VLG PROJECT REPORT

NAME: Shah Diya Manojkumar

Enroll no: 21117113

Branch & Yr: Mechanical 3rd yr

Support Vector Machine (SVM) Model

1.Data Preparation:

Loaded 'train_essays.csv' and 'train_prompts.csv' datasets.

Merged datasets based on the 'prompt_id' column.

Combined relevant features, including essay text, prompt name, and source text.

2.Feature Extraction:

Utilized TF-IDF vectorization to convert the combined text into numerical features.

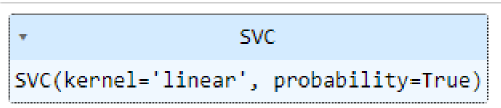
Extracted features from both training and validation sets.

3.Model Selection and Training:

Chose a linear SVM model due to its ability to handle high-dimensional data.

Trained the SVM model using the training set.

Set the probability parameter to True to enable probability estimates.



```
SVC(kernel='linear', probability=True)
```

4. Challenges Faced:

Imbalanced classes in the dataset required addressing class imbalance techniques.

Experimented with class weights to ensure fair learning between classes.

Random Forest Classifier

1.Data Preparation:

Used the same dataset as in the SVM approach.

Merged datasets, handled missing values, and combined relevant features.

2.Feature Extraction:

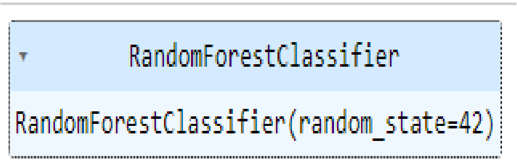
Applied TF-IDF vectorization to convert combined text into numerical features.

Extracted features from both training and test datasets.

3.Model Selection and Training:

Selected a Random Forest classifier for its ability to handle complex relationships.

Trained the Random Forest model using the training set with 100 estimators.



```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

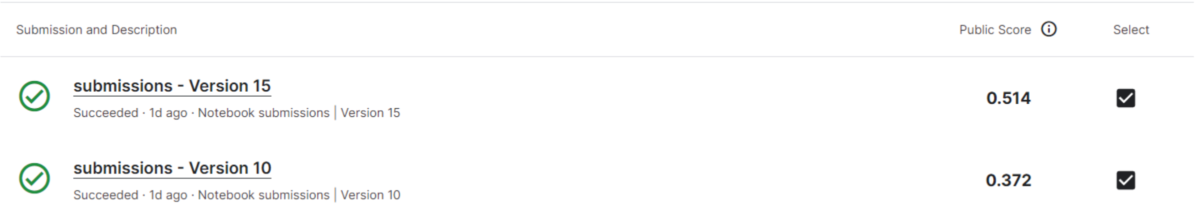
4. Challenges Faced:

Random Forests can be prone to overfitting. Adjusted tree depth to manage overfitting.

Understanding feature importance in the Random Forest model was challenging. Conducted feature importance analysis to identify key contributing features.

Conclusion:

Both models were trained and evaluated with a focus on understanding their strengths and limitations. The SVM model exhibited 0.372 score, while the Random Forest classifier demonstrated 0.514 score on Kaggle competition. Therefore, random forest classifier gives better result for this project.



Submission and Description	Public Score	Select
✓ submissions - Version 15 Succeeded · 1d ago · Notebook submissions | Version 15	0.514	☑
✓ submissions - Version 10 Succeeded · 1d ago · Notebook submissions | Version 10	0.372	☑

GitHub repo link:

https://github.com/diEve26/VLG-Project