Exploratory Data Analysis: Real vs Automatically Generated Irish Folk Tunes

Shpat Cheliku

Research problem

Research Questions

- Use features to discover **differences** and **commonalities** between the original songs from *The Session* and the computer-generated songs with *Folk-RNN* and use those features to do classification.

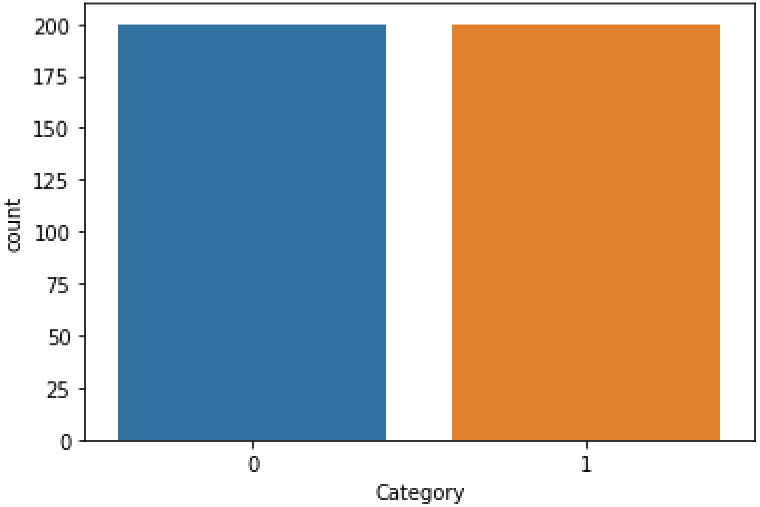
- Which features can tell us about the **artificialness** of the generated tunes?

Data

Dataset

- 200 real tunes from **The Session**
- 200 generated tunes from **Folk-RNN**

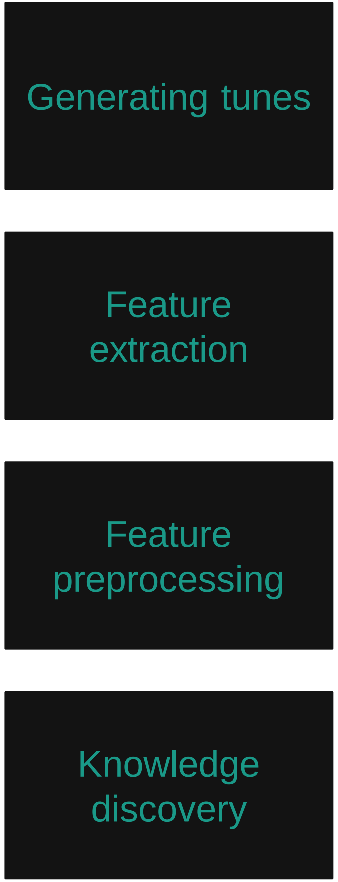
- Generated = 0
- Real = 1



Methods

Methods

- **Folk-RNN** (already trained recurrent neural network)
  - 3 LSTM layers with 512 units
  - Using mini-batch size of 50 and dropout of 0.5
- **jSymbolic**
  - Feature extraction (**1495** features)
- **Preprocessing**
  - Drop columns with **standard deviation = 0**
  - Feature scaling for classification purposes (MinMax Scaler)
- **1495** features reduced to **10-30** features



