of the "frame count" in (figure 6). Thanks to the representation of the box plot cited and a more detailed count it was possible to verify that inside it there are 35 wrong records(these outliers are considered errors since the frame count can't be negative).

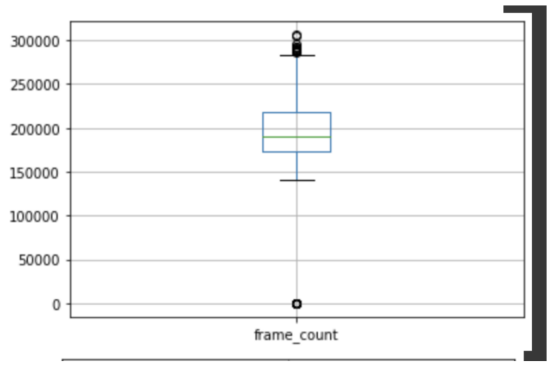

Figure 6: Box plot of frame count

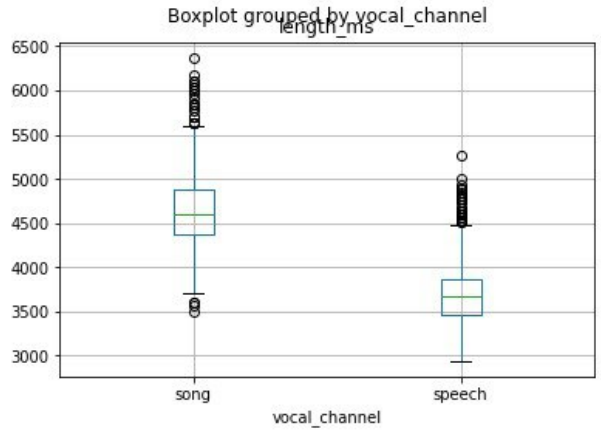

Figure 7: Box plot of lenght_ms looking to vocal channel

In (Figure 7) we may look at how the two types of vocal channel work in terms of duration: we can see that "song" has a much longer duration (near to 4500 milliseconds) in respect of speech which has less duration. Before analyzing the variables through a scatter plot it is important to make another consideration within this section: looking at outliers, through a box plot, we have noticed many of them are characterized by the missing values of "actor" and "vocal channel". In the end, the other outliers were considered special cases and not like errors like in the previous case(for instance, for the feature "channels", 2446 records belong to the channel "mono", instead of 6 records belonging to the channel "stereo-audio").

2.0.5 Continues attribute's analysis through scatter plots

Following the analysis with Box plot we wanted to continue the study of the dataset through the use of a scatter plot (or dispersion graph) thanks to which we could deduce some peculiar characteristics of RAVDESS:

- Scatter plot between "intensity" and "mfcc_min" compared to the emotional intensity: it is clear that high values of "intensity" correspond to high values of "mfcc_min" first of all for strong emotional intensity interpretation(probably this observation derives to the high level of correlation between two continuous mentioned variables).

- In (figure 9), scatter plot between "mfcc_min" and "max" compared to the emotional intensity(high values of "mfcc_min" corresponds to high values of "max" which identify a "strong" emotional intensity). For this purpose, we may say that there is a transitive rule between "mfcc_min", "intensity" and "max" such that also at a high level of "intensity" we may see a high level of "max"(and "mfcc_min), similar to what showed in (figure 8-9). This property was also noticed for the "std" feature.

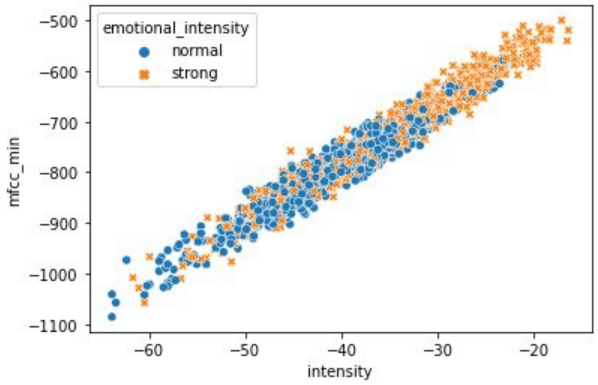

Figure 8: Scatter plot of intensity and min value of mfcc compared to emotional intensity

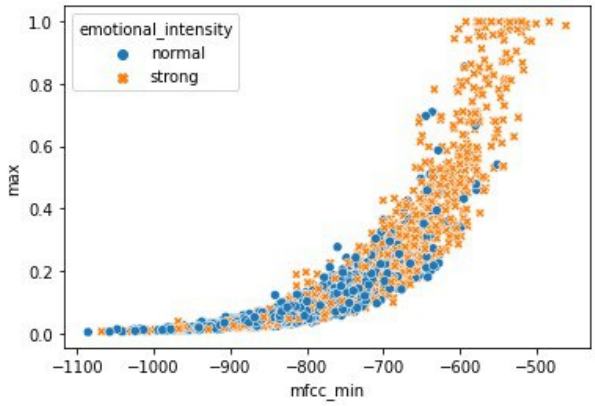

Figure 9: Scatter plot of mfcc_min and max

3 Data preparation

After the analysis of the dataset using graphical representations, we made changes that allowed us to obtain the best results necessary for the development of the project and understanding of the dataset itself. In the first instance we made the following deletions:

- elimination of the variables which contain a constant value;
- elimination of the "actor" because it contains 1126 missing values;
- elimination of "intensity" as it included 816 missing values and, through the correlation matrix, we noticed that this variable had a high positive and high negative correlation with other variables;
- elimination of "channels" since it has been noticed that it had a unitary correlation with "frame width";
- elimination of 35 incorrect frame count records.

We also tried to compensate for the missing values or deleted them through the use of other variables: the missing values of the "vocal_channel" have been filled using a classification model(as we argue in chapter 5). Following the decisions made for the study of the dataset, the correlation matrix was then recalculated , from which we found an interesting correlation equal to one between the "lenght_ms" and "frame_count" variables.

As a result of the values that the new correlation matrix has given us, we have proceeded to eliminate the variable "frame count" as this had a correlation equal to one with the variable "lenght_ms", the latter is able to provide more detailed and complete information about our dataset. For this reason, we have inserted again the previous deleted rows. The interesting variables we decide to not delete are "frame_width" and "sc_min", these have a high number of constant values except for some of them. In fact, they may suggest some useful information for other types of tasks

4 Clustering

This section describes the application of the main clustering algorithms (K-Means,K-Means variants, DB-Scan, Hierarchical) on that dataset and their respective results.

