Music Genre Classification Project with PCA and Logistic Regression

A Machine Learning Project Predicting Music Styles through Dimensionality Reduction

Case Description

Objective

In the era of digital streaming, there's an increasing need to categorize and recommend music based on genres. By analyzing various musical features extracted from tracks, we can delve deeper into their defining patterns. In this music genre classification project, you'll work with a dataset containing various musical features extracted from tracks across different styles. Please note that while this music genre classification dataset is extensive, it is incomplete. A significant portion of the records lacks specific genre information. Your primary task is to predict the genres of these unlabeled tracks. To accomplish this, you'll employ Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. By transforming the abundant features into principal components, you'll streamline the data, making it more manageable and revealing patterns that are not immediately obvious in the raw data. The principal components you've derived will form the foundation for the next step in the project—employing a supervised machine learning algorithm, with a focus on the well-known logistic regression technique.

Why Principal Component Analysis (PCA)?

Music tracks are complex entities with numerous inherent features. Some of these features might be correlated. For instance, specific rhythm patterns might be prevalent in rock and blues. In this machine learning project, PCA can assist in reducing redundancy by transforming correlated musical features into a set of linearly uncorrelated variables or principal components. Reducing dimensionality can drastically improve the performance of classification algorithms by eliminating noise. The features in the music genre classification dataset are designed to be intuitive and accessible allowing you to focus on the core concepts of PCA and machine learning without the need for specialized audio knowledge.

Prepare to dive deep into the layers of musical data and discover the patterns that help outline the musical genres.

Project requirements

You'll need Python 3 or a newer version and can choose any IDE (Jupyter Notebook, Spyder, PyCharm, Visual Studio, etc.). You’ll also need to have the following Python libraries installed:

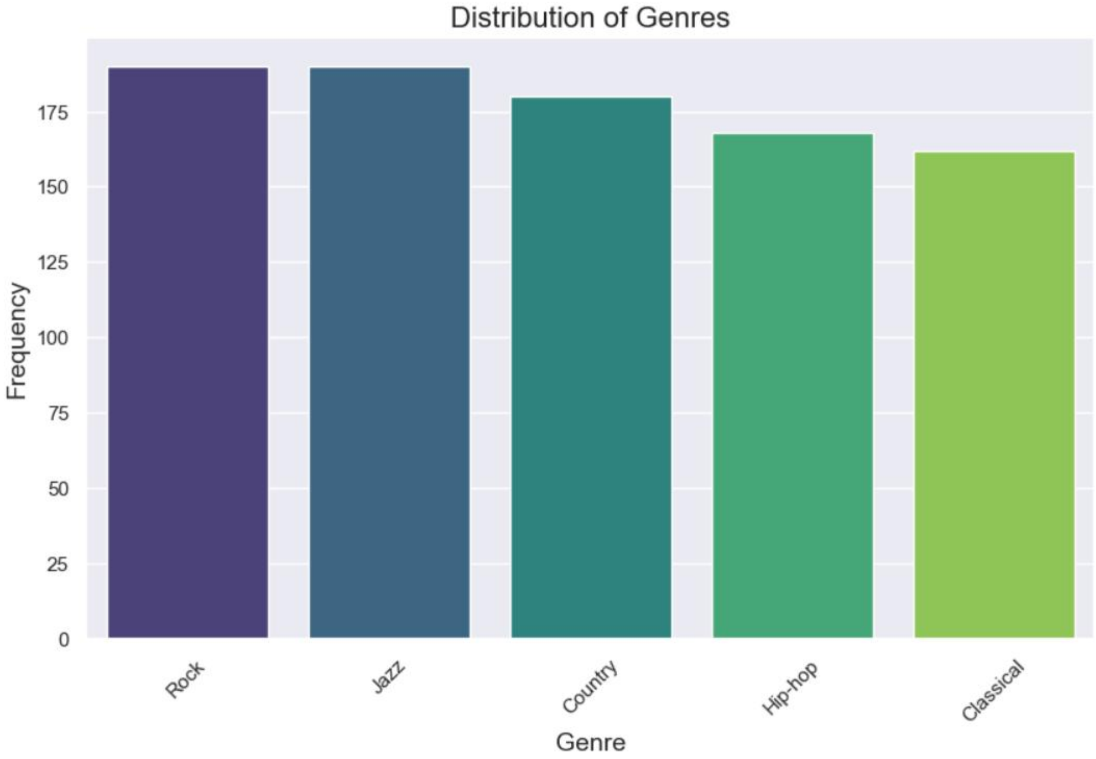
- Pandas
- NumPy
- Matplotlib
- Scipy
- Scikit-learn
- Seaborn

Project files

The music data for the Music Genre Classification project can be found in the music dataset mod.csv file, and the data legend is provided in the Music Data Legend.xlsx.

