

Shrey Anand, Ansul Atriek, Andrew Emerson, Amanul Haque,
Suvodeep Majumder, Vidhisha Jaswani

1. Introduction

- What makes a song a hit?
- Predict hit or not using the **acoustic properties** of the song.
- Data features:** energy, key, loudness, tempo, album, artist.

2. Data Description

1. **Data Collection:** 3 Phases with 10,000 instances and **Spotify API** to collect features.

2. **Data Pre-processing:**
Normalization, Imputation, removing duplicate data and feature selection.

3. Technical Section

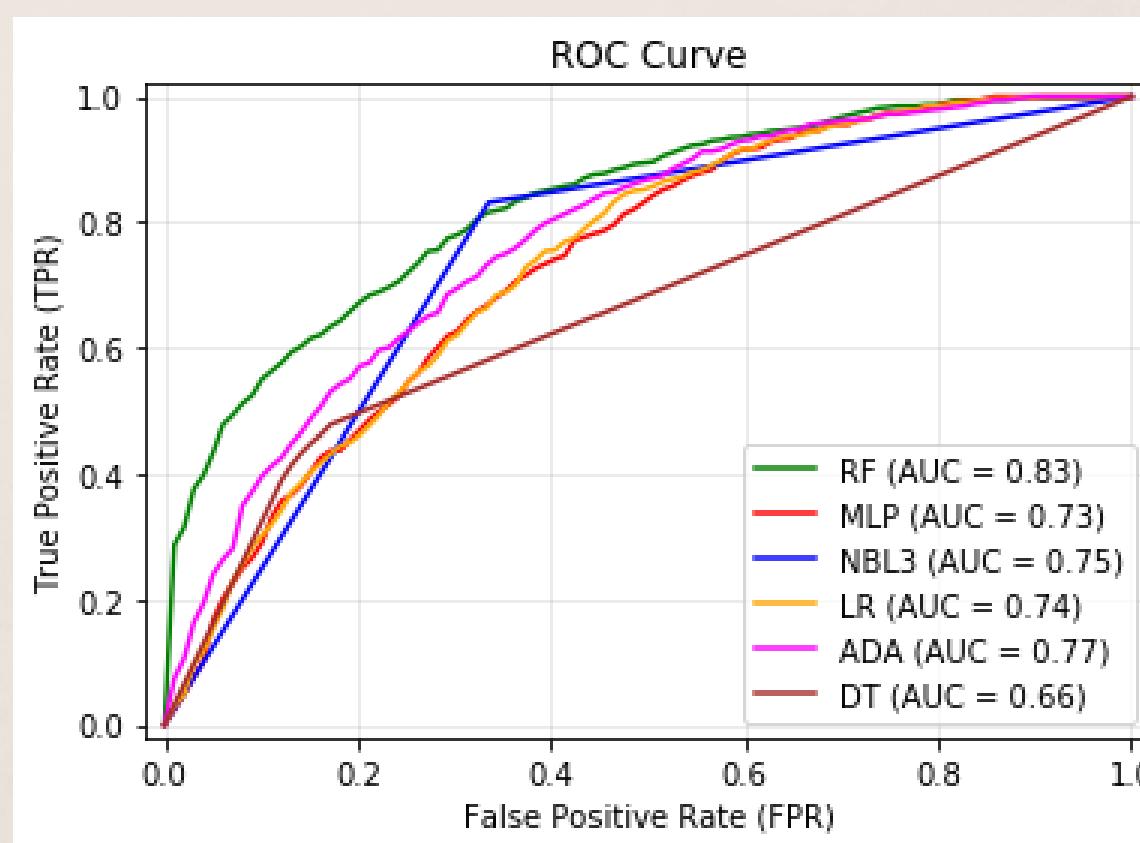
The algorithm from built-in libraries are:

- SVM
- Naive Bayes
- Logistic Regression
- Decision Trees
- Random Forests

Implemented the following algorithms from scratch:

- Neural Network
- Naive Bayes
- Multilayer tree based Naive Bayes using error correction.

