MUSIC GENRE CLASSIFICATION

Introduction:

This project focuses on building a machine learning model to classify music genres based on extracted features from the various aspects of the sound. Using data preprocessing, feature engineering, and multiple classification algorithms, we evaluated and compared model performances to identify the most accurate approach.

1.Dataset Overview

This Dataset was taken from the Kaggle repository. Dataset contains 9990 rows of samples and 60 columns of features including audio characteristics such as *chroma_stft_mean*, *rms_mean*, *spectral_centroid_mean*, and genre labels.

2.Data preprocessing

a. Import the library such as pandas , numpy, random matplotlib, seaborn needed for data preprocessing as well as EDA.

b. Read the csv file of the dataset using pd.read_csv(“filepath.csv”).

c. Find the total number of columns,rows and also use the head() and tail() to analyze the data.

d. Use describe() to see the central tendency of the dataset.

e. Use info() to check the datatype of all columns.

f. Find the null values in each column and handle them correctly if any present (here no values were found).

g. After observing the datatypes we understand that most of the columns are of float64.

h. Make one copy of the dataset and work on it so your original data will not get lost.

i. Now using one for loop collects all non float values(except Genre column) into a list and removes that nonfloat value to improve the accuracy of our model.

j. Now Check the most common Genre in the dataset using value_counts(). The samples of each genre are about the same .

3. MODEL EVALUATION:

Classification is a type of supervised learning which involves using various classification models to assign a label based on the other features.Model Evaluation can be performed using accuracy score,precision score,recall score and F1 score

- **Accuracy Score:**The accuracy score is a commonly used metric to evaluate the performance of a classification model. It measures the proportion of correct predictions (both true positives and true negatives) out of all the predictions made by the model.

  Accuracy=(TP+TN)/(TP+TN+FP+FN)

- **Precision Score:** The Precision Score is a performance metric used to evaluate classification models, particularly in situations where the cost of false positives is important. Precision measures the accuracy of positive predictions, i.e., how many of the instances predicted as positive are actually positive.

  Precision=(TP)/(TP+FP)

- **Recall Score:** The **Recall Score** (also known as Sensitivity or True Positive Rate) is a performance metric used in classification problems, particularly when the cost of false negatives is high. Recall measures how well a model identifies all relevant positive instances in the dataset, i.e., how many actual positive instances the model correctly predicts.

  Recall=(TP)/(TP+FN)

- **F1 Score:** The F1 Score is a performance metric used in classification problems that combines both Precision and Recall into a single value. It is particularly useful when the dataset has imbalanced classes or when both false positives and false negatives are important to minimize.

  F1 Score=2(Precision*Recall)/(Precision+Recall)*

The Following are some of the classification models utilized to classify to the genre based on the other features and their performance is evaluated using accuracy score,precision score,recall score and F1 score :

1. **Gaussian Naive Bayes:**

Gaussian Naive Bayes (GNB) is a variant of the Naive Bayes classifier that is specifically used when the features of the data are continuous and are assumed to follow a Gaussian (normal) distribution. It is a probabilistic classifier based on Bayes' Theorem, and it assumes that the features are conditionally independent given the class label.

Accuracy Score: 0.425

Precision Score: 0.439

Recall Score: 0.425

F1 Score: 0.402

2. **Logistic Regression:**

Logistic Regression is a statistical method used for binary classification tasks. Despite its name, it is a classification algorithm, not a regression algorithm. It is used to predict the probability that an input belongs to a particular class (usually a binary class, e.g., 0 or 1, true or false, etc.).

Accuracy Score: 0.304

Precision Score: 0.296

Recall Score: 0.304

F1 Score: 0.269

3. **Decision Tree Classifier:**

A Decision Tree Classifier is a non-linear model used for both classification and regression tasks in machine learning. It works by recursively splitting the data into subsets based on the feature values, leading to a tree-like structure. Each internal node of the tree represents a decision (based on a feature), and each leaf node represents a class label (in classification tasks). The model is interpretable and easy to visualize, making it useful for both small and large datasets.

