Choice of model:

In this project, we will be training three models and compare them to find the most suitable model for our classifier.

For our first two models, we will be using Logistic Regression and Random forest classifiers since they are prominent choices for classification. Since Logistic Regression is typically a binary classifier and we need to classify multiple genres, we will be incorporating it with the OneVsRest classifier so that a separate model is trained for each genre.

We will then be training a feedforward neural network for our third model. This neural network will be tuned by finding optimal hyperparameters using a combination of trial and error and Grid Search.

Dataset:

Finding a suitable dataset for this project was difficult since there were only two freely that are easily accessible.

Dataset option1:

- books: 91,892
- average length of summary: 157
- hierarchy levels for genres
- languages: English, Spanish, German, Indonesian

Dataset option2:

- books: 16,000
- average length of summary: 520
- no hierarchy levels for genres
- languages: English

Although dataset 1 has considerably more data to train with, we will be using dataset 2 since there are more features (words) for training. Upon analysing the first dataset, we also find that it is in multiple languages whereas dataset 2 is only in English. For this project we will only be training for English features so dataset 2 is more desirable. It also proved to be a difficult task to merge the sub genres in dataset 1 due to the different hierarchy levels in comparison to merging just a few sub genres in dataset 2.