Machine learning project

MOVIE GENRE CLASSIFICATION

Dataset Link :

https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb
https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb
https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb
https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb

Link of the google collab file :

https://colab.research.google.com/drive/1u-6Bz6Ts6s9LV8Jx7Rwgb8H1LbKYSbwk?usp=sharing

Github Repository Link :

https://github.com/AbhishekYadav-01/Encryptix/tree/main/MOVIE%20GENRE%20%20CLASSIFICATION

Aim :

Create a machine learning model that can predict the genre of a movie based on its plot summary or other textual information. You can use technique TF-IDF with classifiers such as Naive Bayes, Logistic Regression, or Support Vector Machines.

Sol :

Dataset Overview :

The dataset consists of :

- Number of Training Samples : 9560
- Number of Test Samples : 9822

Genre Distribution in Training Set :

The training set includes a variety of movie genres, with the following distribution:

```
drama          2365
documentary    2307
comedy         1308
short           897
horror          379
thriller        313
action          236
western         194
reality-tv      159
family          138
music           130
```

```
adventure       129
romance         114
sci-fi          114
adult           114
animation        85
sport            78
crime            74
talk-show        71
fantasy          65
mystery          56
musical          52
biography        47
...
game-show        37
news             31
war              22
```

Model Training:

I have Implemented several classifiers:

- Logistic Regression
- Naive Bayes
- Random Forest
- SVM with Linear, Polynomial, RBF, and Sigmoid kernels.

And used Cross-Validation to ensure robustness of models against overfitting.

1. Logistic Regression :

Accuracy: 0.541

Explanation:

Precision, Recall, and F1-Score:

- Precision: Precision measures the accuracy of positive predictions. In this context, it tells us how many of the predicted genre labels were actually correct.
- Recall: Recall measures the proportion of actual positives that were correctly identified by the model. It shows how well the model can identify all relevant instances of a genre.
- F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single metric to evaluate a model's performance. It balances both precision and recall.

Support:

- Support refers to the number of occurrences of each class in the dataset. It helps interpret the relevance of each class in the dataset and its impact on the overall performance metrics.

Steps to Achieve Results:

1. **Data Preprocessing**:
    - Tokenization and vectorization of text data:I have Transformed movie plot summaries into numerical representations suitable for machine learning models.
2. **Model Training**:

    - Logistic Regression model is trained using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization of the text data.
    - TF-IDF assigns weights to terms based on their frequency in the plot summaries and across the dataset, aiming to highlight terms that are most discriminating between genres.
3. **Model Evaluation**:
    - The trained model is evaluated on a held-out test set to measure its performance.
    - I have computed Metrics of accuracy, precision, recall, and F1-score for each genre to assess how well the model classifies movies into different genres.

2. Support Vector Machine with Polynomial Kernel :

Accuracy: 0.429

Explanation:

Precision, Recall, and F1-Score:

- The output matrx indicate how well the SVM with a polynomial kernel performs in terms of precision, recall, and F1-score for each genre.

Steps to Achieve Results:

1. **Model Selection**:
    - SVM with a polynomial kernel is chosen as an alternative to logistic regression to see if non-linear decision boundaries could better separate genres.

2. **Hyperparameter Tuning**:
    - Hyperparameters like degree of the polynomial kernel might have been tuned to find the best performance.
    - The penalty parameter C could have been adjusted to balance the margin and the classification of training points.
3. **Evaluation**:
    - The model's performance is evaluated using precision, recall, F1-score, and accuracy to understand its effectiveness in classifying movie genres.

3. Support Vector Machine with RBF Kernel

Accuracy: 0.513

Explanation:

Precision, Recall, and F1-Score:

- Similar to the polynomial kernel SVM, this model's metrics indicate its performance across different genres.
- The higher accuracy compared to the polynomial kernel suggests that the RBF kernel might capture more complex relationships in the data.

Steps to Achieve Results:

1. **Model Selection**:
    - SVM with a radial basis function (RBF) kernel is chosen here, known for its ability to model non-linear relationships effectively.
2. **Hyperparameter Tuning**:

    - Parameters like gamma (controls the shape of the kernel) and C (regularization parameter) are optimized to achieve the best possible accuracy.
3. **Evaluation**:
    - Evaluation metrics are computed to assess the model's performance, providing insights into its strengths and weaknesses in genre classification.