For this purpose, the numerical features have been normalized through the StandardScaler

4.1 Cluster analysis through K-means Algorithm

4.1.1 Optimal K choice

To identify the best k parameter for the KMeans, the different steps followed are :

- k has been varied between 2 and 29;
- For each value of k the SSE has been calculated to represent all values returned in a graph;
- Silhouette has been calculated for each k value;
- At this point the corresponding k value was taken into account by making a trade-off between the SSE values and the Silhouette value.

4.1.2 Experiment using all the continue features of the dataset

After normalizing the values of the variables, only the continuous variables remaining in the dataset are taken into account ('frame_width', 'length_ms', 'zero_crossings_sum', 'mfcc_mean', 'mfcc_std', 'mfcc_min', 'mfcc_max', 'sc_mean', 'sc_std', 'sc_min', 'sc_max', 'sc_kur', 'sc_skew', 'stft_mean', 'stft_std', 'stft_min', 'stft_kur', 'stft_skew', 'mean', 'std', 'min', 'max', 'kur', 'skew'). The graph in (figure 10) shows the variations of the SSE; based on the elbow of the function and the various silhouette values obtained, it was decided to take a value of k equal to 7.

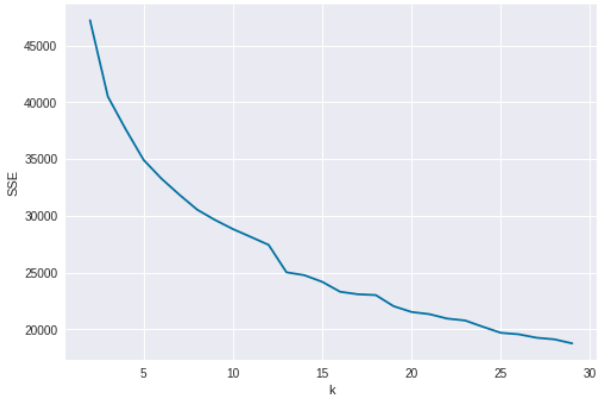

Figure 10: SSE graph

In (Figure 11) shows the coordinates relative to the centroids of the variables considered.

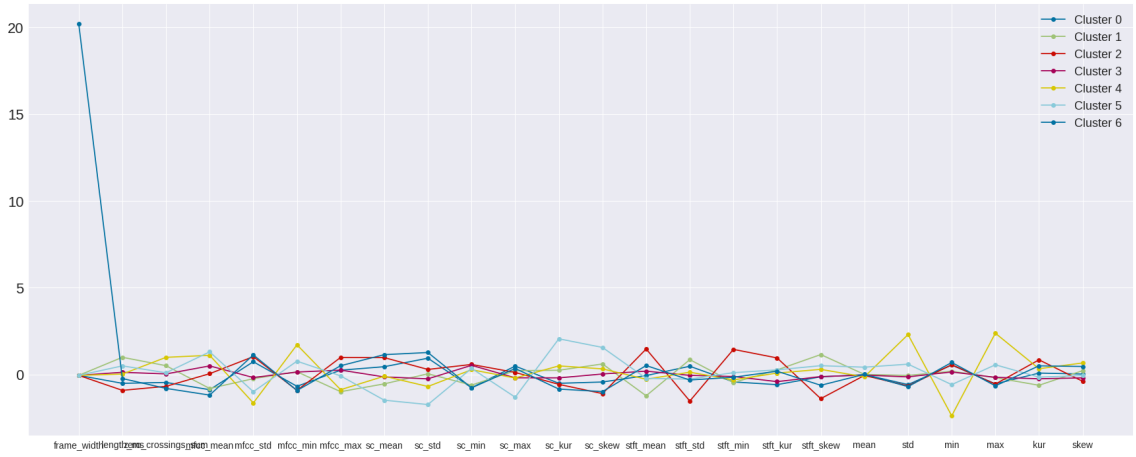

Figure 11: Parallel Plot of all the continuous Attribute

As you can see from the graph in the figure, the centroids are not well separated, this is because probably this experiment takes into account the very large size of the dataset. Therefore, we try a different approach to reduce further the size of the dataset by considering only a subset of continuous variables.

4.1.3 Experiment using subsets of continuous variables of the dataset

The experiments conducted follow these steps :

- A fixed variable class is defined: "zero_crossings_sum" and "length_ms";
- Several combinations of different variables related to audio statistics are added to this subset;
- The application of K-Means is tested using the method explained in 4.1.1;
- The parallel coordinate graph is used to display the distribution of centroids.

It turns out that the subset capable of returning the best graph of parallel coordinates at the centroid separation level is the following : "zero_crossings_sum " , "length_ms" , "mfcc_min" , "mfcc_max" , "mfcc_mean" , like (figure 12) shows. This is probably because the number of outliers present in these variables was particularly lower than in the others.

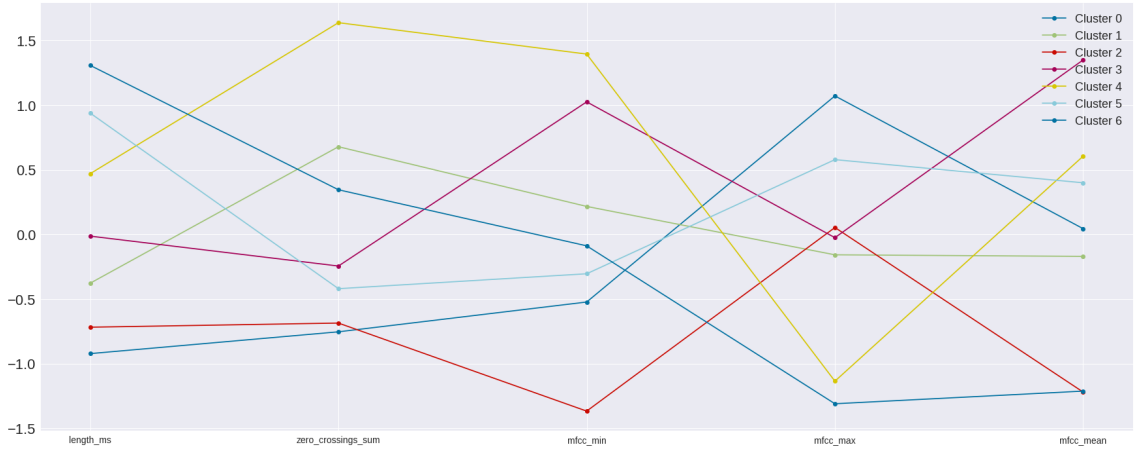

Figure 12: Parallel plot of : "zero crossings sum ", "length ms", "mfcc min" , "mfcc max", "mfcc mean

In this case, k = 7 has been chosen, as given in the SSE curve in (figure 13).