Part 1: Data Exploration

Data exploration is a foundational step in any machine learning project. This phase focuses on understanding the structure, distribution, and completeness of the dataset. The dataset used in this project included multiple features corresponding to musical attributes and a Genre column representing the track's genre. Roughly 11% of the entries had missing values in the Genre column. Unlike many real-world datasets that suffer from class imbalance, this dataset was relatively balanced across genres. Initial exploration involved examining the structure of the dataset using Python’s pandas library and visualizing the Genre distribution with seaborn's barplot. These steps helped provide a clear understanding of the data’s composition and confirmed its readiness for machine learning applications after minor cleaning.

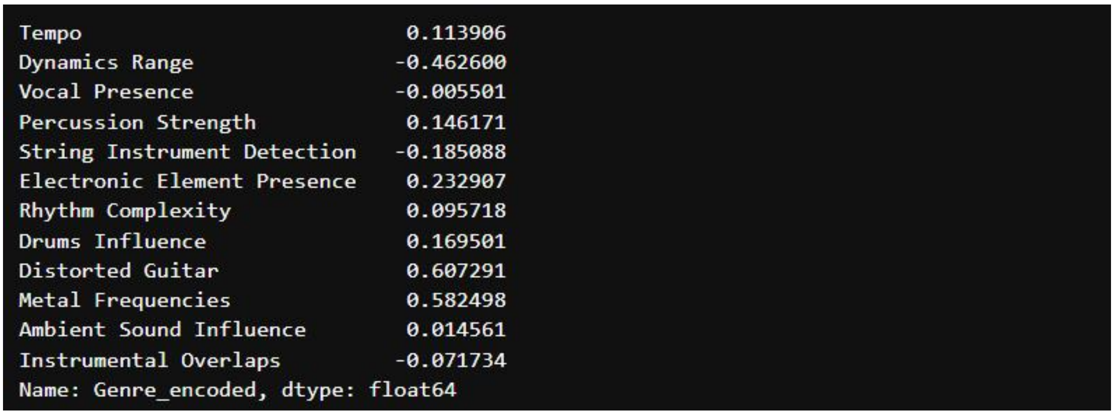


Part 2: Correlation Analysis

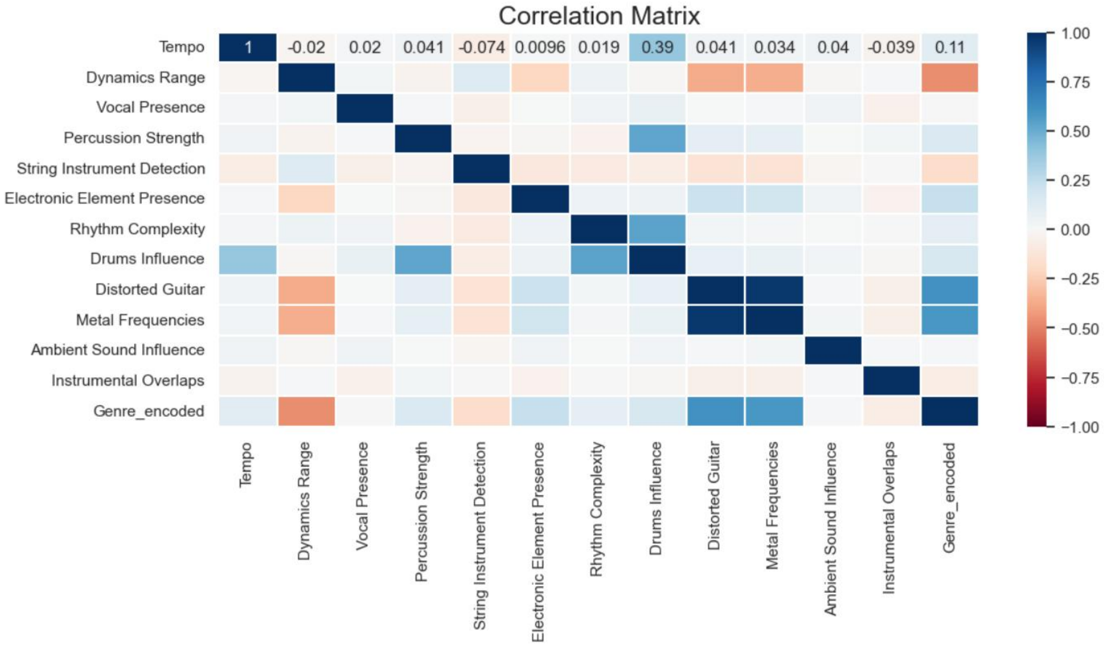
Correlation analysis helps in understanding the relationships between features and the target variable (musical genres). Below are the correlation of each predictor with the target (Genre) variable. Distorted Guitar and Metal Frequencies emerged as the most influential factors, strongly associated with heavier genres like rock and metal. This underscores their critical role in differentiating such genres from others. Dynamics Range and String Instrument Detection, on the other hand, are negatively associated with the encoded genres, suggesting a stronger link to softer styles like classical and acoustic music.