Generating tunes

Feature extraction

Feature preprocessing

Knowledge discovery

Results

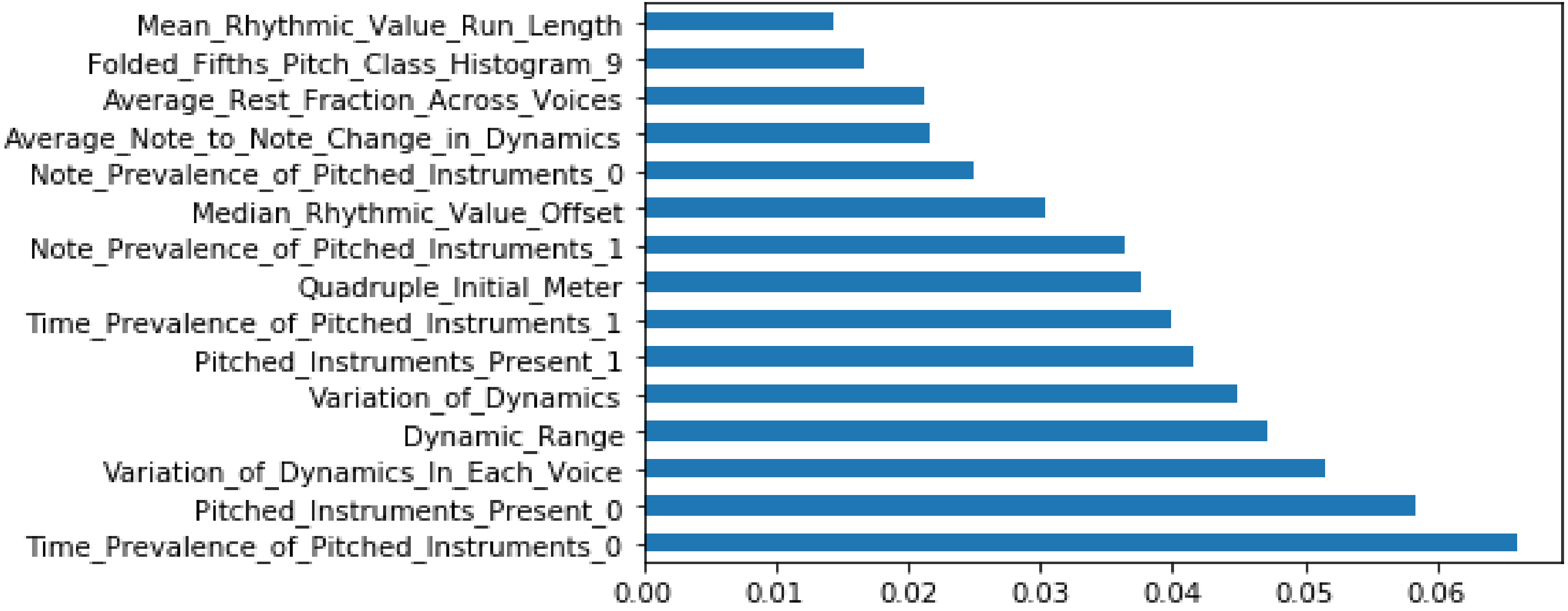
jSymbolic Feature Definitions

- Number of Pitches
  - Number of unique pitches that occur at least once in the piece. Enharmonic equivalents are grouped together for the purpose of this calculation.
- Range
  - Difference in semitones between the highest and lowest pitches.

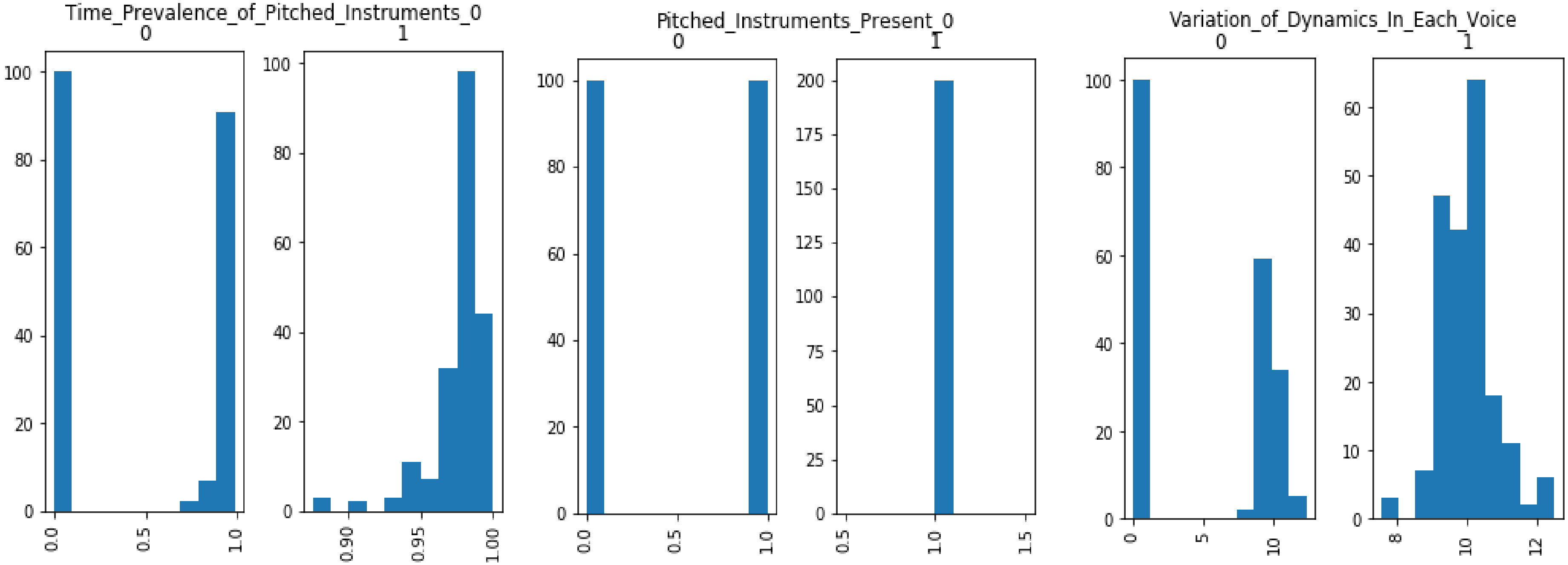
Feature importance

- We were left with **683** features after preprocessing.
  - The other **811** features didn’t give any information about the tunes.
- Techniques for feature selection/importance:
  - **Extra Trees Classifier** for Feature Selection
  - Feature Selection based on **Pearson Correlation**
  - **Univariate Feature Selection** using Chi-Squared statistical test

