Project 2 – Machine learning applied to Tango

Develop a model that helps Tango DJs to select songs

1.1 General background

Tango is an Argentinian dance that has gained popularity in a number of countries in the world especially in Europe. Many cities hold regular social events called ***milongas***, which are attended by people with various levels of dancing, ages, and social backgrounds. Just as any other dance party, a milonga requires a DJ which is the key element to keep the atmosphere pleasant and control the energy of the evening. The role of the DJ is not trivial because there are several types of songs, many orchestra and different eras. There are a number of universal guidelines that are followed in every milonga:

- Songs are organized in groups of 4 songs called ***tandas***;
- Each tanda comprises songs of the same orchestra, singer, similar year, and similar tempo;
- **Orchestras that have been extremely influential to the tango culture should be played at least once;**
- **The energy should increase gradually during the night, reach a climax, and then decrease;**

The two last guidelines can be further explored to help tango DJs to select suitable tandas, as illustrated in Figure 1.

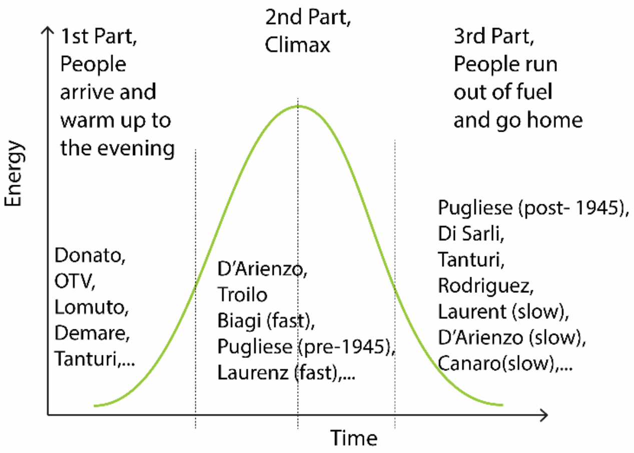


Figure 1. An exaggerated view of the flow of energy during the evening

Figure 1 shows that orchestras with lower energy should be played during the 1st and 3rd part of the milonga, and orchestras that are more influential and have higher energy should be played during the main period of the milonga (2nd Part).

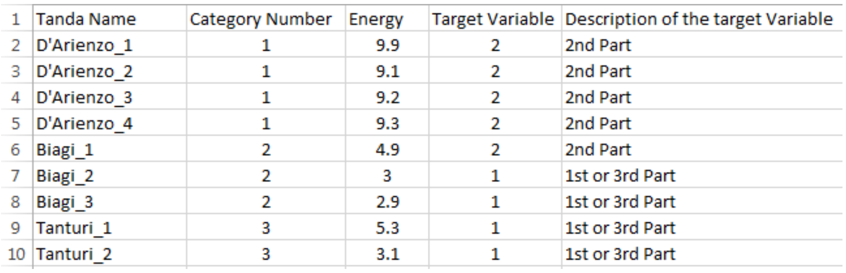
1.2 Objective

The overall objective of this project is to develop model based on machine learning that could be used to select tandas at different periods of a milonga.

1.3 Project methodology and approach

A new machine learning approach is investigated based on **Classification**. The selected algorithms are **linear Support Vector Machine (SVM)** and **Naive Bayes (NB)**. The project is developed in **python** and the machine learning algorithms are applied using the open source library **sciKit-learn**.

Using the first two guidelines referred in Section 1.1, tandas are organized in iTunes and exported to excel in form of a table (*i.e.* each tanda is an instance of a dataset). A category number is given for each orchestra, *i.e.* orchestras that had high, medium and low influence are given 1, 2 and 3, respectively. Remember that each tanda has only songs of the same orchestra, and thus the category number is an attribute of an instance in the dataset. For each tanda, an energy value is given by personal experience. Therefore, this is a project with two independent variable (*i.e.* category number and energy). For the target variable, the problem was simplified by attributing the value 1 to a tanda that should be played in the 1st or 3rd part of the milonga, and the value of 2 to a tanda that should be played in the 2nd part of the milonga. A print screen of a sample region of the dataset is shown in Figure 2 for illustration.



