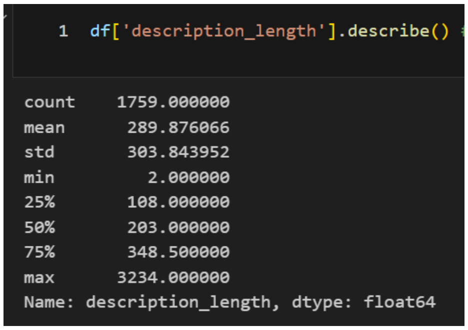
**Task:**

1. **Data Preprocessing:**
    - The data is checked for null values.
    - The description column is examined for the distribution of description length, which ranges between 2 to 3234 and plotted histplot. Explored the truncation function to use in Deep Learning models to truncate till 500 words.
    - The texts in description column are converted to lowercase, removed punctuations, removed stopwords, tokenized and lemmatized.
    - The labels are encoded with Label Encoder.

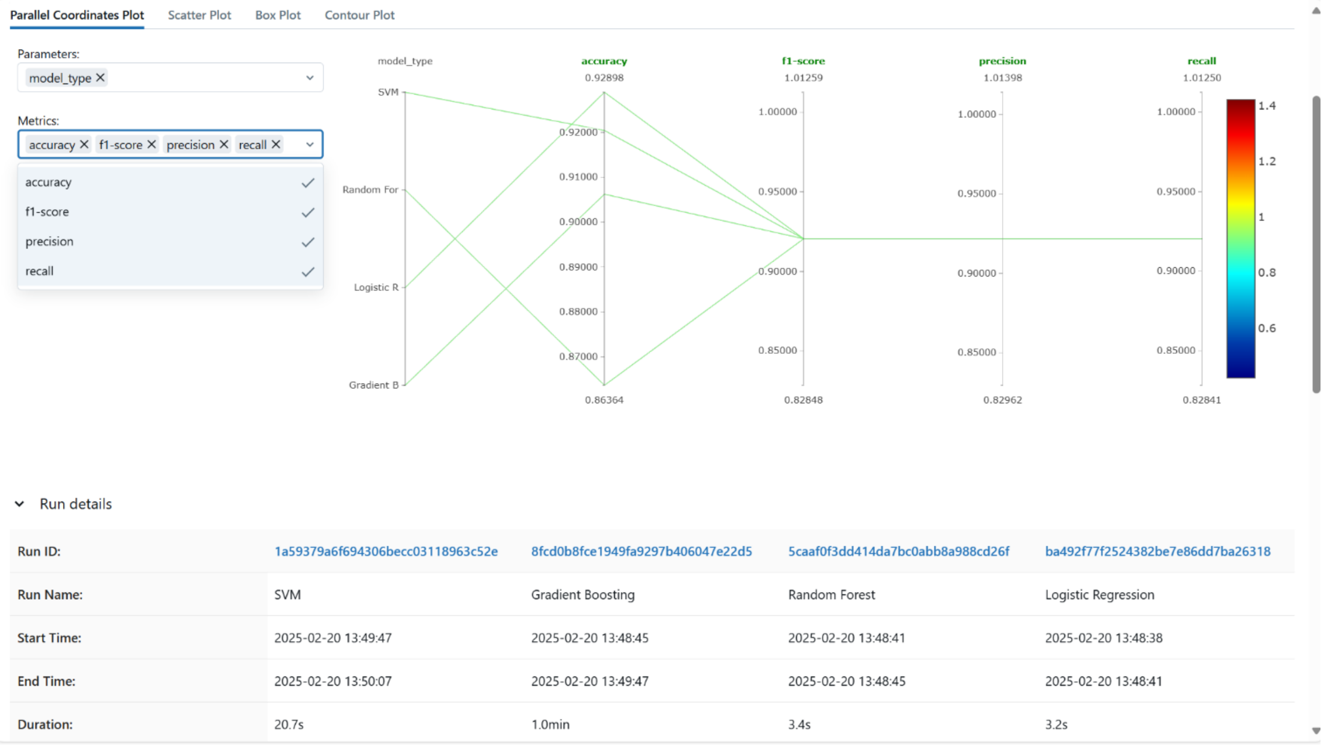


```
df['description_length'].describe()
```

```
count    1759.000000
mean      289.876066
std       303.843952
min         2.000000
25%       108.000000
50%       203.000000
75%       348.500000
max      3234.000000
Name: description_length, dtype: float64
```