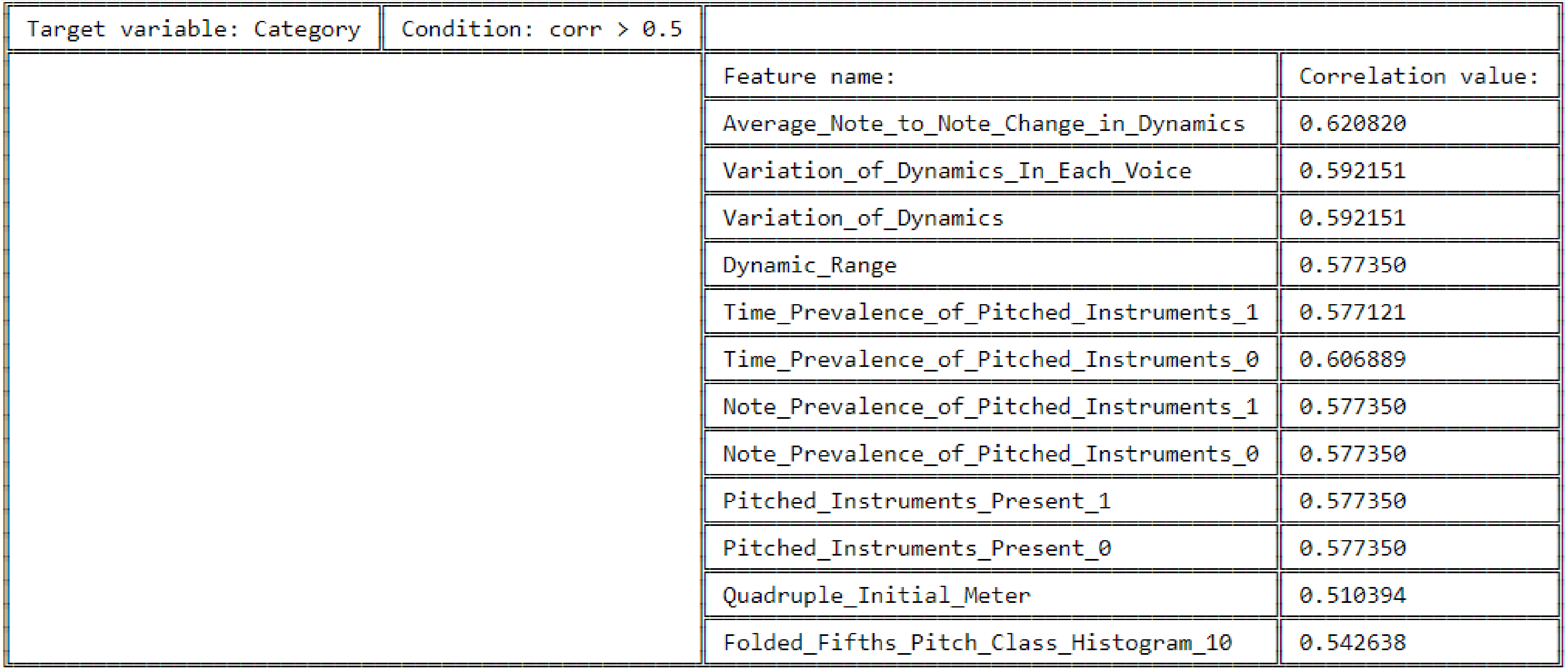
Extra Trees Classifier - Top 15 Features



Extra Trees Classifier - Top 3 Features



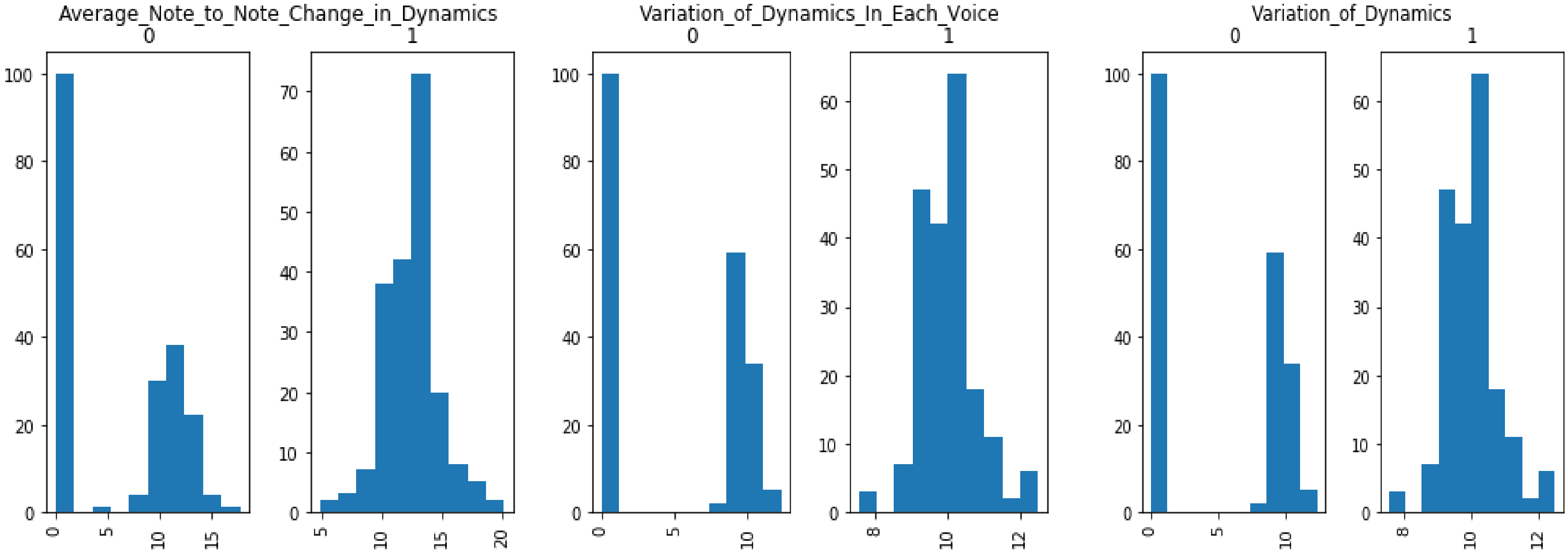
Pearson Correlation - Features with Corr > 0.5