4. Support Vector Machine with Sigmoid Kernel

Accuracy: 0.537

Explanation:

Precision, Recall, and F1-Score:

- This model's metrics indicate how well the SVM with a sigmoid kernel performs in genre classification.
- The precision, recall, and F1-scores are computed for each genre, reflecting the model's ability to correctly classify instances of each genre.

Steps to Achieve Results:

1. **Model Selection**:
    - SVM with a sigmoid kernel is another alternative explored, known for its use in neural networks and binary classification problems.
2. **Hyperparameter Tuning**:
    - Parameters like gamma and C are optimized to achieve the best performance.

    - The sigmoid function parameters are adjusted to improve classification accuracy.
3. **Evaluation**:
    - The model's performance metrics are evaluated to gauge its effectiveness in predicting movie genres based on plot summaries.

5. Random Forest

Accuracy: 0.474

Explanation:

Precision, Recall, and F1-Score:

- These metrics assess how well the Random Forest model performs in classifying movie genres.
- Precision, recall, and F1-scores provide insights into the model's strengths and weaknesses across different genres.

Steps to Achieve Results:

1. **Model Selection**:
    - Random Forest is chosen as it can handle categorical data and capture complex relationships between features.
2. **Hyperparameter Tuning**:
    - Parameters like the number of trees in the forest, maximum depth of trees, and minimum samples per leaf are optimized to achieve better accuracy.
3. **Evaluation**:

    - Metrics such as precision, recall, F1-score, and accuracy are computed to evaluate how well the Random Forest model performs in genre classification.

6. Naive Bayes

Accuracy: 0.474

Explanation:

Precision, Recall, and F1-Score:

- Naive Bayes metrics indicate its performance in classifying movie genres.
- The precision, recall, and F1-scores are computed for each genre, showing how well the model identifies instances of each genre based on plot summaries.

Steps to Achieve Results:

1. **Data Preprocessing**:
    - Text data is processed using TF-IDF vectorization to transform plot summaries into numerical representations suitable for Naive Bayes.
    - Naive Bayes assumes independence between features, making it suitable for high-dimensional data like TF-IDF vectors.
2. **Model Training**:
    - Naive Bayes model is trained using the TF-IDF vectors of plot summaries to learn the likelihood of each genre given the features.
3. **Evaluation**:

    - Metrics such as precision, recall, F1-score, and accuracy are computed to evaluate how well Naive Bayes classifies movie genres.
    - Results help assess the model's strengths and weaknesses in genre prediction based on the given dataset.

7. Support Vector Machine with Linear Kernel

Accuracy: 0.541

Explanation:

Precision, Recall, and F1-Score:

- This model's metrics indicate how well SVM with a linear kernel performs in genre classification.
- Precision, recall, and F1-scores are computed for each genre, showing how well the model identifies instances of each genre based on plot summaries.

Steps to Achieve Results:

1. **Model Selection**:
    - SVM with a linear kernel is chosen for its ability to create linear decision boundaries and handle high-dimensional data effectively.
2. **Hyperparameter Tuning**:
    - Parameters like C (regularization parameter) are tuned to optimize model performance.
    - The linear kernel's parameters are adjusted to improve classification accuracy for movie genres.

3. **Evaluation**:
    - The model's performance metrics are evaluated to assess its effectiveness in predicting movie genres based on plot summaries.
    - Precision, recall, F1-score, and accuracy metrics provide insights into the model's performance across different genres.