Moderate influences include features like Electronic Element Presence and Percussion Strength, which align with rhythm-heavy and electronic genres. However, attributes such as Vocal Presence and Ambient Sound Influence showed minimal impact, indicating their limited importance in predicting genres accurately.

These findings emphasize the value of specific musical attributes in classification while highlighting areas for refinement. They provide a clear understanding of the relationship between features and genres, aiding future improvements in the model.



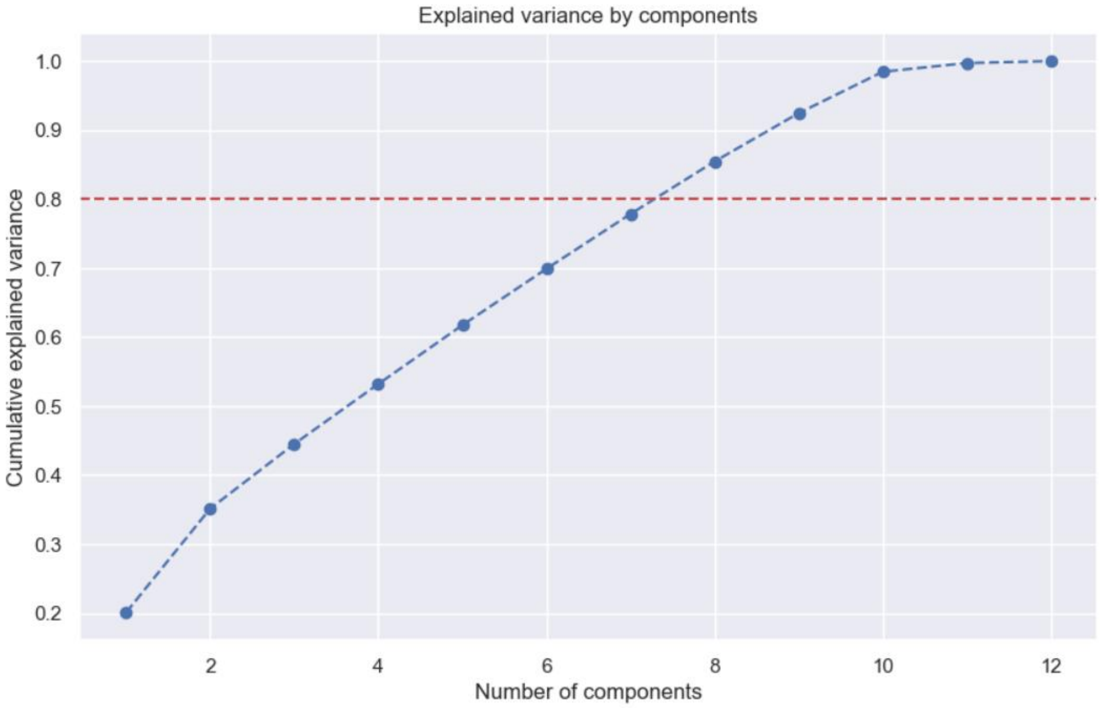
```
Tempo                          0.113906
Dynamics Range                -0.462600
Vocal Presence                -0.005501
Percussion Strength            0.146171
String Instrument Detection   -0.185088
Electronic Element Presence    0.232907
Rhythm Complexity              0.095718
Drums Influence                0.169501
Distorted Guitar               0.607291
Metal Frequencies              0.582498
Ambient Sound Influence        0.014561
Instrumental Overlaps         -0.071734
Name: Genre_encoded, dtype: float64
```