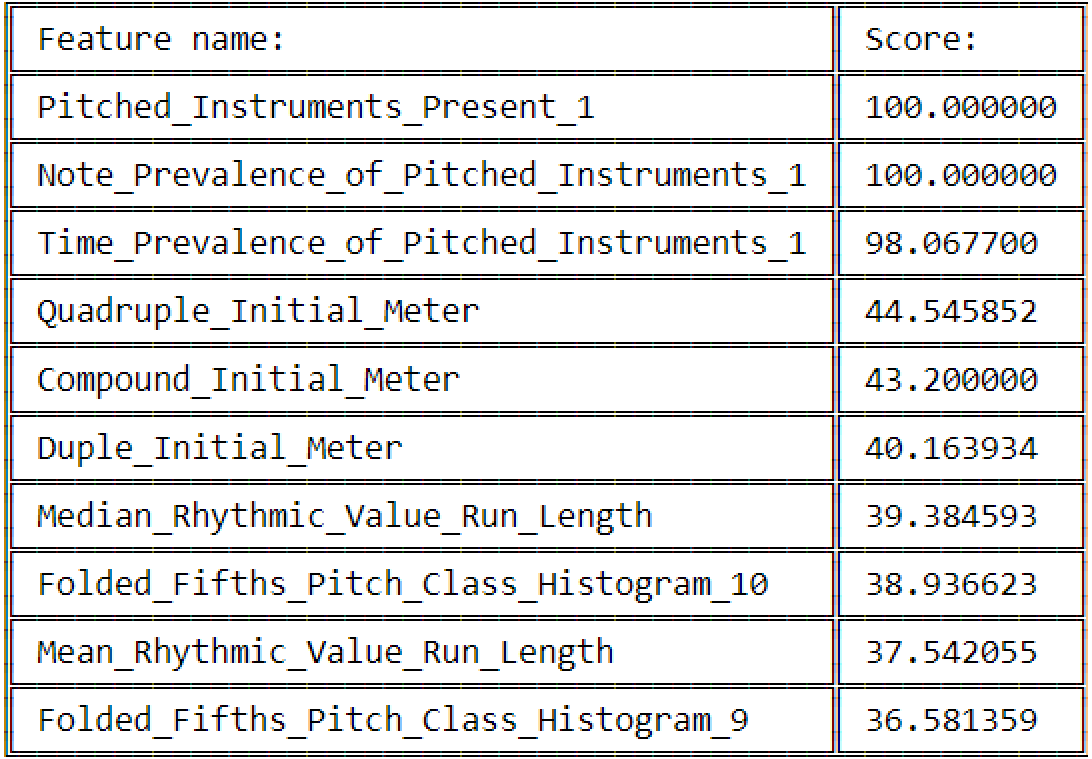
Target variable: Category	Condition: corr > 0.5

Feature name:	Correlation value:
Average_Note_to_Note_Change_in_Dynamics	0.620820
Variation_of_Dynamics_In_Each_Voice	0.592151
Variation_of_Dynamics	0.592151
Dynamic_Range	0.577350
Time_Prevalence_of_Pitched_Instruments_1	0.577121
Time_Prevalence_of_Pitched_Instruments_0	0.606889
Note_Prevalence_of_Pitched_Instruments_1	0.577350
Note_Prevalence_of_Pitched_Instruments_0	0.577350
Pitched_Instruments_Present_1	0.577350
Pitched_Instruments_Present_0	0.577350
Quadruple_Initial_Meter	0.510394
Folded_Fifths_Pitch_Class_Histogram_10	0.542638

