Non-Probabilistic Classification Algorithms Performance Study

Fontecha del Ser, Álvaro

Course: **Machine Learning**

Universidad Politécnica de Madrid
College of Science and Mathematics
Western Mindanao State University

POLITÉCNICA

Presentation Outline

Introduction

Spotify's Track Dataset [3] has been selected as the object of this study. It contains several parameters for 114000 tracks with 20 variables.

Table 1: Example Datapoint

index	Unnamed: 0	track_id	artists	album_name	track_name
0	0	5SuOikwiRyPMVoIQDJUgSV	Gen Hoshino	Comedy	Comedy
popularity	duration_ms	explicit	danceability	energy	key
73	230666	0	0.676	0.461	1
loudness	mode	speechiness	acousticness	instrumentalness	liveness
-6.746	0	0.143	0.0322	1.01e-06	0.358
valence	tempo	time_signature	track_genre		
0.715	87.917	4	acoustic		

Our goal is to train a prediction algorithm that classifies the popularity estimator between popular or not. For details about the methods and models employed consult [1].

Distribution of Values

Preprocessing includes discarding irrelevant and text variables, treating null values and encoding popularity in a binary variable.

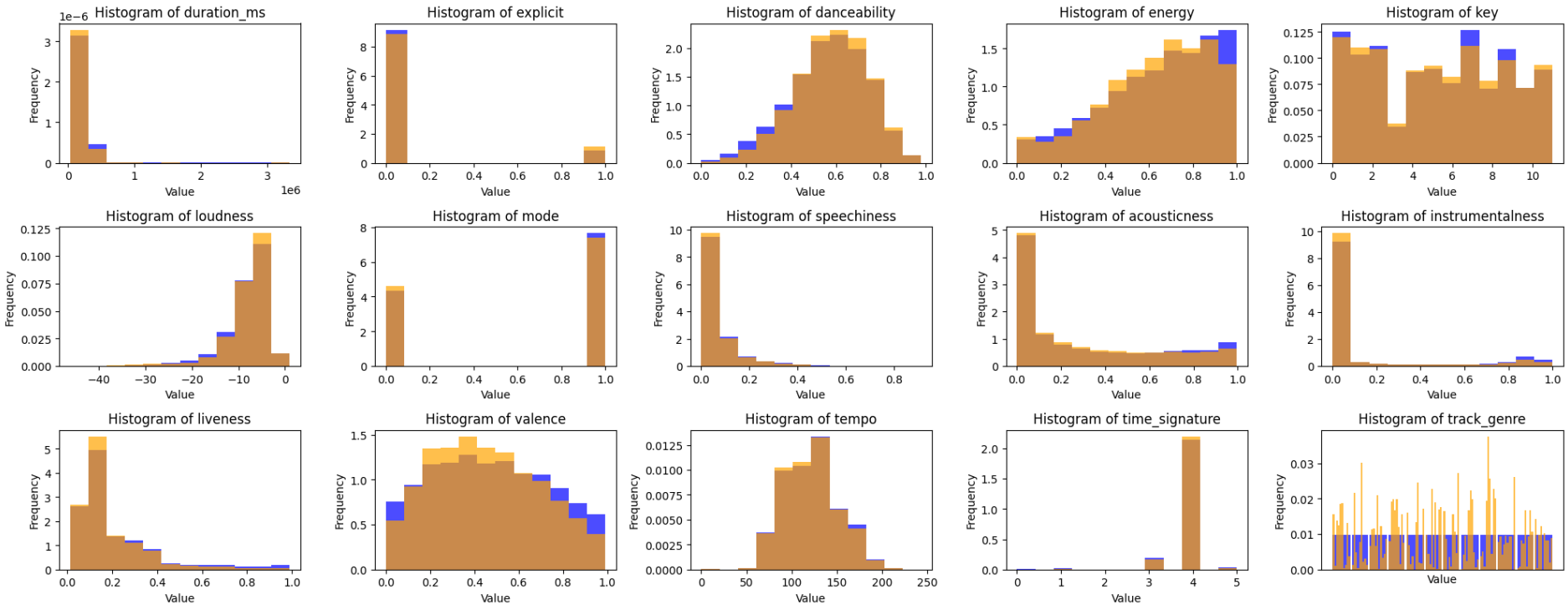

Figure 1: Distribution of variables across all data points (blue) and within the positive class instances (orange)