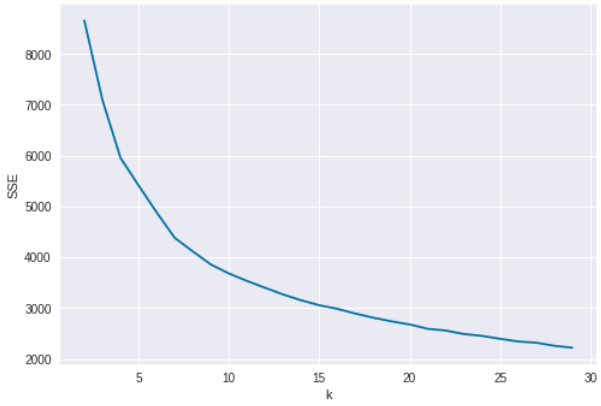

Figure 13: SSE graph

We take an SSE value of 4382.68072 and a silhouette of 0.229685. This subset has been replicated in all subsequent experiments using different algorithms

4.1.4 Observation

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
500	388	285	334	308	282	355

Figure 14: Clusters distribution

The (figure 14) shows the distribution of the records for each cluster; from a first analysis, we can see that the elements are well distributed among them.

Through this scatter plot in (Figure 15) which represents how the 7 clusters (related to the variables 'zero_crossings_sum' and 'mfcc_min') are distributed in the Cartesian space and their respective centroids, we can see that some of them are well separated, while others are not.

Figure 15: Scatter plot of Zero crossing sum and Mfcc min

As a follow-up analysis, it was decided to compare each categorical variable with the labels returned by clustering. In particular, was considered relevant how cluster 5 (seeing figure 16 - 17) can group some records related to all emotions except the surprise and some records characterized only by the intervention of women actors(because as we will see in chapter 7, women have the "zero_crossings_sum" greater than men).

Labels	0	1	2	3	4	5	6
emotion							
angry	33	66	115	131	2	8	21
calm	51	14	10	5	112	82	102
disgust	65	68	18	9	18	2	12
fearful	55	66	70	93	23	32	37
happy	81	57	49	67	19	45	58
neutral	57	10	4	2	38	39	38
sad	64	48	17	16	72	74	85
surprised	94	59	2	11	24	0	2

Figure 16: Emotion distribution in the clusters

Labels	0	1	2	3	4	5	6
sex							
F	102	306	237	47	227	282	3
M	398	82	48	287	81	0	352

Figure 17: Sex distribution in the clusters

As we see in (Figure 5) of section 2.0.3, we may generalize that cluster 5 gathers records that contain "song" (characterized by a high length in milliseconds) by looking at the lack of "surprised" and "disgust" emotions(only two records) in the cluster.

4.2 K-MEANS variants

4.2.1 Bisecting K-MEANS

For this type of experiment, k is chosen as shown in 4.1.1, focusing on the choice of the cluster which we want to divide, is based on the biggest SSE. The cluster with k equal to 10 and SSE equal to 4190.6005 was chosen, with a silhouette of 0.10.

The distribution in (figure 18) is the following:

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
213	175	232	216	362	255	291	247	223	238

Figure 18: Cluster distribution

from which it can be inferred that some data are well distributed between clusters.

bkmeans_labels	0	1	2	3	4	5	6	7	8	9
emotion										
angry	2	5	52	11	47	78	110	36	3	32
calm	60	102	19	50	1	30	1	8	76	29
disgust	25	24	45	26	2	17	7	34	0	12
fearful	18	24	73	26	24	44	65	54	9	39
happy	13	21	93	28	18	67	38	35	8	55
neutral	48	25	23	37	0	9	0	5	23	18
sad	50	63	47	46	3	43	7	23	61	33
surprised	29	26	91	3	6	2	2	30	0	3

Figure 19: Cluster distribution referred to emotion

From the table in (figure 19), we can see how cluster 8, most likely, can discriminate songs, not including emotions surprised and disgust.

4.2.2 X-Means

For this type of experiment a k = 20 was chosen as the maximum number of clusters and at the end of the computation the algorithm returns 4 clusters, distributed in the following way (figure 20) :

Cluster 0	Cluster 1	Cluster 2	Cluster 3
345	724	536	847

Figure 20: Clusters distribution

Although again the data are well distributed among the 4 clusters, but no interesting properties could be found as in the previous case. In addition, the silhouette index is very low, amounting to 0.0019.

4.3 Cluster analysis using DBScan Algorithm

For these experiments, it is necessary to accurately choose the value to be assigned to two parameters: MinPoints and Epsilon. To solve this problem it was chosen to use as a MinPoints number a value higher or equal to the size of the dataset plus one, then in this case 6. Next, k-NearestNeighbor was tested, where k is the number of MinPoints fluctuated at each iteration (up to 60). This technique considers the average distance between noise points and the nearest k points. In particular, the values are plotted in a graph and from this we take into account the elbow of the returned function, from which we get the value of Epsilon.

However, the experiments are not successful, in particular with the variation of the number of MinPoints and Epsilon have found the following problems :

- we obtain maximum 2 clusters;
- In some cases noise points were very high;
- Many elements were concentrated in one cluster;
- In some cases very low silhouette values;

4.4 Cluster analysis using Hierarchical Algorithm

The Euclidean distance has been used to implement the model in question, making it easier to compare and interpret the results.

The algorithm is executed using different configurations and in particular trying to vary :

- The connection policy between clusters: single, ward, average and complete;
- The use of a connectivity matrix (derived from the graph obtained from the first 100 neighbors of each node) for the calculation of distances.

Based on the dendrograms obtained in the (figures(from 21 to 24)), the choice of cluster size was made.