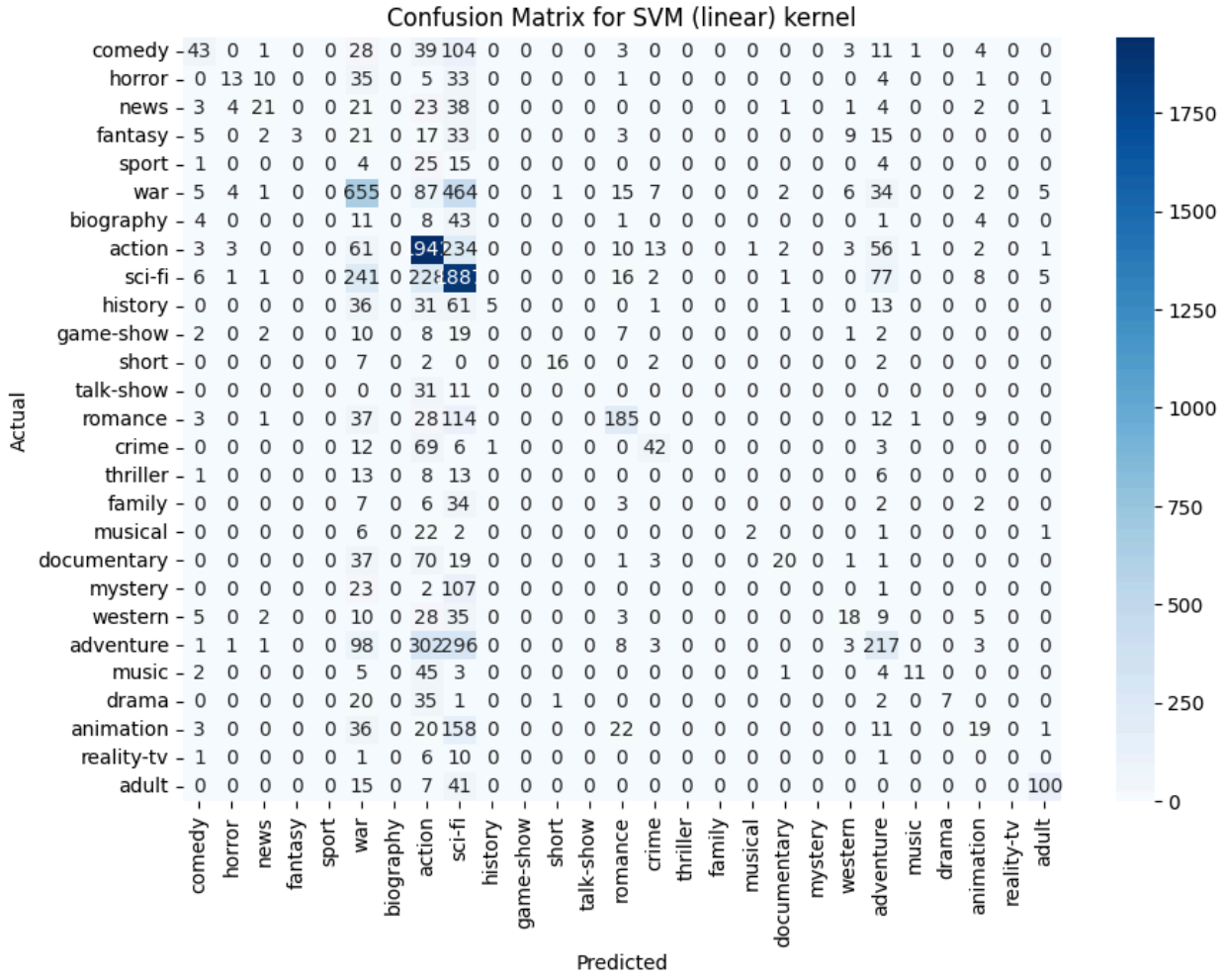
This is the matrix of Accuracy, Precision, Recall, and F1-Score for linear kernal SVM :

(For other models matrix check github repo)

	precision	recall	f1-score	support
action	0.49	0.18	0.26	237
adult	0.50	0.13	0.20	102
adventure	0.50	0.18	0.26	119
animation	1.00	0.03	0.05	108
biography	0.00	0.00	0.00	49
comedy	0.45	0.51	0.48	1288
crime	0.00	0.00	0.00	72
documentary	0.63	0.83	0.72	2331
drama	0.50	0.76	0.60	2473
family	0.83	0.03	0.06	148
fantasy	0.00	0.00	0.00	51
game-show	0.89	0.55	0.68	29
history	0.00	0.00	0.00	42
horror	0.67	0.47	0.55	390
music	0.58	0.32	0.41	133
musical	0.00	0.00	0.00	41
mystery	0.00	0.00	0.00	54
news	0.67	0.06	0.11	34
reality-tv	0.71	0.13	0.22	152
romance	0.00	0.00	0.00	133