Correlation coefficient between variables

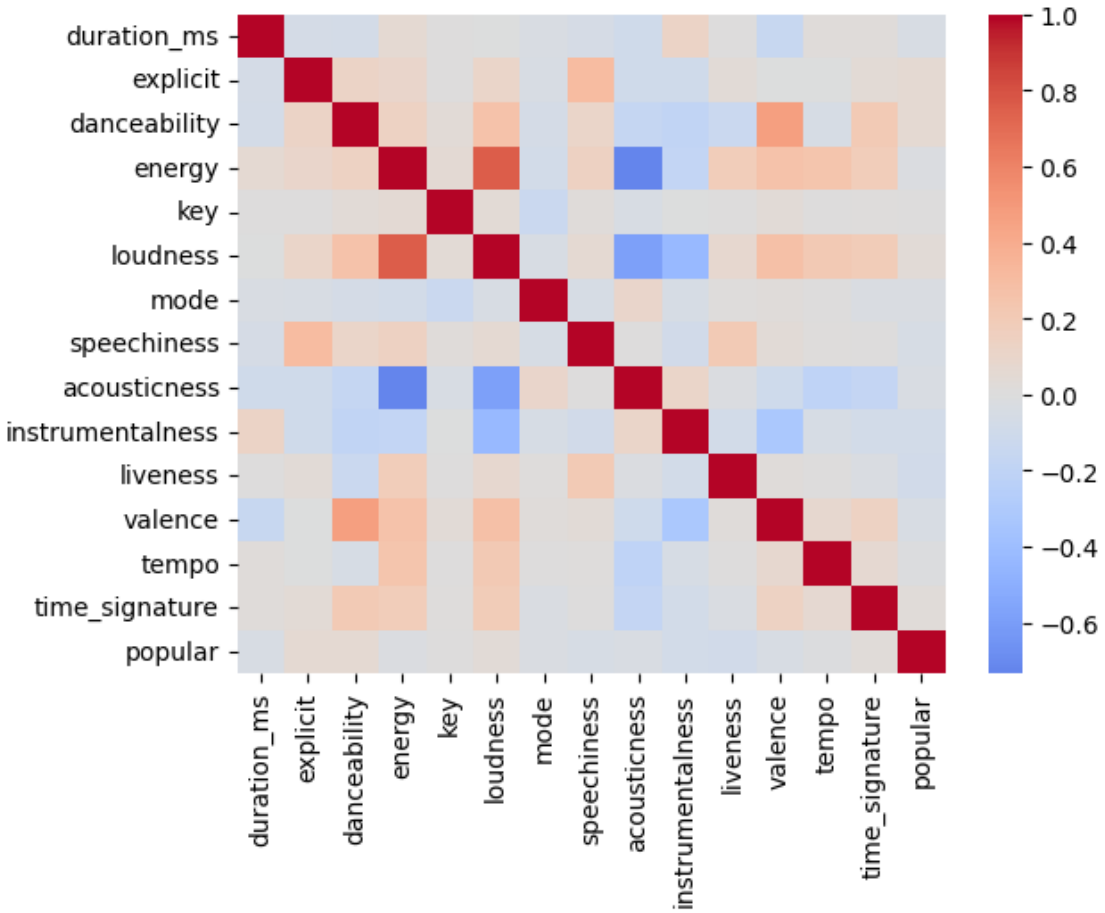

Figure 2: Correlation Heatmap

Methodology

- The **software** employed has been Python for data manipulation and Weka [2] for model training and evaluation
- The **metric** chosen to evaluate the models has been F_1 measure as recall was considered to be as important as accuracy.
- A **10-fold cross-validation** algorithm was selected to estimate the performance due to the size of the dataset.
- The algorithms that will be employed are: **IBk** ($k = 10$); **RIPPER**; **Multilayer Perceptron** (3 hidden layers); **SVM**(Polynomial kernel) and **C4.5**.
- For FSS we have employed **Gain Ratio Evaluation**, **CFS** and **Wrapper** approaches.

Feature Selection Subsets

Table 2: Variable Subsets

	track_genre	explicit	liveness	instrumentalness	duration_ms	valence	danceability	energy	acousticness	speechiness	loudness	time_signature	tempo	mode	key
Gain Ratio	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
CFS	✓														
Wrapper (K-NN)	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓		
Wrapper (RIPPER)	✓	✓	✓	✓	✓			✓	✓	✓					
Wrapper (MLP)	✓	✓	✓	✓	✓			✓		✓					
Wrapper (SVM)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
Wrapper (C4.5)	✓		✓	✓		✓	✓	✓	✓	✓			✓		

Performance Estimations

Table 3: F_1 Measure across models and Feature Subsets

	Original Variables	Univariate FSS	Multivariate FSS	Wrapper FSS
IBk	0.781	0.781	0.753	0.754
RIPPER	0.672	0.673	0.751	0.775
MLP	0.754	0.760	0.760	0.783
SVM	0.753	0.753	0.753	0.753
C4.5	0.782	0.785	0.753	0.787

Conclusions

- FSS works to reduce the dimensionality of data and complexity while keeping accuracy.
- In some cases CFS may work to reduce dimensionality too much, not yielding an improvement in accuracy. In this case this is due to a highly correlated class: the track genre.
- In Wrapper approaches a sustantial improvement to accuracy may be also achieved even if it requires more time.
- In the case of K-NN the reduction of dimensionality through FSS is actually detrimental to accuracy.

Conclusions (continued)

The algorithm that consistently outperforms the others is C4.8. With it F_1 measures of up to 0.787 may be achieved. This may be to a number of factors:

- Presence of categorical variables that model conceptual categories, **not numerical values**.
- Dataset may be **subclassified** within genres.
- Ease of understanding for extrapolating conclusions.

In conclusion, in our study we have attained values of 0.784 accuracy and 0.799 recall, correctly classifying the popularity of 78.4% of songs and correctly identifying 79.9% of popular ones.

List of References

[1] C. Bielza and P. Larrañaga. *Data-Driven Computational Neuroscience: Machine Learning and Statistical Models*. Cambridge University Press, 2021.

[2] Eibe Frank et al. *Weka: Data Mining Software in Java*. https://www.cs.waikato.ac.nz/ml/weka/. Version 3.9.4. 2016.

[3] Maharshi Pandya. *Spotify Tracks Dataset*. https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset. 2023.