Accuracy Score: 0.648

Precision Score: 0.650

Recall Score: 0.649

F1 Score: 0.649

4. **Random Forest Classifier:**

   Random Forest is an ensemble learning method that combines multiple decision trees to create a more robust and accurate model. Each individual tree in a Random Forest is trained on a random subset of the data, and during the tree-building process, a random subset of features is chosen at each split. This randomness helps ensure that the individual trees are diverse, making the ensemble more powerful and less prone to overfitting

   Accuracy Score: 0.863

   Precision Score: 0.864

   Recall Score: 0.864

   F1 Score: 0.862

5. **AdaBoost Classifier:**

   AdaBoost (Adaptive Boosting) is an ensemble learning method that combines multiple weak learners (typically decision trees) to create a stronger model. AdaBoost works by sequentially applying weak learners to iteratively re-weight the training data, with each subsequent learner focusing more on the mistakes made by previous ones. The final model is a weighted combination of the weak learners, resulting in a model that performs significantly better than a single weak learner.

   Accuracy Score: 0.686

   Precision Score: 0.688

   Recall Score: 0.687

   F1 Score: 0.687

6. **GradientBoostingClassifier:**

   Gradient Boosting is a powerful ensemble learning technique that combines the predictions of several base models (typically decision trees) in a sequential manner to create a more accurate and robust final model. Unlike bagging methods like Random Forest, where trees are built in parallel and combined, boosting builds trees sequentially, with each new tree trying to correct the errors of the previous ones.

   Accuracy Score: 0.821

   Precision Score: 0.822

   Recall Score: 0.687

   F1 Score: 0.687

7. **Support Vector Machine(SVM):**

   Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It is one of the most powerful and versatile machine learning algorithms, particularly for high-dimensional spaces. SVM works by finding the optimal hyperplane that best separates the data into different classes.

   Accuracy Score: 0.314

   Precision Score: 0.279

   Recall Score: 0.315

   F1 Score: 0.275

8. **XGBoost Classifier:**

   XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of the gradient boosting algorithm. It is one of the most popular machine learning algorithms used in data science competitions (e.g., Kaggle) and real-world applications due to its high performance and flexibility. XGBoost is optimized for speed, scalability, and accuracy, making it one of the top choices for classification and regression tasks.