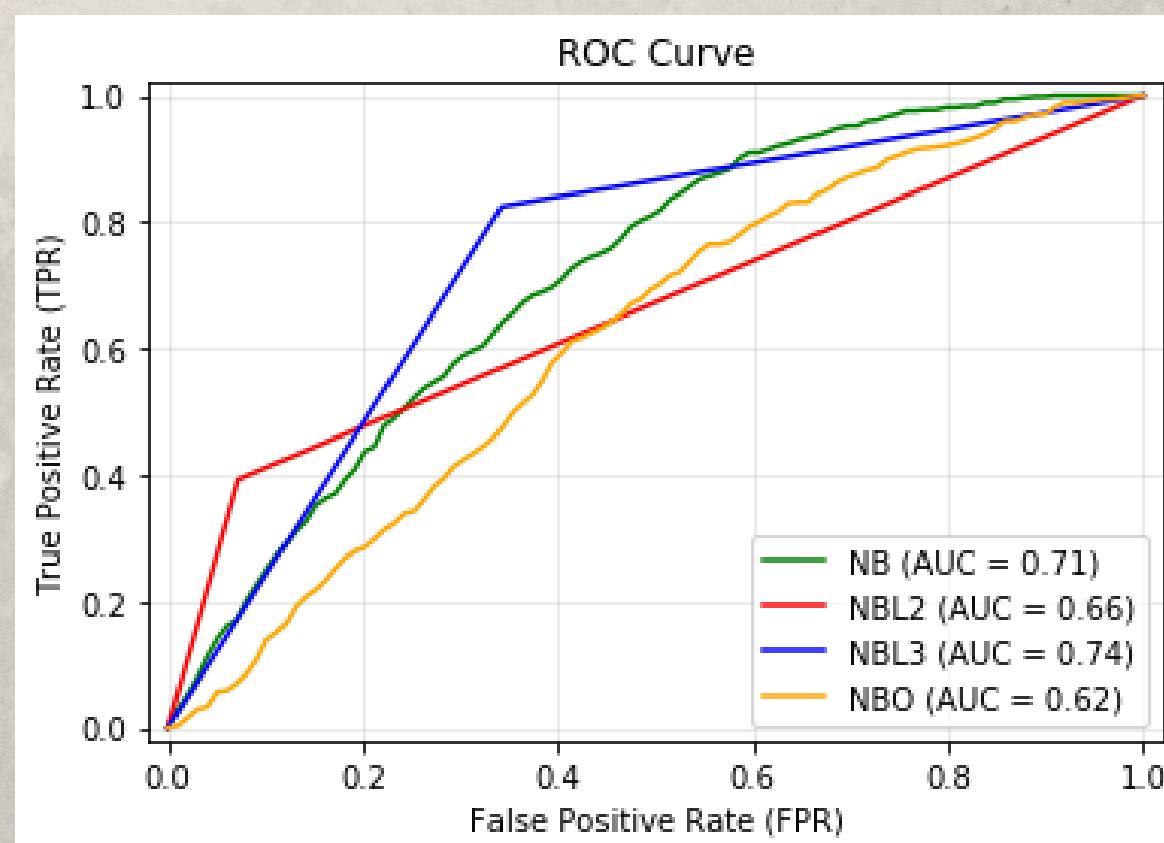
4. Results



Learner	Accuracy	F-measure	AUC
Naïve Bayes	0.697	0.689	0.691
Naïve Bayes ^{**}	0.575	0.621	0.617
2-Layer Naïve Bayes [*]	0.822	0.800	0.622
3-Layer Naïve Bayes [*]	0.693	0.727	0.744
Neural Network ⁺	0.814	0.735	0.734
Neural Network ^{**+}	0.802	0.770	0.606
Decision Tree	0.767	0.770	0.628
SVM	0.816	0.733	0.552
Logistic Regression	0.816	0.733	0.707
AdaBoost	0.814	0.752	0.714
Random Forest ⁺	0.860	0.831	0.835

*Implementation from scratch

⁺Implemented without Feature Selection.



5. Parameter Choices

Naive Bayes

- priors: prior probabilities of the classes.

Neural Network

- hidden_layer_sizes: ith element represents the number of neurons in ith hidden layer.
- activation: activation function for the hidden layer.
- learning_rate: learning rate schedule for the weight updates.
- max_iter: maximum number of iterations.

6. Conclusions

- The **feature selection** only marginally affected our results.
- Logistic regression** produces the best results.
- SVM** performs the worst of all classifiers as displayed by the low AUC metric.
- Decision Trees** work satisfactorily.
- This project has helped with our understanding of data collection, pre-processing, data visualization, and the technical details of implementing various algorithms - a complete data mining experience.