*Figure 1. Distribution of Description Length*

2. **Model Selection:**
    - GridSearch is employed for hyperparameter tuning and finding the better perfoming parameters in Logistic Regression and Random SearchCV is used for Random Forest Classifier.
    - Ensembling learning method, stacking classifier is used, where base models are Logistic Regression, Random Forest Classifier(50 estimators) and Gradient Boosting Classifier(50 estimators) and Logistic Regression as meta-learner.
    - MLFlow is employed for evaluating various model performances using the evaluation metrics accuracy, precision, recall, f1-score.



Parameters: model_type

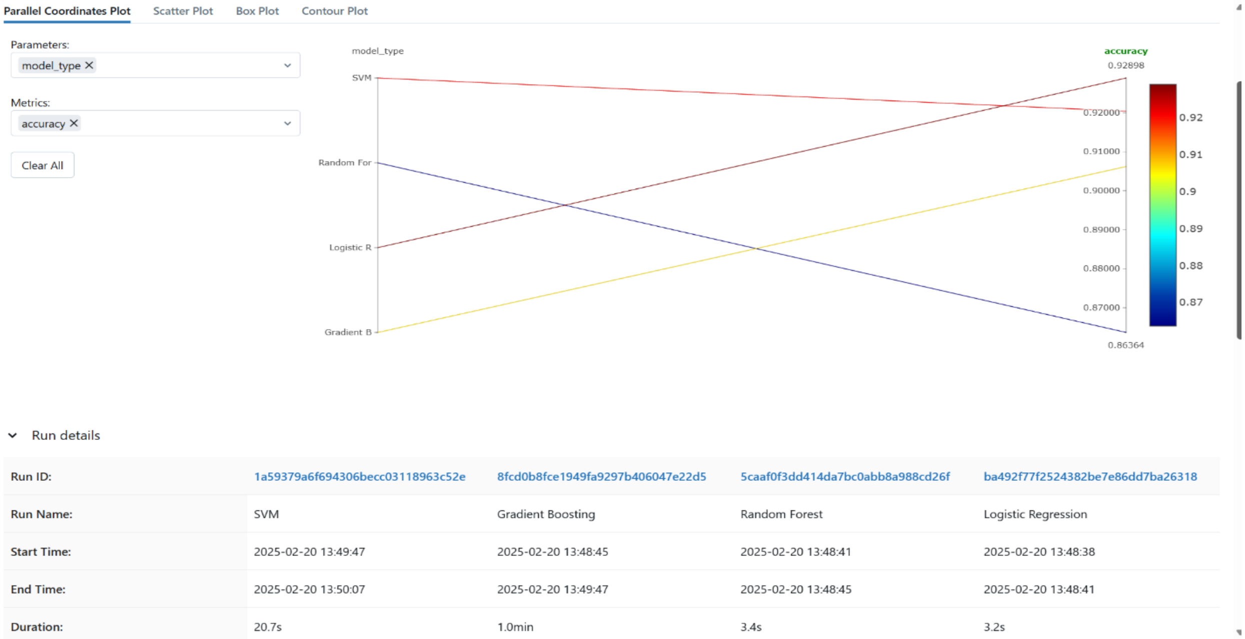
Metrics: accuracy, f1-score, precision, recall