Accuracy Score: 0.886

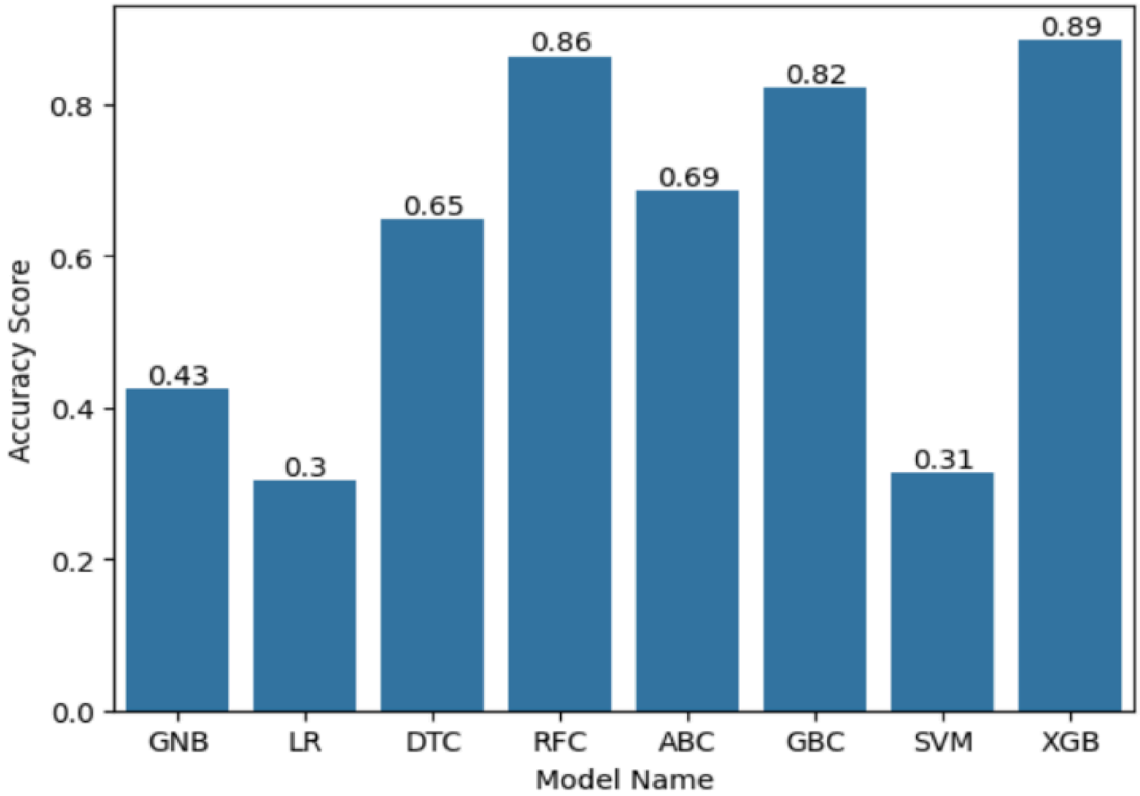
Precision Score: 0.887

Recall Score: 0.886

F1 Score: 0.886

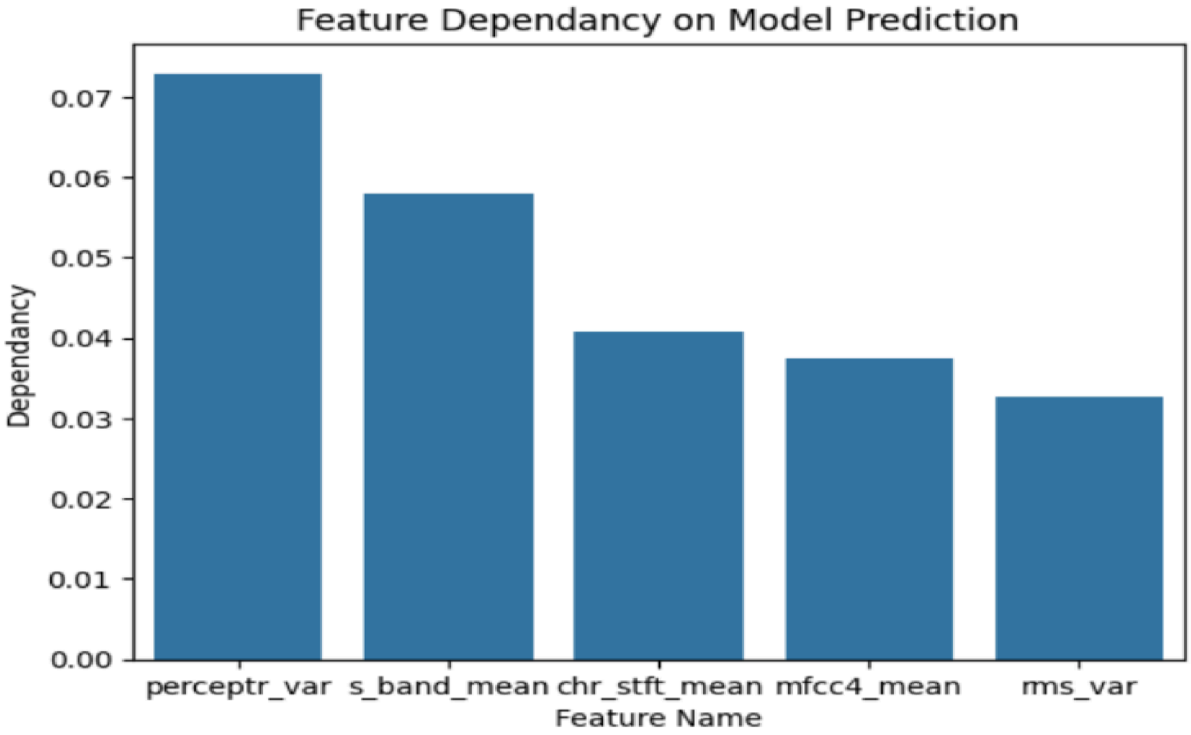
4.Visualization:

a) From the above model evaluation by Gaussian Naive Bayes, Logistic Regression, Decision Tree Classifier, Random Forests Classifier, AdaBoost Classifier, GradientBoostingClassifier, SVC Algorithm, Xgboost Classifier from all these classifiers we plotted a bar graph to visualize accuracy of these algorithms. So, We can easily find out which classifier has maximum accuracy, which is given by the Xgboost algorithm.

