**Project**

**1. Overview**

* Training data set N=4363
* Each song has 264 features:
  + Statistics: consist of different types: mean, variants
  + Bands: amount of compromised time series within a song
    - 168 values of rhythmic
      * 24 bands, 7 statistics
    - 48 values of chrome
      * 12 bands, 7 statistics
    - 48 values of MFCC
      * 12 bands, 7 statistics
* 10 possible classes

Colums of feature

[0:168] includes 0, excludes 168: rhythmic

[168:216]: chrome

[216:264] MFcc

First, all of one statistic (first all means: mean of 1 band, mean of 2 band..)

Histogram:

* amount of classification distribution: Y

**2. What we found out**

Imbalanced data set:

* The training data is et is very imbalanced. According to <https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/> , imbalanced data set can be handled by using: What to do against imbalance data set:
  + Oversample/ Undersampling
    - Undersampling is not an option because we have to less examples
  + Penalized Models
  + Anomaly Detection
* According to <https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation> is it important to do first cross-validation and then oversample by using the function smote
  + That’s the reason why the normal oversampling do not fit well
* Outcome is that that validation is the real validation and not a fake one if using smote after cross validation
* Nevertheless, even the accuracy of the training data is the real one, the accuracy of the test data do not increase because I found out that
  + Dummy solution has accuracy of 50%, and just consist of class 1 objects.
* We now the distribution oft he test data: 50% class 1 => almost the same as the training set 2178/4363 => ca. 50%
* EVEN THAT APPROACH WAS NOT SUCCESS, IT IS WORTH TO MENTIONED IT IN THE REPORT
* SMOTE, def oversampling/undersampling is not usefull

Dimension reduction:

* It is more helpful to use a supervised dimension reduction instead of an unsupervised for supervised machine learning problem (what classification is)
  + LDA > PCA
    - We have to show that PCA is not better?
* For more information: <https://stats.stackexchange.com/questions/161362/supervised-dimensionality-reduction> at the end of the page

Decision tree:

* High fake validation for normal oversampled data

**3. Best scores:**

1. LDA features reduction for each of the three features mainly.
2. LDA features reduction to nine dimension (using 14 Features => clean=’wrong’, using data just for training. Logistic regression, solver unimportant
3. LDA features reduction to nine dimension (using 14 Features => clean=’wrong’, using 80% of data for training!
4. Without feature reduction (using 14 features, ca 252 colums) Logistic regression newton solver

**4. To do:**

* **Try semi supervised learning using the test dataset (recommended in the project paper**
* Showing PCA is not better
* Detecting outliners using correlation matrix
  + Finding more wrong data
  + I read that logistic regression do not like data with high correlation
* I messed up with the name in kaagle,( have to include our team number 772(you shoud register in mycourse)