Dataset Analysis Questions

Note for Students: These questions should be answered based on the 'train.csv' dataset. Please provide your observations, reasoning, and any supporting evidence. You will be evaluated based on the logic and reasoning you use in your answers.

The answers must be written down on a separate document and submitted separately . You can use any word processor to write your answers, such as Microsoft Word, Google Docs, or Notepad. But submit the document as a pdf file.

When asked questions like 'Train a model on the dataset with and without the correlated columns and compare the accuracies'. Make sure to show the training code for both cases on the notebook.

The notebook along with the answer document will be evaluated.

Please note that correlation/pca and methods asked here may or may not improve the accuracy of the model. You will be evaluated based on your reasoning, your experiments, and your observations.

Q1: Selecting a Validation Split Method

- **What is a good validation split method, and what are you planning to use for this dataset, and why?**
  - Discuss different validation split methods, such as k-fold cross-validation, stratified sampling, or a simple train-test split.
  - Explain your choice for this dataset based on its size, class distribution, and other relevant characteristics. Justify why your selected method is appropriate.

Q2: Exploring Column Correlations and Impact on Accuracy

- **Is there a correlation between any of the columns? If so, does removing one of the columns affect the accuracy, and how?**
  - Investigate potential correlations between features. Identify which pairs of columns show a significant correlation. Indicate which columns have correlation and to what degree
  - Assess how removing one of the correlated columns influences the accuracy of a machine learning model. Explain the reasoning behind any observed effects.
  - Train a model on the dataset with and without the correlated columns and compare the accuracies. And note down the accuracies in the document in this format ex. `Accuracy with column: 0.75` and `Accuracy without column: 0.70`

Q3: Analyzing Class Distribution and Addressing Imbalance

- **Plot the distribution of the Genres and discuss if there is any class imbalance.**
  - Create visualizations like histograms or bar plots to visualize the distribution of genres within the dataset.
  - Determine if there is an imbalance in the class distribution and, if so, provide insights on which classes are affected.
  - If there is no imbalance, indicate that there is no imbalance and provide a reason for your conclusion.
  - Please paste the plot in the document and write down your observations.
- **In case there is an imbalance, how would you solve it? List two methods.**
  - Address the issue of class imbalance. Discuss and propose at least two methods for balancing the class distribution, such as oversampling, undersampling, or using synthetic data generation techniques.

Q4: What is overfitting and how will you address it?

- **What is overfitting, and how will you address it?**
  - Explain what overfitting is and how it can be identified.
  - Discuss at least two methods for preventing overfitting.

Q5: What is underfitting and how will you address it?

- **What is underfitting, and how will you address it?**
  - Explain what underfitting is and how it can be identified.
  - Discuss at least two methods for preventing underfitting.

Ensure that your answers are supported by data, visualizations, and clear reasoning. Your insights and explanations will be key to understanding the dataset and making informed decisions for model training and evaluation.