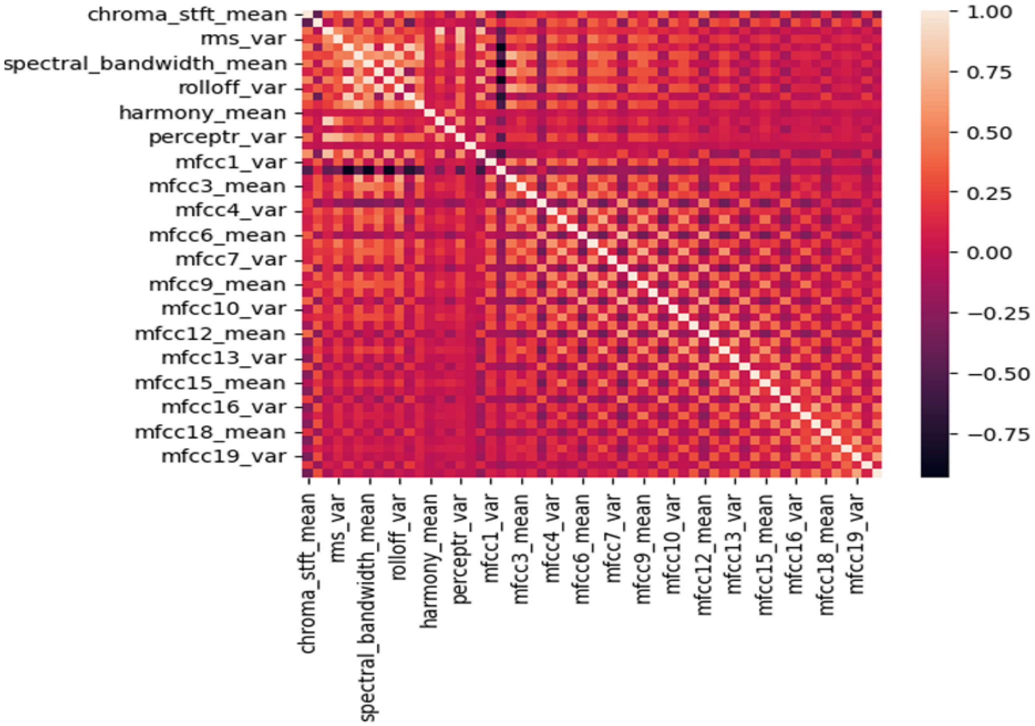


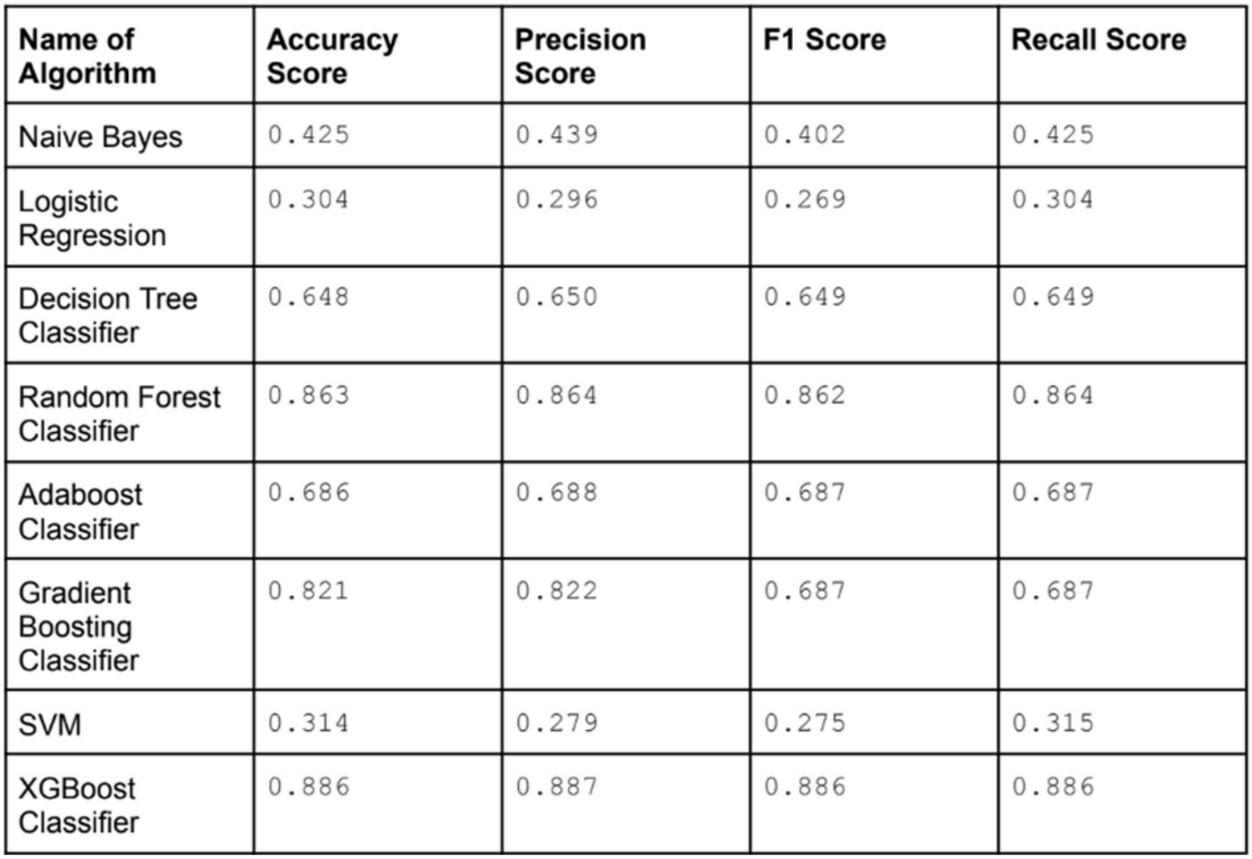
b)After that We have printed the top 5 features on which model accuracy is mostly dependent Using nlargest() function.

c)We have plot a Bar Plot for top 5 feature Dependency on Model Prediction.For visualizing the most dependent Features which is perceptr_var.



d)We have plot a heatmap for finding correlation between features.





Name of Algorithm	Accuracy Score	Precision Score	F1 Score	Recall Score
Naive Bayes	0.425	0.439	0.402	0.425
Logistic Regression	0.304	0.296	0.269	0.304
Decision Tree Classifier	0.648	0.650	0.649	0.649
Random Forest Classifier	0.863	0.864	0.862	0.864
Adaboost Classifier	0.686	0.688	0.687	0.687
Gradient Boosting Classifier	0.821	0.822	0.687	0.687
SVM	0.314	0.279	0.275	0.315
XGBoost Classifier	0.886	0.887	0.886	0.886

CONCLUSION:

The XGBoost Classifier has the highest scores across all metrics (Accuracy: 0.886, Precision: 0.887, F1 Score: 0.886, Recall: 0.886), making it the most effective model for this dataset. Random Forest follows closely, with slightly lower but still strong metrics, indicating it as a solid alternative. Other ensemble models, like Gradient Boosting and AdaBoost, also perform well but are outshined by XGBoost and Random Forest.

In contrast, simpler models like Naive Bayes, Logistic Regression, and SVM show significantly lower accuracy and other metric scores, indicating that they do not capture the patterns in this dataset as effectively as the ensemble methods. This suggests that ensemble learning, especially boosting and bagging methods, are better suited for this dataset, likely due to their ability to reduce overfitting and improve model robustness.

5.Reference:
https://github.com/ABSounds/MusicGenreClassification/blob/main/GenreClassificationML.ipynb

6.Dataset:

https://www.kaggle.com/code/jvedarutvija/music-genre-classification