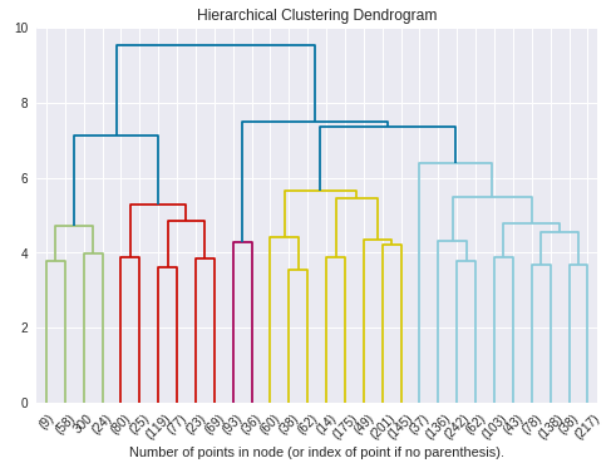

Figure 21: Hierarchical clustering with complete linkage

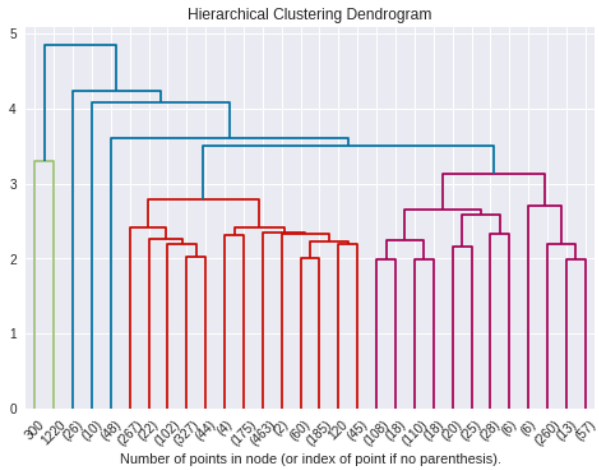

Figure 22: Hierarchical clustering with average linkage

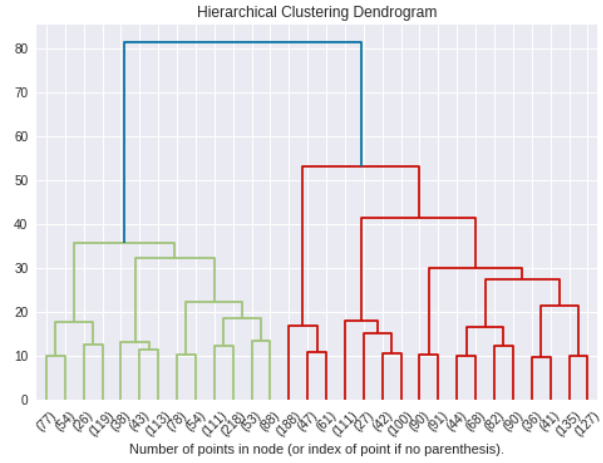

Figure 23: Hierarchical clustering with Ward linkage

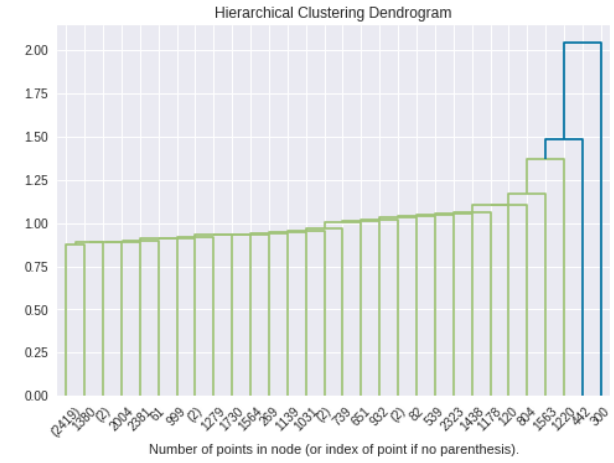

Figure 24: Hierarchical clustering with Single linkage

As we can see from the dendrogram, the single linkage does not give satisfying results, it was therefore decided to focus the analysis on the other configurations. For each configuration tested, the silhouette was calculated to assess the goodness of the parameters.

In (figure 25) the various criterion were compared :

Method	Cluster's Dimension	Silhouette Value
Average + connectivity	2414 , 2 , 26 , 10	0,13
Complete	2449 , 1 , 1 , 1	0,22
Average + connectivity	2414 , 2 , 26 , 10	0,13
Complete	1094 , 744 , 129 , 393 , 92	0,12
Complete + connectivity	186 , 22 , 1 , 10 , 2233	0,11
Ward	1072 , 804 , 296 , 280	0,22
Ward + connectivity	655 , 938 , 568 , 291	0,20

Figure 25: Methods used for linkage

The following considerations can be deduced from the table:

- Adopting the average and complete linkage configurations to the connectivity matrix results in poor performance;
- For the Ward criterion, the performance is identical using or not the connectivity matrix. You also get balanced clusters compared to other configurations.

4.5 Best cluster obtained comparison

Given the poor performance of the clusters obtained through DBScan, it was decided to make a final comparison only between the cluster obtained through K-Means, their variants, and the clusters obtained through hierarchical clustering, summarized in (figure 26)

Algorithm	Number of Clusters	Silhouette Value
K-Means	7	0,22
Bisecting K-Means	10	0,10
X-Means	4	0,0019
Hierarchical	4	0,22

Figure 26: Comparison cluster results

From the final comparison, the algorithm of K- MEANS was considered better, as it guarantees the best trade-off between the value of silhouette and significance with regard to the internal distribution of data.

5 Classification

This section describes the results of the following classification algorithms: Naive Bayes(Gaussian), K-NN, and Decision Tree. The above models require the setting of a target variable that will be the starting point of the classification procedure. It was considered interesting to analyze the variable "emotion" because it was considered more relevant from an analytical point of view; however also following the setting of hyperparameters suggested by GridSearch, they presented the phenomenon of overfitting (due to their imbalance): we're able to generalize very well on the training data but were characterized by poor performance on the validation set(although One vs. One and One vs. Others methodologies were also tested and repeated stratified K fold validation was used to balance the classes in the experiments). Through appropriate reflections,

it was concluded that we wanted to use the variable "vocal_channel" so that if good results were obtained(in terms of F1 score and accuracy), the returned labels could be considered to replace the missing values of this variable.

5.1 Pre-processing

The first operation conducted was to perform a **One Hot Encoding** operation on the categorical variables. The numeric variables were normalized using the StandardScaler only for the experiments with the KNN(necessary for distance calculation). Next, records with missing values of the variable "vocal_channel" are isolated, which will then be useful for a possible prediction. For all three algorithms examined, the splitting of the dataset in the proportion 70-30 is performed, thus 70% training set and 30% test set. The GridSearchCV (passing as input the obtained training set data) is used to find the best configuration of hyper-parameters that can maximize the F1 performance evaluation metrics, a solution deduced from the reflection that the use of Accuracy metrics alone could provide very good but very general results. Parameter optimization is inferred from the use of the **"Repeated K-Fold cross-validation" methodology**(with k equal to 5 and numbers of repetitions equal to 10, the repetition makes it possible to test different folds). The groups of hyper-parameters tested in the two different algorithms using it are shown in (figure 27) :

Decision Tree	K-Nearest Neighbors
criterion: gini, Entropy	"n_neighbors": range(3,30,2)
Max_depth:None, 2, 5, 10, 15, 20	'weights': ['uniform', 'distance']
min_samples_split': [2, 5, 10, 15, 20]	'metric': ["euclidean", "manhattan", "minkowski"]
'min_samples_leaf': [1, 5, 10, 15, 20]	

Figure 27: Configurations tested

We decide to choose a k-odd number to avoid possible situations of parity in the discrimination of the class.

