BANANA PROBLEM STATEMENT

I) PROBLEM STATEMENT:

Develop a model to classify songs genre into one of two categories: 'Hip-Hop' or 'Rock.' Use labeled song dataset, and perform a series of data preprocessing and analysis steps, including feature correlation analysis, data normalization, and dimensionality reduction via Principal Component Analysis (PCA). Conclude by visualizing the data to gain insights into the distribution and separation of the genres based on the extracted features.

II) INSTRUCTION & ASSUMPTIONS:

1. Dataset Quality:

- The dataset contains a balanced representation of 'Hip-Hop' and 'Rock' songs, **or** appropriate class balancing techniques will be applied.

2. Evaluation Metrics:

- Standard classification metrics (accuracy, precision, recall, F1-score) are appropriate for evaluating the model performance. These metrics will provide insights into the models' ability to correctly classify the genres.

III) DELIVERABLES:

This project provides hands-on experience in data preprocessing, feature engineering, model training, and validation. By addressing challenges such as class imbalance and overfitting, you will build robust machine learning models capable of accurately classifying 'Hip-Hop' and 'Rock' songs based on audio features. This system has the potential to significantly enhance user experiences in various music-related applications.

- Implementation Outline:
    - Clean the data
    - Exploratory data visualization of the data
    - Feature reduction (Correlation and Principal Component Analysis)
    - Use ML algorithms like logistic regression and decision tree

IV) REQUIREMENTS:

1) Technical environment:

The analysis and modeling will be performed using Python, with libraries such as scikit-learn for machine learning tasks, pandas for data manipulation, and matplotlib or seaborn for visualization.

The working environment (e.g., Jupyter Notebook) supports the necessary libraries and tools for data analysis and model training.

2) Data Collection and Preparation:

Obtain a labeled dataset containing audio features for 'Hip-Hop' and 'Rock' songs. Ensure that the dataset includes sufficient examples for each genre to facilitate model training and evaluation.

NOTE: Audio data of almost 5000 different songs are used. Data needed is present in two files, one CSV file which contains basic information about the track along with the genre and a JSON file containing muscial features like danceability and acousticness. This data was compiled by The Echo Nest. Load the dataset into your working environment (e.g., Jupyter Notebook).

V) WORKFLOW / METHODOLOGY:

1. DATA PREPROCESSING:

i) Feature Correlation:

- Analyze the correlation between features to understand their relationships and potential redundancy.
- Visualize the correlation matrix using a heatmap to identify highly correlated features.

ii) Normalization:

- Use scikit-learn’s `StandardScaler` to normalize the data, ensuring that all features contribute equally to the model performance.

iii) Dimensionality Reduction:

- Apply PCA to the scaled data to reduce dimensionality, highlight the most significant features.
- Visualize the data in the reduced dimensional space to observe the separation between the genres.

2. MODEL TRAINING AND VALIDATION:

- Implement techniques to handle any class imbalance in the dataset, ensuring that the models are not biased towards the majority class.

i) Logistic Regression:

- Train a Logistic Regression model to classify the songs based on the preprocessed features.
- Evaluate the model using metrics such as accuracy, precision, recall, and F1-score.

ii) Decision Tree:

- Train a Decision Tree model and compare its performance with Logistic Regression.
- Evaluate the model using metrics such as accuracy, precision, recall, and F1-score.

iii) Cross-Validation (Optional):

- Use cross-validation techniques to validate the models, reducing the risk of overfitting and ensuring generalizability to unseen data.
- Report the cross-validation results to provide a more reliable estimate of model performance.

3. EVALUATION AND VISUALIZATION:

i) Model Evaluation:

- Evaluate the models using appropriate metrics such as accuracy, precision, recall, and F1-score.
- Compare the evaluation metrics of the Logistic Regression and Decision Tree models to determine which model performs better.

ii) Visualization:

- Plot the principal components to visualize how well the PCA separates the two genres.
- Visualize the decision boundaries of the models to understand their classification behaviour.

VI) OUTCOME OF THE PROJECT

By completing this project, Students will gain practical experience in data preprocessing, feature engineering, model training, and validation. Students will also learn advanced techniques to address common machine learning challenges such as class imbalance and overfitting, providing a solid foundation for building robust classification systems.