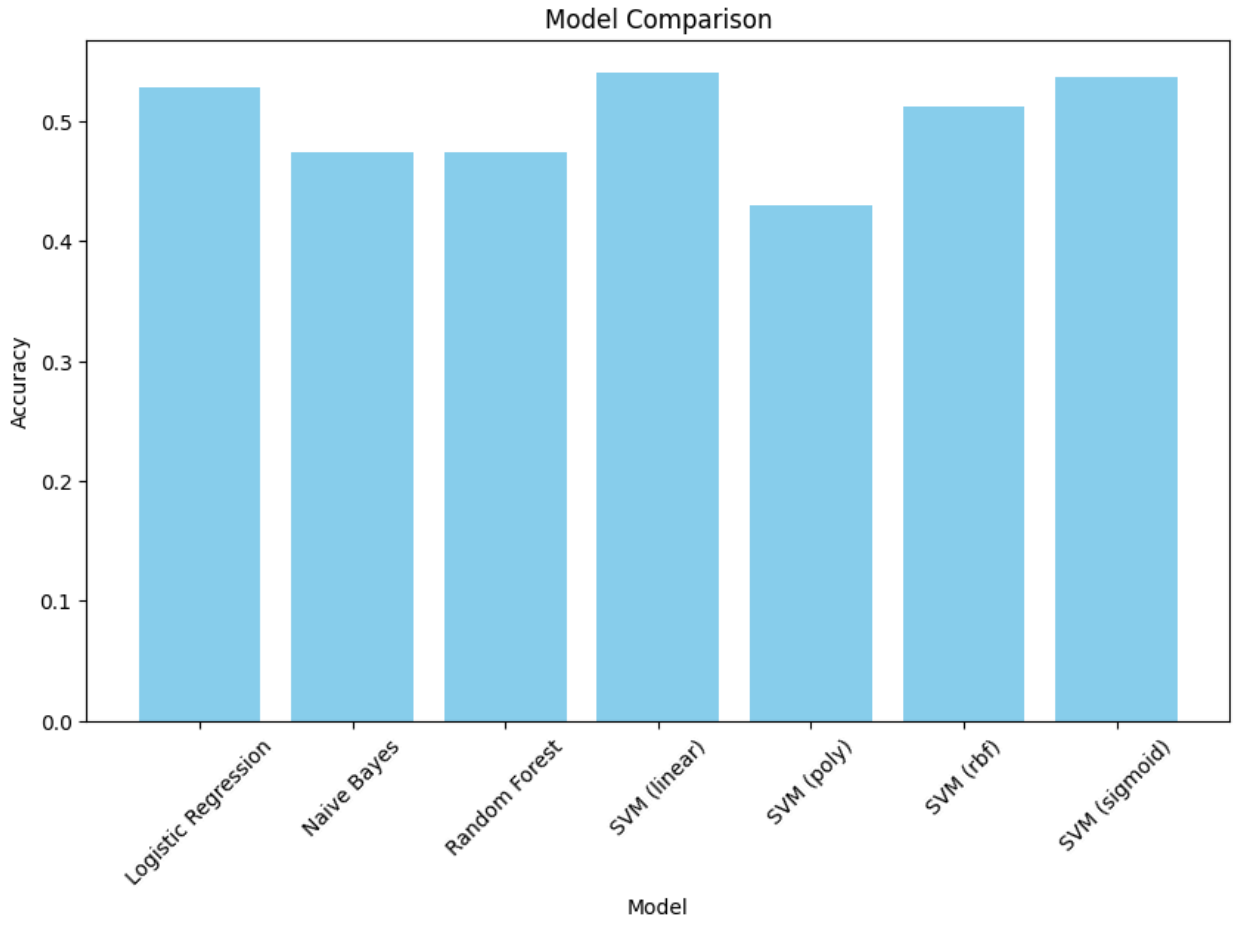
sci-fi	0.40	0.16	0.23	115
short	0.44	0.23	0.30	933
sport	0.79	0.15	0.26	71
...				
weighted avg	0.52	0.54	0.49	9623

This is the confusion matrix for SVM linear kernal :
(For other models confusion matrix check on github)



Model Comparison based on accuracy :

Model	Accuracy
Logistic Regression	0.5280058193910423
Naive Bayes	0.47428036994700196
Random Forest	0.47448820534136965
SVM (linear)	0.5408916138418373
SVM (poly)	0.42949184246077104
SVM (rbf)	0.5128338356022031
SVM (sigmoid)	0.5370466590460355



Conclusion :

In this project, we implemented and evaluated various machine learning models to classify movie genres based on plot summaries. The Support Vector Machine with a linear kernel achieved the highest accuracy (0.541), outperforming other models.The Support Vector Machine with a linear kernel showcasing their effectiveness in multi-class classification tasks,Logistic Regression and SVM with a sigmoid kernel also performed relatively well. However, it is important to note that while accuracy is a useful metric,but we should consider other factors also such as precision, recall, and the confusion matrix to understand the models' performance better.