5.2 K-NN

We apply, as announced, the GridSearchCV algorithm on the classifier obtaining the following results:

	Best Parameter set
N_Neighbors	11
Metric	"Manhattan"
Weights	"Distance

Figure 28: K-NN best configuration

Then we apply the inferred hyperparameters to the classifier that is trained on the training set and predicts on the test set. For the performance evaluation of the same, using the conventional metrics, we obtain the following class-discriminated results:

	Precision	Recall	F1-score	Support
class song (0)	0.88	0.94	0.91	275
Class speech(1)	0.96	0.92	0.94	402

Figure 29: Results returned by KNN

which results in an accuracy level of 93% and an F1-macro of 92%. The above data can also be interpreted from the observation of the ROC that registers an AUC of 99%. From the Confusion Matrix, on the other hand, we can infer how the Class 0 and Class 1 values were almost all correctly predicted except for a few cases that justify the lack of a few percentage points on precision, especially speaking of Class 0.

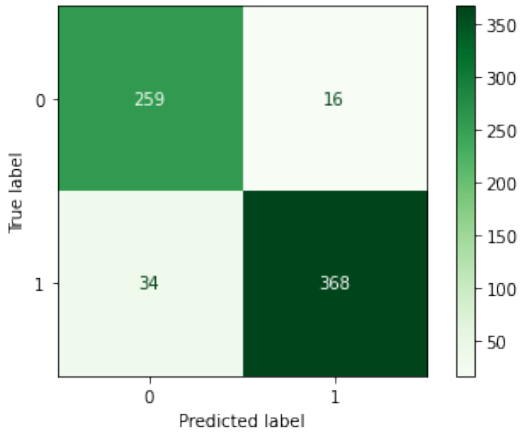

Figure 30: Confusion matrix obtained by K-NN

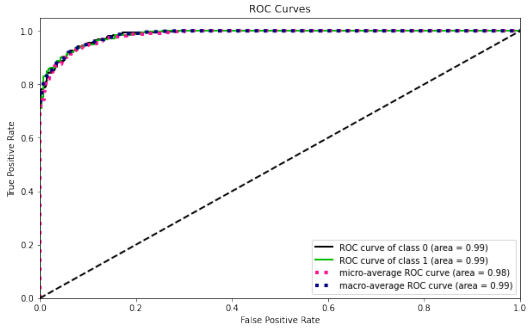

Figure 31: ROC curve obtained by K-NN

5.3 Naive Bayes

For this type of classifier, there are no hyperparameters to set but we will evaluate the general performance that the classifier achieves in class prediction :

	Precision	Recall	F1-score	Support
class song (0)	0.79	0.99	0.88	275
class speech (1)	0.99	0.82	0.90	402

Figure 32: Results returned by Naive Bayes

We obtained a level of accuracy, an F1 macro score of 89% and an AUC of 97%(Figure 34). We note a Precision level for class 0 of 79% which also results in a high number of false predictions for class 0 in the confusion matrix(figure 33). Similar reasoning for the recall of class 1 which records 71 negatively predicted records in this case, a higher rate inferable from a higher correct prediction for class 1.

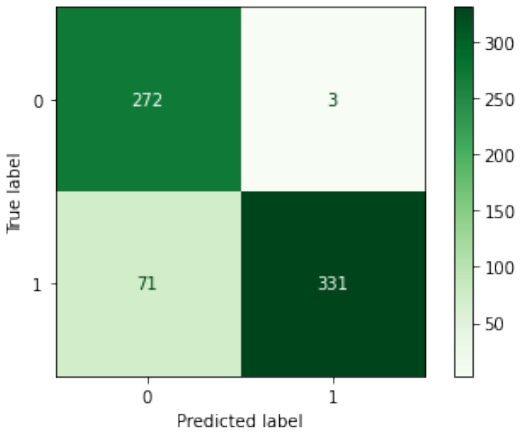

Figure 33: Confusion matrix obtained by Naive Bayes

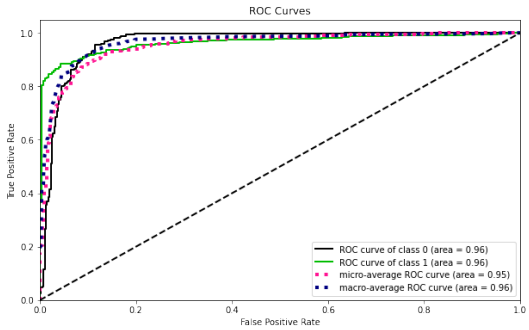

Figure 34: ROC curve obtained by Naive Bayes

5.4 Decision Tree

The hyper-parameters that the GridSearchCV returns to us, as a result, are as follows in (figure 35) :

	Best Parameter set
Criterion	Entropy
Max depth	15
Min samples leaf	1
Min sample split	2

Figure 35: Decision Tree best configuration

By applying the hyperparameters obtained from GridSearchCV we obtain performance values for the three main metrics as in (figure 36):

	Precision	Recall	F1-score	Support
class song (0)	0.92	0.95	0.93	275
Class speech (1)	0.96	0.94	0.95	402

Figure 36: Results returned by Decision tree

We obtained an accuracy and F1-macro value of 94% and AUC of 94%. Confirmation of these results can be seen in the confusion matrix and the roc curve (figure 37-38) :

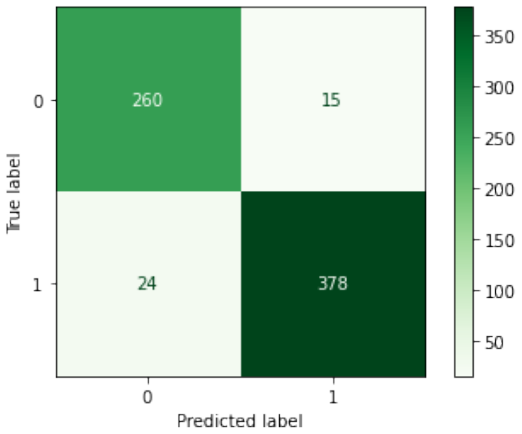

Figure 37: Confusion matrix

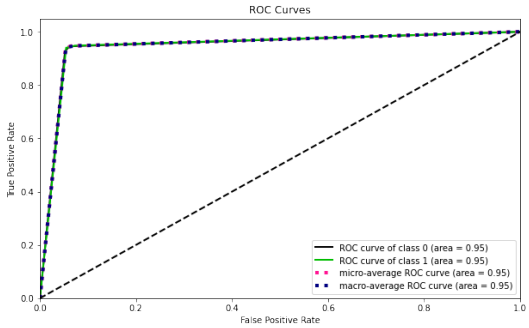

Figure 38: ROC curve of DT

We note the importance of the feature "lenght_ms" in the order of about 65% and 9 other features comprising for the most part statistical variables as visible from the graph and 'tree in the figures :

