**Classification**

SV42/2020 Danilo Babić, SV61/2020 Jelena Miković

# Problem

The goal is to predict the genre of music based on lyrics. The options were folk, pop and rock.

# Data

The training data set consisted of 4800 lyrics for folk, pop and rock songs.

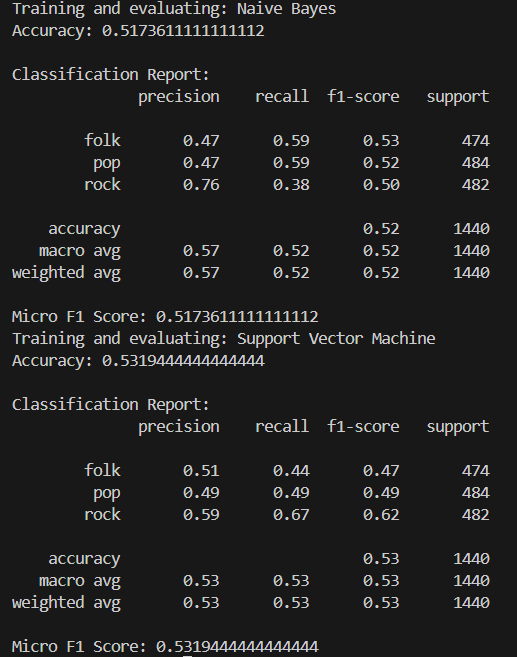
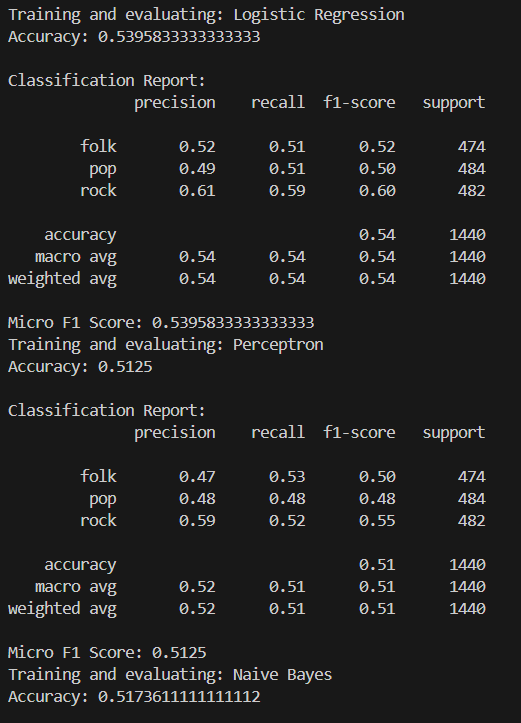
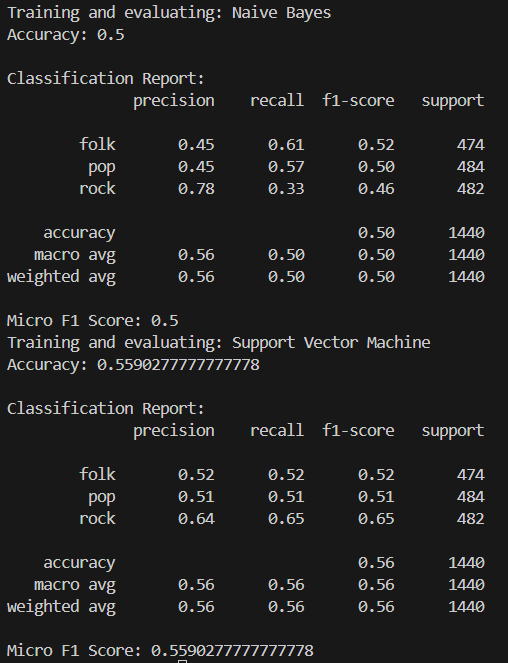
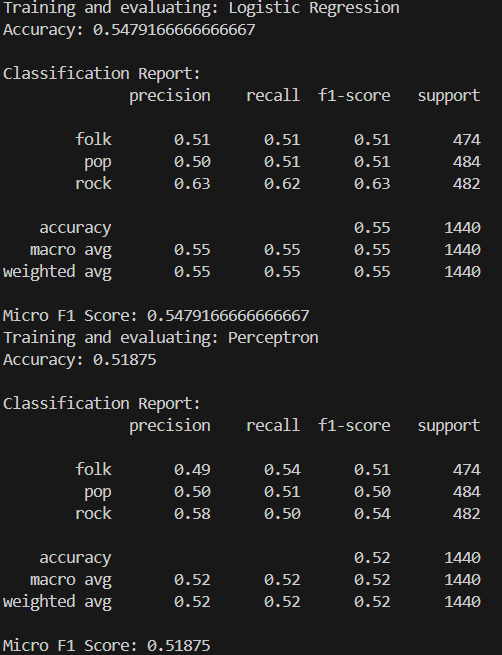
# Pre-processing

All of the lyrics are pre-processed. Text pre-processing involves several steps:

* Converting text to lowercase.
* Removing non-alphanumeric characters.
* Removing stopwords (the offical ones from google-s dataset, also all of the words were written in both Cyrillic and clipped Cyrillic)

# Vectorization and Prediction Models

**We tried Logistic Regression, Perception, SVM and Naïve Bayes with both TF-IDF and Bag of Words. TF-IDF gave better results for vectorization, and SVM gave best results as a classifier. On the pictures bellow are micro F1 results for all models with both vectorizers (first row Bag of Words, second row TF-IDF).**

**Text data is vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. TF-IDF is a numerical statistic used in information retrieval and text mining to determine the importance of a word in a document relative to a collection of documents. It's calculated by multiplying the term frequency (TF), representing how often a term occurs in a document, by the inverse document frequency (IDF), which measures how unique or rare a term is across the entire document collection, helping to emphasize words that are both frequent within a document and rare across documents.**

**A Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel is employed for prediction. SVM is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points belonging to different classes in a high-dimensional space, maximizing the margin between classes while minimizing classification errors.**

**Parameters that gave best result:**

* C = 1.0 (Regularization parameter)
* RBF kernel

# f1 score and Results Analysis

Results for SVM and TF-IDF on training data were 0.559, and on testing 0.774.

F1 microscore is a composite metric that considers precision and recall across all classes, weighted by the number of true positives, false positives, and false negatives. It's particularly useful in evaluating classification models with imbalanced class distributions, offering a balanced assessment of overall performance.

# References

* <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html>
* <https://en.wikipedia.org/wiki/Support_vector_machine>
* <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>

Note: ChatGPT helped format some of the sentances above, we chose everything we wanted to say, he just made it sound better. ☺