1	Tanda Name	Category Number	Energy	Target Variable	Description of the target Variable
2	D'Arienzo_1	1	9.9	2	2nd Part
3	D'Arienzo_2	1	9.1	2	2nd Part
4	D'Arienzo_3	1	9.2	2	2nd Part
5	D'Arienzo_4	1	9.3	2	2nd Part
6	Biagi_1	2	4.9	2	2nd Part
7	Biagi_2	2	3	1	1st or 3rd Part
8	Biagi_3	2	2.9	1	1st or 3rd Part
9	Tanturi_1	3	5.3	1	1st or 3rd Part
10	Tanturi_2	3	3.1	1	1st or 3rd Part

Figure 2. Sample of the dataset

The test set comprises 71 instance points and the skewness is [-0.05, 0.09]. Due to the low number of data points, only the train set is analyzed in this project.

1.4 Results

Figure 3 (a) shows the results of the SVM algorithm with two regions representing the predicted results (green and red) and the input data plotted as scattered points. As it can observed in this figure, the green region (*i.e.* tandas to be played in the 2nd part of the milonga) is mostly centered on data with category number 1 and higher energy. This green region also extends to some points that have category number 2 and very high energy. The practical interpretation is that tandas with songs that were very influential with medium/high energy or tandas with songs that were somewhat influential with very high energy could be played to reach the climax of the milonga. On the other hand, Figure 3 (a) shows that the red region is centered on data with category number 2 with low energy and category number 3 with all range of energies. The practical interpretation is that tandas with songs that were not as influential could be played in the beginning and the end of the milonga when not all the people have arrived yet or some people have already left.

Figure 3 (b) shows the results of the NB algorithm. This algorithm have an increased green region comparing to the SVM algorithm. An example that this model could be beneficial is in a milonga that have a wider range of ages. In this case it would be a good choice to play more influential tangos with lower energy in the 2nd part of the milonga for the older crowd but also other variety of tangos including less influential but with high energy for the younger crowd.

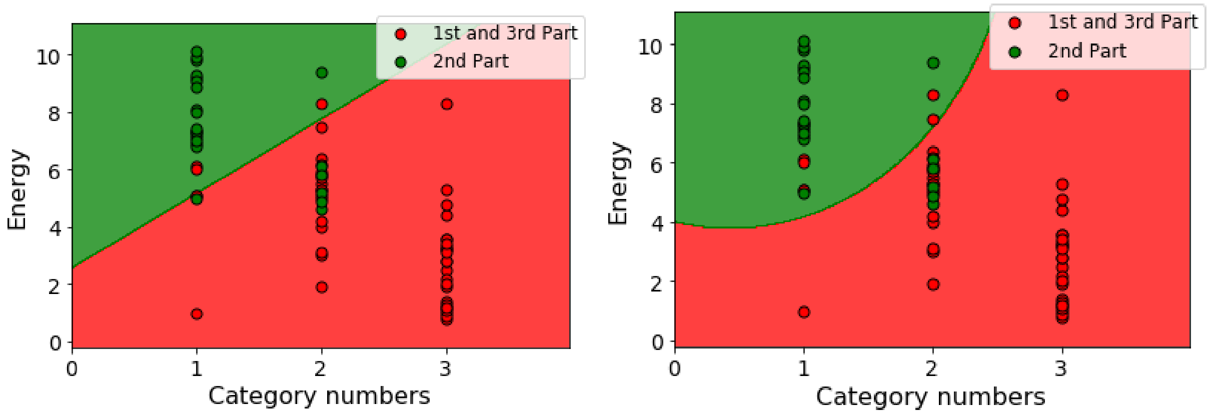


Figure 3. (a) SVM and (b) NB algorithms applied to the dataset

The confusion matrix for SVM algorithm is $\begin{bmatrix} \mathbf{44} & \mathbf{3} \\ \mathbf{6} & \mathbf{18} \end{bmatrix}$ and for the NB algorithm is $\begin{bmatrix} \mathbf{41} & \mathbf{6} \\ \mathbf{5} & \mathbf{19} \end{bmatrix}$