Pearson Correlation - Top 3 Features

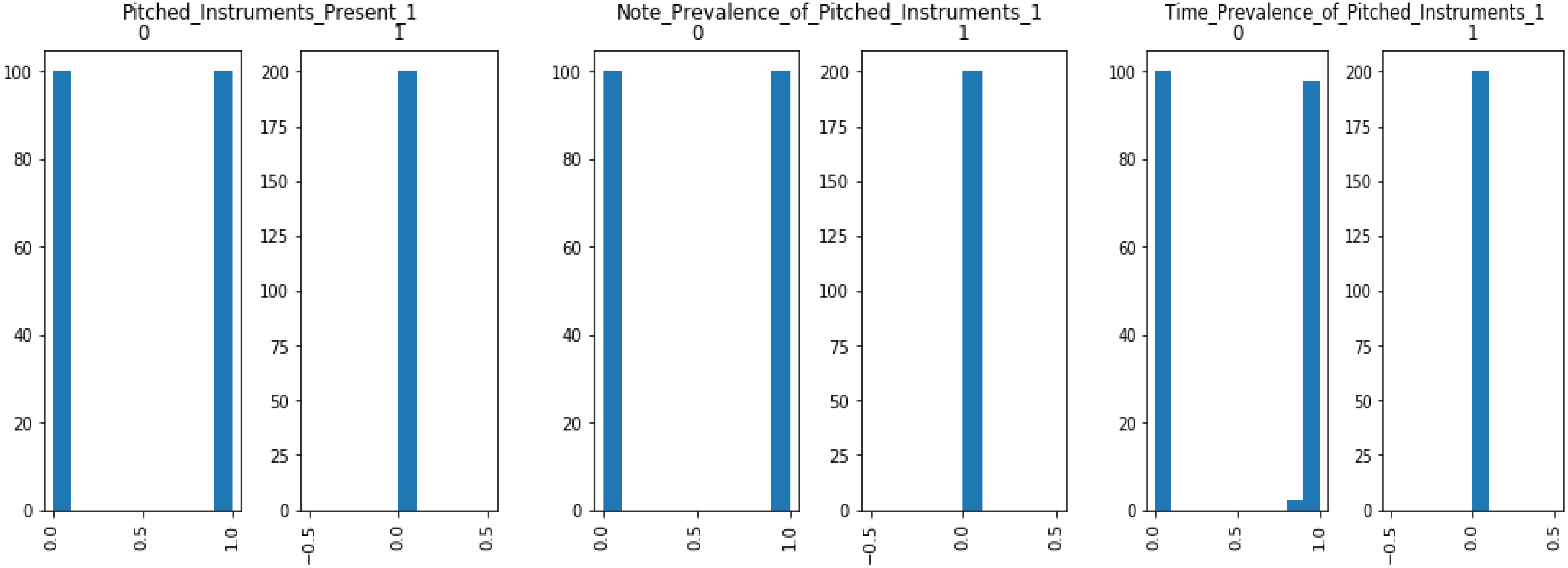


Univariate Feature Selection - Top 10 Features



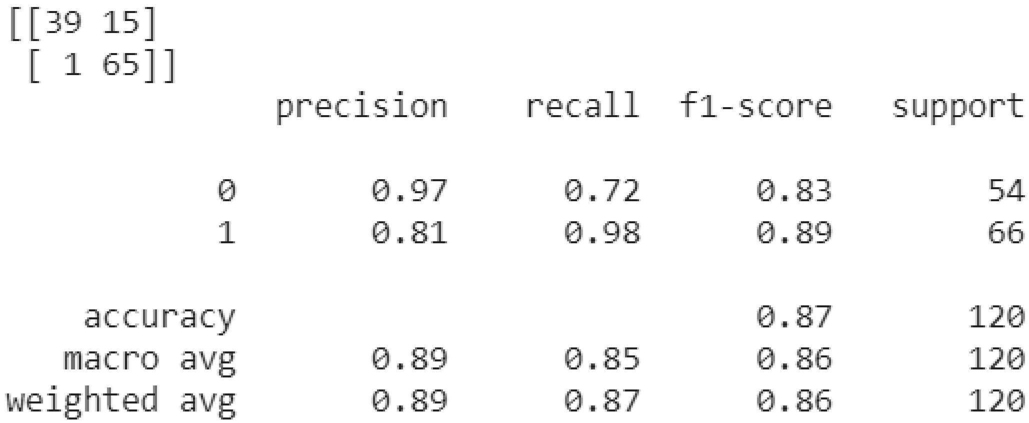
Feature name:	Score:
Pitched_Instruments_Present_1	100.000000
Note_Prevalence_of_Pitched_Instruments_1	100.000000
Time_Prevalence_of_Pitched_Instruments_1	98.067700
Quadruple_Initial_Meter	44.545852
Compound_Initial_Meter	43.200000
Duple_Initial_Meter	40.163934
Median_Rhythmic_Value_Run_Length	39.384593
Folded_Fifths_Pitch_Class_Histogram_10	38.936623
Mean_Rhythmic_Value_Run_Length	37.542055
Folded_Fifths_Pitch_Class_Histogram_9	36.581359

Univariate Feature Selection - Top 3



Classification - Using all 683 Features

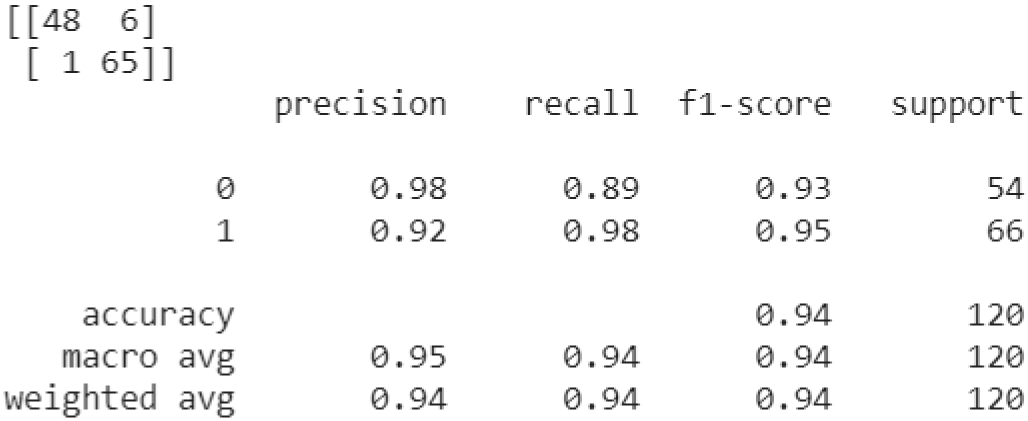
- Classifier: **SVM**
- **Train** set = 280 tunes
- **Test** set = 120 tunes



```
[[39 15]
 [ 1 65]]
              precision    recall  f1-score   support

           0       0.97      0.72      0.83        54
           1       0.81      0.98      0.89        66

    accuracy                           0.87       120
   macro avg       0.89      0.85      0.86       120
weighted avg       0.89      0.87      0.86       120
```