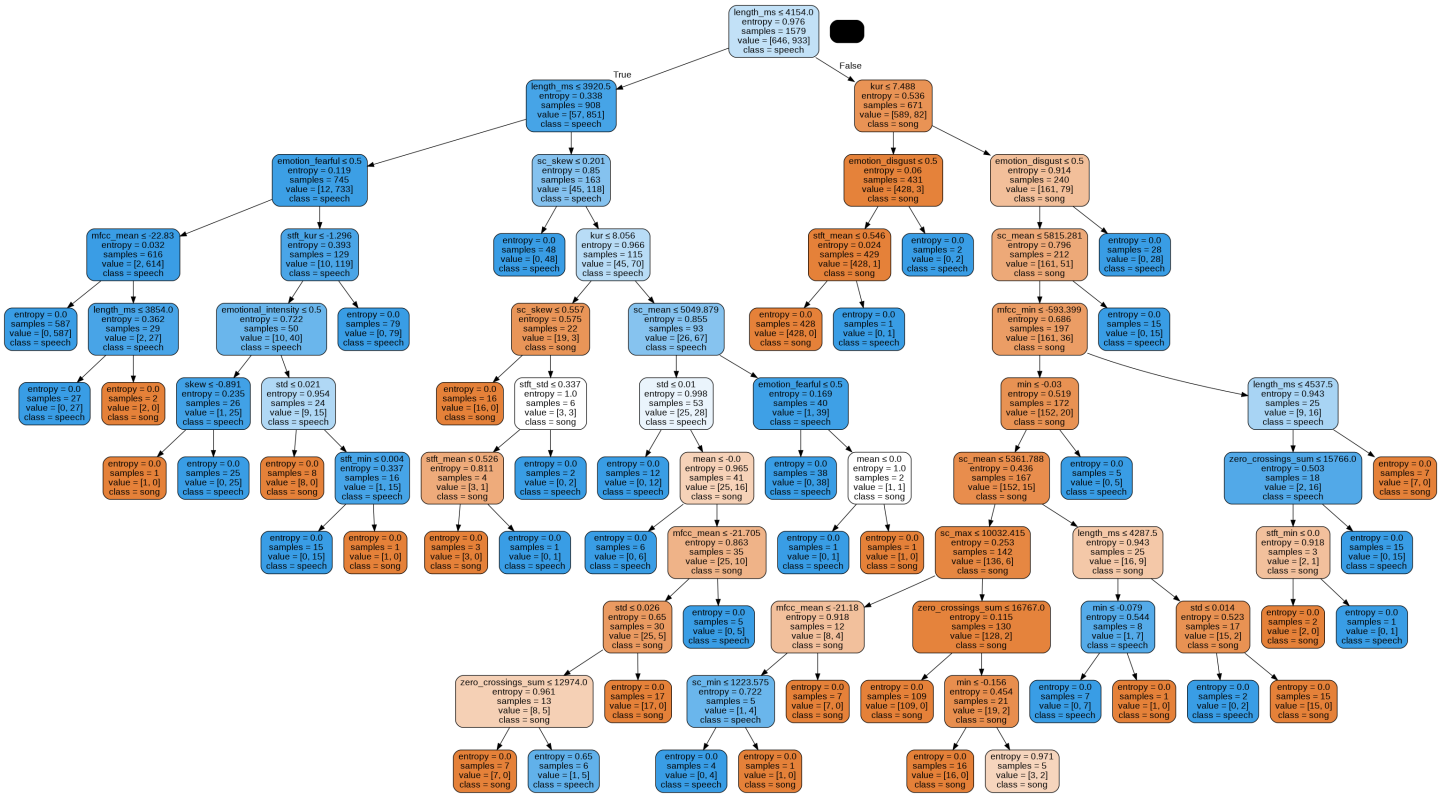

Figure 39: Decision tree figure

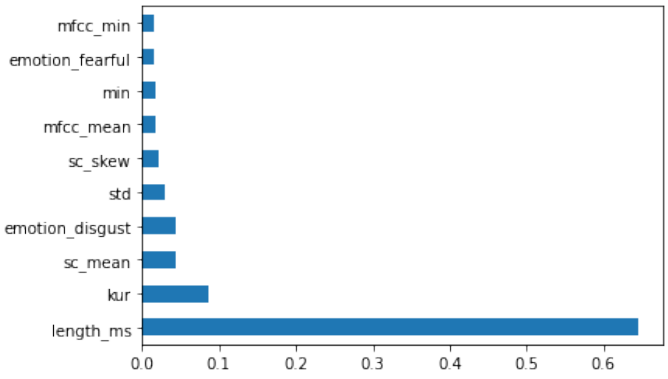

Figure 40: Most relevant features

The Decision tree in (figure 39) shows how the variable "lenght_ms" actually succeeds in distinguishing and classifying records well in the order that they belong to the class "speech" or "song" and also allows us to give strength to the inferred reflections in the data understanding (paragraph 2.0.4). Surprisingly, although playing a little role, the emotion fearful contributed to class discrimination, while it is not surprising that the absence of the emotion of disgust is a relevant feature in song discrimination.

5.5 Conclusions

The predictive algorithm that gave us the best results turns out to be the Decision Tree with an overall F1 macro level of 94% although however Naive Bayes and K-NN also record optimal and satisfactory values. As could be seen from the ROC curves, each of them had curves closer to the ideal ones (those with AUC =1), so they were able to accurately distinguish the two classes. It was therefore decided to replace the missing values with the labels returned by the Decision Tree.

6 Linear regression

The goal of this task was to predict the missing values of the feature "intensity"(dependent variable Y) so that the information obtained from this variable can extract new knowledge for future experiments. The dependent and independent variables considered were normalized with the StandardScaler and as for the classification task, once the records containing null values were separated, the dataset was split into 70% training and 30% test set. Again, Grid search(exploiting K fold cross validation) was used to find the best configuration of parameters regarding K-NN and Decision Tree.

6.1 Univariate regression

The first experiment performed was to consider the independent variable X the mfcc_min being the feature most positively correlated with the chosen Y (97 percent correlation). The results using various regressors are shown below :

Measures	Linear	Ridge	Lasso	KNN	Decision Tree
R^2	0,95	0,94	-0,05	0,94	0,94
MSE	0,049	0,04	0,992	0,053	0,055
MAE	0,180	0,180	0,802	0,186	0,189

Figure 41: Univariate regression results

As can be seen from the results in (figure 41), the best model in terms of R$\hat{2}$ and error minimization is simple linear regression.