Correlation Matrix
Tempo
Dynamics Range
Vocal Presence
Percussion Strength
String Instrument Detection
Electronic Element Presence
Rhythm Complexity
Drums Influence
Distorted Guitar
Metal Frequencies
Ambient Sound Influence
Instrumental Overlaps
Genre_encoded
1
-0.02
0.02
0.041
-0.074
0.0096
0.019
0.39
0.041
0.034
0.04
-0.039
0.11
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Part 3: PCA for Dimensionality Reduction

The analysis revealed moderate correlations between some features, highlighting the potential for redundancy and multicollinearity in the dataset. To address these challenges, we employed PCA to reduce dimensionality. Before applying PCA, the features were standardized using StandardScaler, ensuring equal weightage across attributes.

The PCA analysis revealed that eight principal components could explain approximately 82% of the dataset's variance. This insight guided our decision to retain eight components for model training, striking a balance between dimensionality reduction and information retention.



Part 4: Evaluating Classification Efficacy – PCA-Transformed vs. Original Data

Two separate models were developed to classify genres: one using PCA-transformed features and another using the original features. Both models employed Logistic Regression due to its simplicity and effectiveness in classification tasks. The training and testing datasets were split 70:30 to ensure robust evaluation.

The PCA-based model achieved an accuracy of 54%, slightly outperforming the non-PCA model's 53%. While the improvement was modest, it validated the use of PCA in reducing redundancy and simplifying the model without sacrificing performance. Both models provided consistent classification reports, reflecting the balanced nature of the dataset and the suitability of Logistic Regression for this task.

Part 5: Genre Prediction and Integration

The project's core objective was to predict missing Genre labels. Rows with missing values in the Genre column were isolated and subjected to the same transformations as the training data, including scaling and PCA. The PCA-based Logistic Regression model was then used to predict these genres. Predictions, originally in numeric form, were converted back to their categorical labels using LabelEncoder.inverse_transform.

By reintegrating these predictions into the original dataset, we successfully completed the missing entries. The updated dataset is now fully labeled, enabling its use for downstream tasks like recommendation systems or exploratory analysis of genre trends.

Key Insights and Reflections

This project underscores the value of combining dimensionality reduction with machine learning. PCA proved effective in mitigating multicollinearity and reducing computational complexity while maintaining model performance. The roughly balanced classes ensured fair evaluation and minimized biases that often arise in imbalanced datasets.

Despite its success, the project faced limitations. The accuracy improvements with PCA were modest, suggesting that alternative methods like ensemble models or non-linear classifiers could further enhance performance. Additionally, while PCA reduces redundancy, it may obscure relationships between features and the target variable, limiting interpretability.

Conclusion and Recommendations

This project successfully addressed the challenge of incomplete Genre data while demonstrating the benefits of dimensionality reduction. The PCA-based model performed marginally better than its non-PCA counterpart, confirming its utility for this dataset.

Moving forward, we recommend:

1. **Expanding the Model Pipeline:** Experimenting with ensemble methods or Support Vector Machines for improved accuracy.
2. **Addressing Class Dynamics:** Continuously monitoring class distributions as new data is added to ensure balanced training.
3. **Leveraging the Updated Dataset:** Using the complete dataset for tasks like trend analysis or building music recommendation systems.

By filling in missing Genre data, this project enhances the dataset's utility and opens avenues for advanced analytics. As the music industry becomes increasingly data-driven, such projects are pivotal in bridging data gaps and enabling informed decision-making.