Classification - Using Top 15 Features of ETC

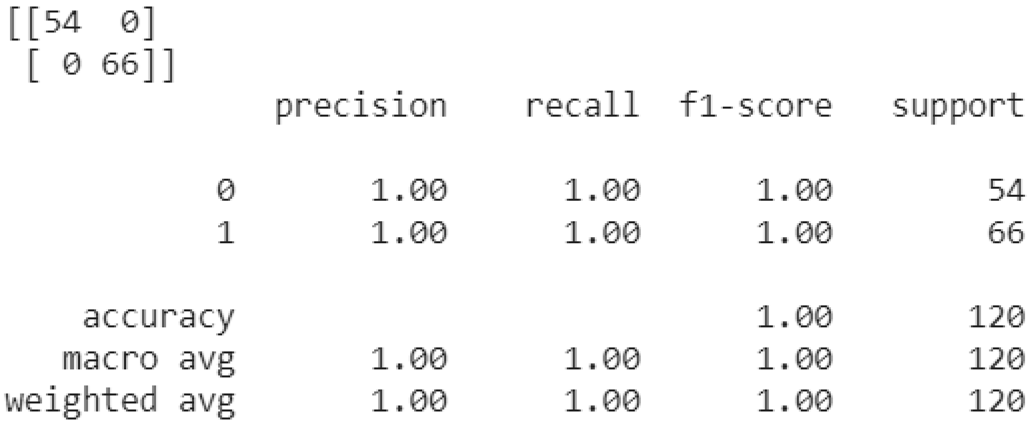
- Classifier: **SVM**
- **Train** set = 280 tunes
- **Test** set = 120 tunes



```
[[48  6]
 [ 1 65]]
              precision    recall  f1-score   support

           0       0.98      0.89      0.93        54
           1       0.92      0.98      0.95        66

    accuracy                           0.94       120
   macro avg       0.95      0.94      0.94       120
weighted avg       0.94      0.94      0.94       120
```

Classification - Using 12 Features of PC

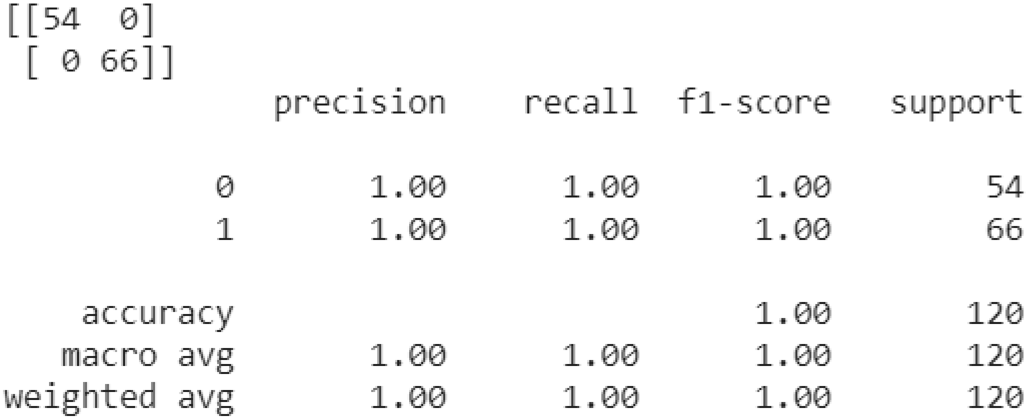
- Classifier: **SVM**
- **Train** set = 280 tunes
- **Test** set = 120 tunes



```
[[54  0]
 [ 0 66]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        54
           1       1.00      1.00      1.00        66

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120
```

Classification - Using Top 10 Features of UFS

- Classifier: **SVM**
- **Train** set = 280 tunes
- **Test** set = 120 tunes

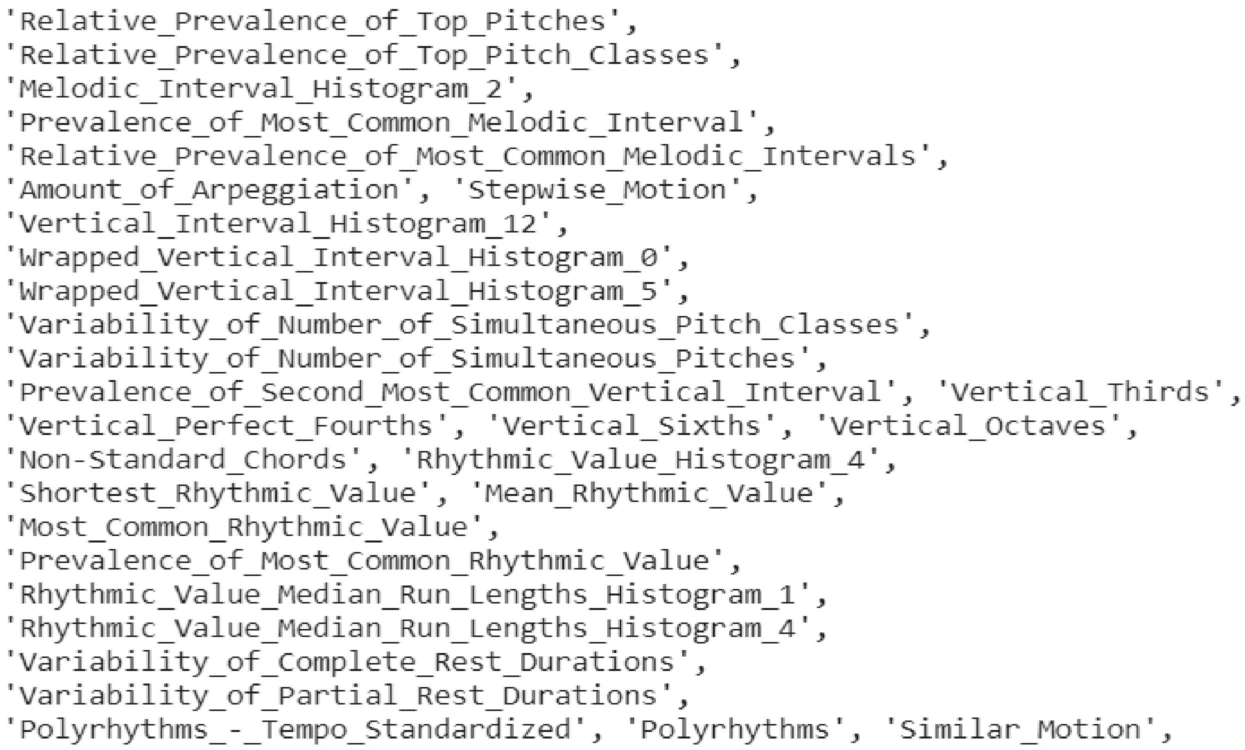


```
[[54  0]
 [ 0 66]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        54
           1       1.00      1.00      1.00        66

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120
```