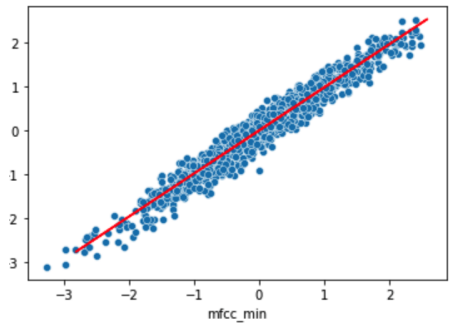

Figure 42: Linear regression plot

As can be seen from (figure 42), in fact, the distance between the line and the points is minimal, therefore the correlation between two variables can be explained by the equation of a straight line

6.2 Multivariate regression

For this type of task, the variables max and std were taken into account being the other two features most correlated with 'intensity (80 % and 84%, respectively). The results using various regressors are shown below(figure 43):

Measures	Linear	Ridge	Lasso	KNN	Decision Tree
R^2	0,66	0,66	0	1	1
MSE	0,327	0,326	0,98	0	0
MAE	0,428	0,438	0,77	0,014	0,004

Figure 43: Multivariate regression results

As can be seen from this table, K-NN and decision Tree perform better than the other regressors, achieving almost perfect results on the test data, as can be seen from the graph below describing Decision Tree performance :

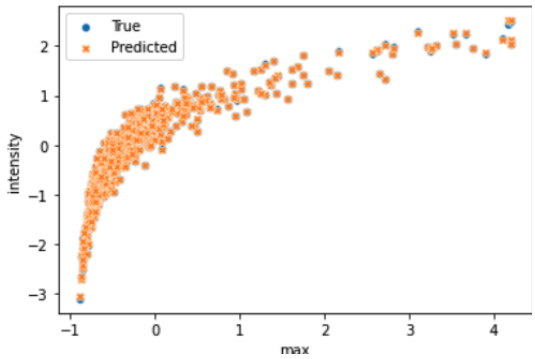

Figure 44: Performance of Decision Tree regressor

Below are the tables(figure 45-46) with the best hyper-parameters found :

KNN	Best Parameter set
N_Neighbors	11
Metric	“Manhattan”
Weights	“Distance

Figure 45: K-NN regressor configurations

Decision tree	Best Parameter set
Criterion	Entropy
Max depth	15
Min samples leaf	1
Min sample split	2

Figure 46: DT regressor configurations

6.3 Conclusion

In conclusion, being the decision tree regressor is the best performing model, it was decided to replace the null values based on the results they returned

7 Pattern mining

The following section explains the methodologies used for the identification of interesting patterns to extrapolate useful information from our dataset

7.1 Pre-processing

The dataset endures modifications: elimination of variables with difficult insights for this particular task, keeping just features like ”vocal_channel, ”emotion”, ”emotional_intensity”, ”sex”, ”lenght_ms”, ”zero_crossings_sum” and ”intensity”. These last, to make the algorithm **Apriori** available, they endure these operations:

- *”lenght_ms”*, discretized according to the quartile;
- *”zero_crossings_sum”*, discretized in 5 bins, each of them labeled with *”Very Low”, ”Low”, ”Medium”, ”High”, ”Very High”* taking into account also the distribution of the attribute;

- *intensity*, discretized in 5 bins, each of them labeled with *"Very Low", "Low", "Medium", "High", "Very High"* taking into account the distribution of the attribute.

7.2 Frequent pattern extraction

To define the level of minimum support, the number of itemset *(frequent, closed and maximal* is plotted with a minimum length in a range [2,5] to show the variation of the "min_supp:

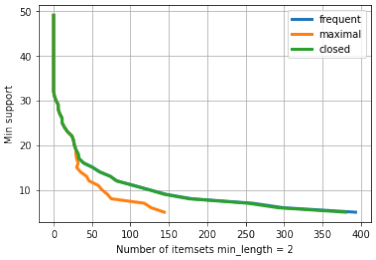

Figure 47: Number of itemset min_lenght=2

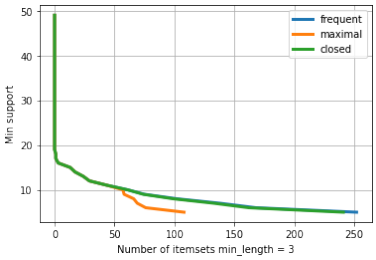

Figure 48: Number of itemset min_lenght=3

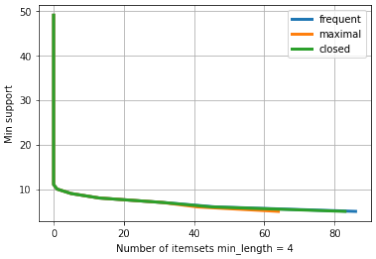

Figure 49: Number of itemset min_lenght=4

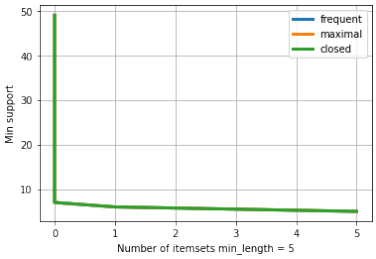

Figure 50: Number of itemset min_lenght=5

looking at Figure [47-50] we notice how, as the minimum number increases in the itemset, these tend to decrease.

For the generation of frequent itemsets, it was taken into consideration a *"min_supp"* of 10 and *"z_min"* equal to 3 (minimum length), below some of them:

	Frequent itemset	Support
1	(2935.999, 3604.0]_Length,normal_EMIntensity,speech_Vc	18.515497553017944
2	(4538.0,6373.0]_Length,song_Vc,normal_EMIntensit	12.561174551386623
3	Low_Intensity,normal_EMIntensity,speech_Vc	16.27243066884176
4	Medium_Intensity,normal_EMIntensity,Low_Zcs	15.415986949429037

Figure 51: Frequent with min_supp=10 and z_min=3

the results show that the songs have a greater average length compared to the speech with a "normal emotional intensity" (referred to in rows 1-2 in the table), it turns out that a low value of "intensity"(in dCb) is associated at a "normal emotional intensity" and this association is made in the presence of a

speech (row 3 in the table), as well as an intensity in volume decibels "average" associated with a normal emotional intensity develops a zero Crossings sum low. To gain more knowledge, it was decided to lower the min_support to 5%, below some :

	Frequent itemset	Support
1	Medium_Zcs,Medium_Intensity,F	18.931484502446981
2	Medium_Zcs,song_Vc,F	8.564437194127244
3	High_Intensity,song_Vc,F	5.34257748776509

Figure 52: Frequent with min_supp=5 and z_min=3

By decreasing the level of support we get rules less trivial and therefore more significant:

- women have an average zero crossing sum, which allows us to assume that the variable "zero_crossings_sums" can be seen as a substitute for a voice frequency, introducing a new analysis cue for the dataset;
- Songs have a higher intensity than speech.