Common Features between Real and Generated

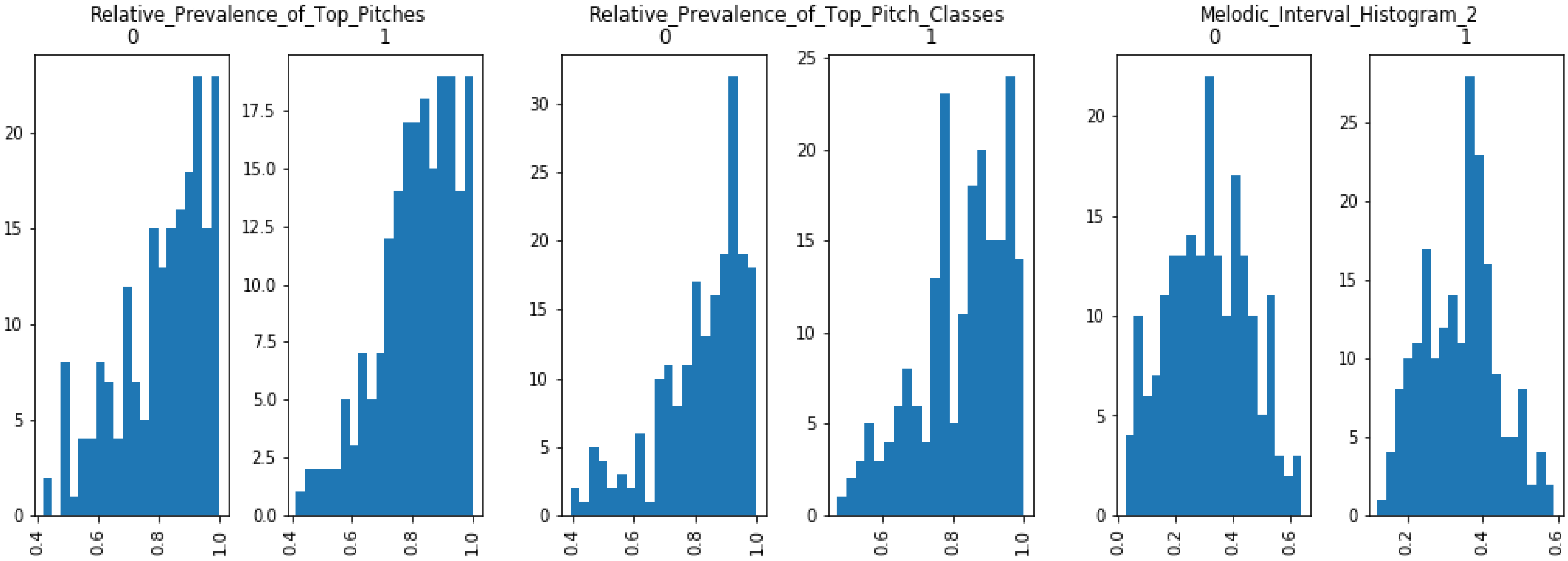
- Used the combined dataset of 400 songs (real + generated)
- Analysed and selected the features that have a **very low standard deviation**
- The resulting set was **30** common features

Common Features between Real and Generated



```
'Relative_Prevalence_of_Top_Pitches',
'Relative_Prevalence_of_Top_Pitch_Classes',
'Melodic_Interval_Histogram_2',
'Prevalence_of_Most_Common_Melodic_Interval',
'Relative_Prevalence_of_Most_Common_Melodic_Intervals',
'Amount_of_Arpeggiation', 'Stepwise_Motion',
'Vertical_Interval_Histogram_12',
'Wrapped_Vertical_Interval_Histogram_0',
'Wrapped_Vertical_Interval_Histogram_5',
'Variability_of_Number_of_Simultaneous_Pitch_Classes',
'Variability_of_Number_of_Simultaneous_Pitches',
'Prevalence_of_Second_Most_Common_Vertical_Interval', 'Vertical_Thirds',
'Vertical_Perfect_Fourths', 'Vertical_Sixths', 'Vertical_Octaves',
'Non-Standard_Chords', 'Rhythmic_Value_Histogram_4',
'Shortest_Rhythmic_Value', 'Mean_Rhythmic_Value',
'Most_Common_Rhythmic_Value',
'Prevalence_of_Most_Common_Rhythmic_Value',
'Rhythmic_Value_Median_Run_Lengths_Histogram_1',
'Rhythmic_Value_Median_Run_Lengths_Histogram_4',
'Variability_of_Complete_Rest_Durations',
'Variability_of_Partial_Rest_Durations',
'Polyrhythms_-_Tempo_Standardized', 'Polyrhythms', 'Similar_Motion',
```

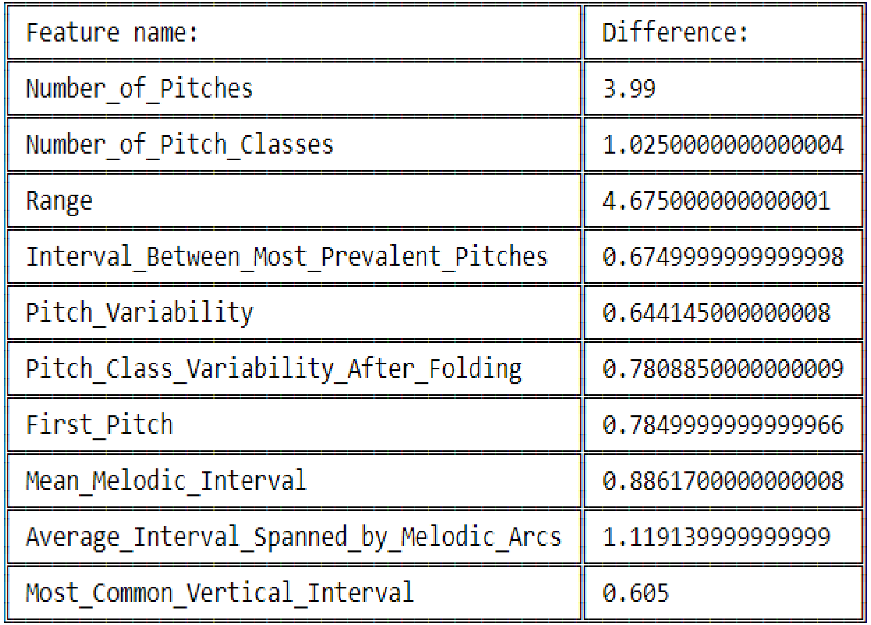
Common Features between Real and Generated



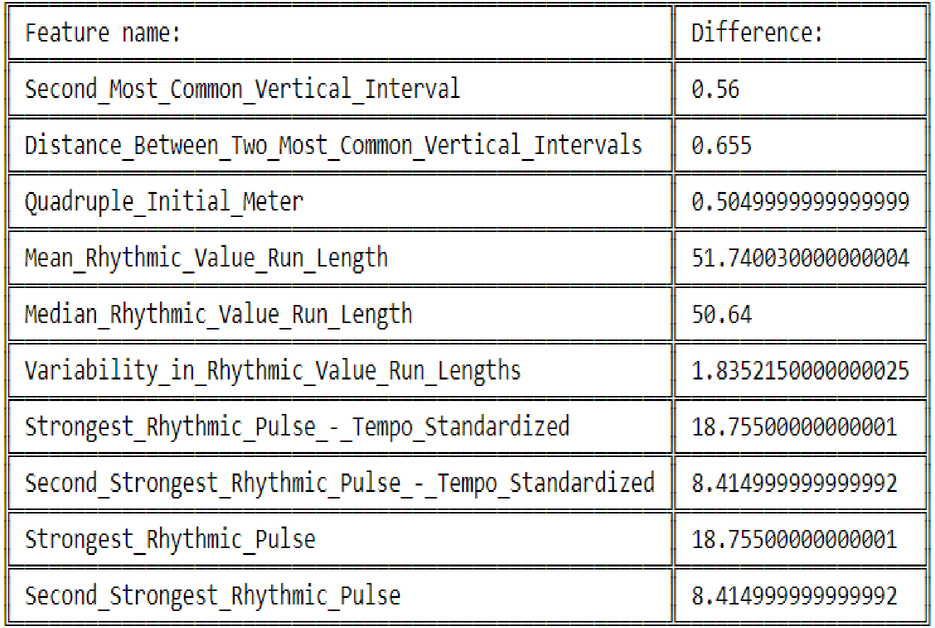
Artificialness of the Generated Tunes

- We calculated the **means** of all relevant **features** for both **Real and Generated** tunes.
- Then we calculated the **differences** between those **means**
  - difference = generated[x].mean() - real[x].mean()
- We took the features with **difference > 0.5** as features indicating **artificialness** of tunes.
- Reason: **Folk-RNN** tune generator **focused more** on those features than a real person does when creating real tunes.

Artificialness of the Generated Tunes - 20 Features



Feature name:	Difference:
Number_of_Pitches	3.99
Number_of_Pitch_Classes	1.0250000000000004
Range	4.675000000000001
Interval_Between_Most_Prevalent_Pitches	0.6749999999999998
Pitch_Variability	0.644145000000008
Pitch_Class_Variability_After_Folding	0.7808850000000009
First_Pitch	0.7849999999999966
Mean_Melodic_Interval	0.8861700000000008
Average_Interval_Spanned_by_Melodic_Arcs	1.119139999999999
Most_Common_Vertical_Interval	0.605



Feature name:	Difference:
Second_Most_Common_Vertical_Interval	0.56
Distance_Between_Two_Most_Common_Vertical_Intervals	0.655
Quadruple_Initial_Meter	0.5049999999999999
Mean_Rhythmic_Value_Run_Length	51.740030000000004
Median_Rhythmic_Value_Run_Length	50.64
Variability_in_Rhythmic_Value_Run_Lengths	1.8352150000000025
Strongest_Rhythmic_Pulse_-_Tempo_Standardized	18.75500000000001
Second_Strongest_Rhythmic_Pulse_-_Tempo_Standardized	8.414999999999992
Strongest_Rhythmic_Pulse	18.75500000000001
Second_Strongest_Rhythmic_Pulse	8.414999999999992

Future work

Future Work

- Team up with a more **musically inclined** individual to gather more insight.
- **Analyse** the **artificialness** of the features in a more detailed way and from other points of view.
- Writing the obtained results and answering the research questions in the project report.

Thank you!