7.3 Rules extraction

Also in this section, you can see how, by changing confidence and supports parameters, you get different rules. The analysis is conducted for a minimum support set (5%, 10%, 15%), as shown:

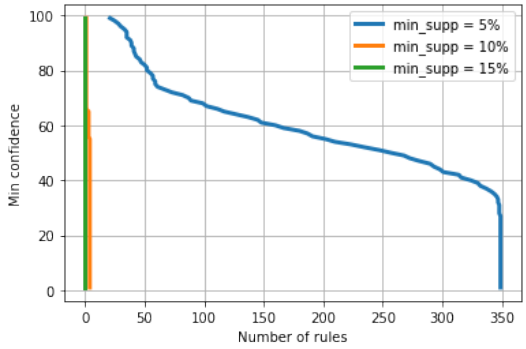

Figure 53: Number of rules

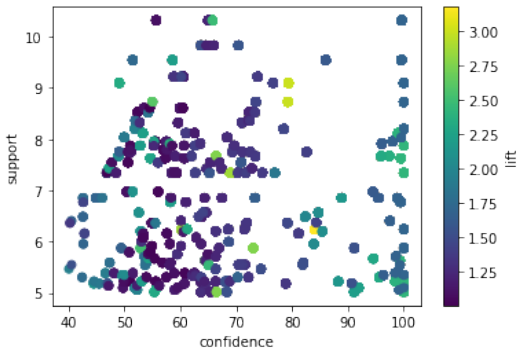

Figure 54: Parameters distribution

In the graph on the left, for min_supp=10% and min supp=15% the number of rules extracted is close to zero while for min_supp=5% a considerable number of rules are extracted. From the right graph, we can see that a considerable number of rules is extracted in a range of support from 5 to 8.5 and a range of confidence from 40 and 80 that contributes to generating more rules with a lift greater than 1, in some cases greater than 2. The following figure shows histograms showing how taking a min_supp=5 into account and starting from a min_conf=80% the distribution of the rules varies.

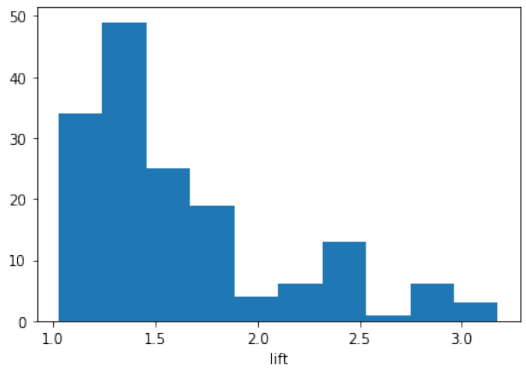

Figure 55: Distribution of lift

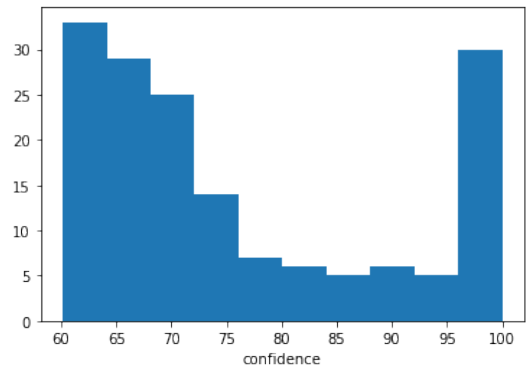

Figure 56: Distribution of confidence

Here are some rules extracted that are interesting (figure 57):

Consequent	Antecedent	Support_abs	Support_perc	Confidence	Lift
song_Vc	(4538.0,6373.0]_Lenght,Medium_intensity,strong_EMIntensity	125	5.09787928	91.911764	2.229155757
song_Vc	(4538.0,6373.0]_Lenght,F,normal_EMIntensity	180	7.34094616	100.0	2.42532146
Low_Zcs	song_Vc,M,normal_EMIntensity	204	8.3197389885	71.08013937	1.3163784119

Figure 57: Rules extracted

you can see that the first two have a lift greater than 1, then it has a positive correlation, while the last little bigger than one then antecedent and consequent are almost independent.

7.4 Prediction of the target variable

In order to compare the results obtained by the predictive algorithms it was decided to try to classify the variable "vocal_channel" using the rules obtained, using a min_confidence=60% and a min_supp=5%(we have taken into account this threshold of min_supp because it allows us to reach e greater number of rules). Of the returned rules have been taken into account those with lift greater than 1 and confidence greater than 78%, in (figure 58) below we report the most interesting.

consequent	antecedent	support_ab	support_pe	confidence	lift
speech_Vc	Very Low_Zcs, (2935.999, 3604.0]_Length ,Low_Intensity	144	5.87275693 3115824	100.0	1.70159611 38
speech_Vc	Very Low_Zcs,Low_I ntensity,M	127	5.17944535 07340945	78.8819875 7763976	1.34225283 51

Figure 58: Rules extracted

As it can be seen from the considerations made at the beginning, the length in milliseconds of the songs is greater than the speech, also in this circumstance, it is clear that men have a "zero_crossings_sum" much lower than women. It has also been noted that the rules having as a consequence "speech_Vc" are far greater than those having as a consequence "song_Vc" and despite this, they have a lower lift (they are all lower than 2). Regarding the prediction of the target variable, the following procedure was followed:

- an additional "prediction" column is created in the dataset initially filled only with the value "speech" being the most recurring value in the rules;
- all the rules having as a consequence "song" have been taken into consideration, the rows for which the conditions expressed by the antecedents were verified have been localized in the dataset;
- follows that the "prediction" column corresponding to these rows has been changed to the value of "song"
- At this point the metrics of accuracy, precision, recall, F1-score and the confusion matrix are calculated (comparing the value reported in the column "prediction" with the actual value of "vocal_channel") as shown in (figure 59-60):

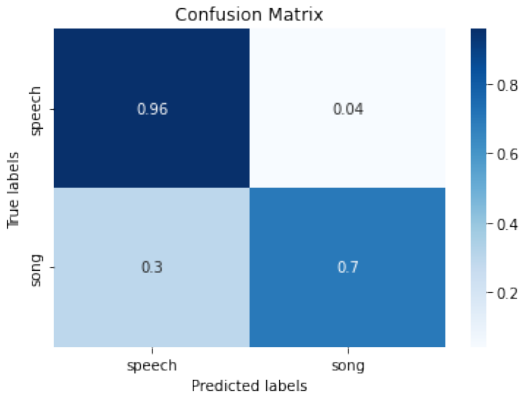

Figure 59: Confusion matrix

Accuracy	0.85
Precision	0.8219
Recall	0.9604
F1-score	0.8858
TN-FN	711 - 57
TP-FP	1384-300

Figure 60: Metrics

In conclusion, it is noted that the accuracy of the method used is lower than the ones obtained by the Decision tree analyzed in the previous section, this despite it discriminates well between the two classes.