Run details

| | | | | |
|---|---|---|---|---|
| **Run ID:** | 1a59379a6f694306becc03118963c52e | 8fcd0b8fce1949fa9297b406047e22d5 | 5caaf0f3dd414da7bc0abb8a988cd26f | ba492f77f2524382be7e86dd7ba26318 |
| **Run Name:** | SVM | Gradient Boosting | Random Forest | Logistic Regression |
| **Start Time:** | 2025-02-20 13:49:47 | 2025-02-20 13:48:45 | 2025-02-20 13:48:41 | 2025-02-20 13:48:38 |
| **End Time:** | 2025-02-20 13:50:07 | 2025-02-20 13:49:47 | 2025-02-20 13:48:45 | 2025-02-20 13:48:41 |
| **Duration:** | 20.7s | 1.0min | 3.4s | 3.2s |

*Figure 2.Evaluation of Machine Learning Models using MLFlow with accuracy, precision, recall, f1-score*

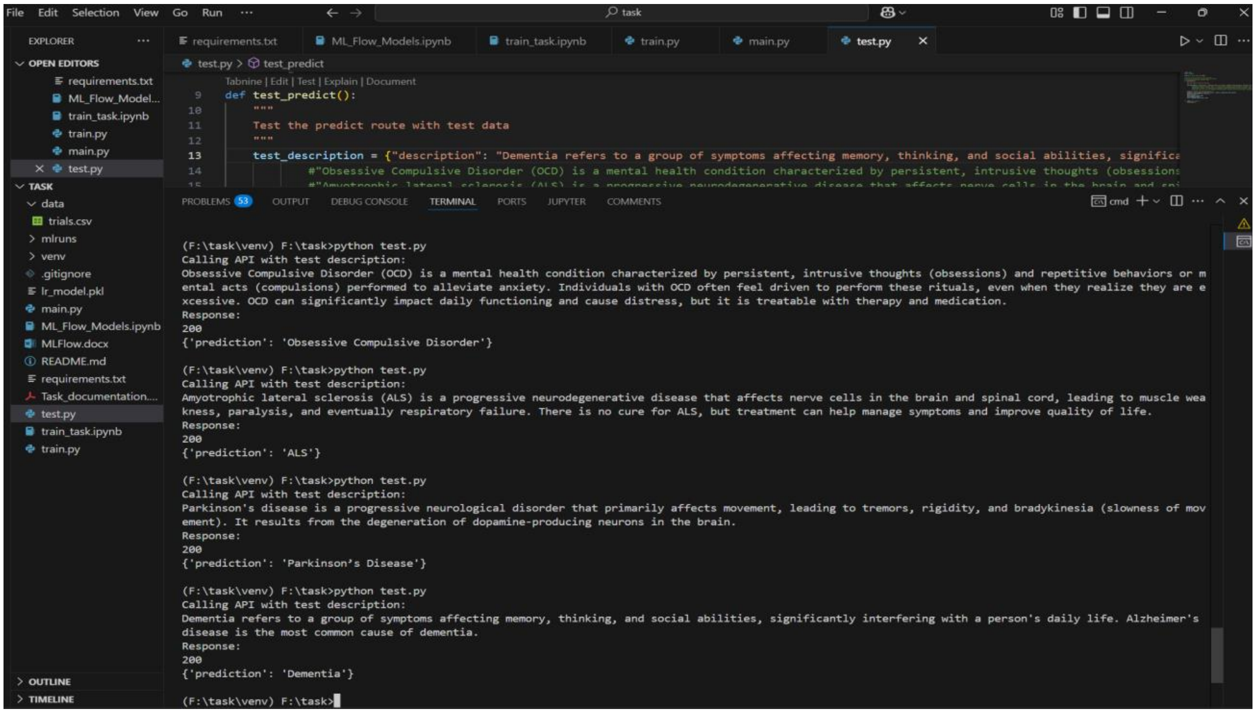
- The models evaluated are Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, in which Logistic Regression showed better performance with accuracy 92.89.



| Run ID: | 1a59379a6f694306becc03118963c52e | 8fcd0b8fce1949fa9297b406047e22d5 | 5caaf0f3dd414da7bc0abb8a988cd26f | ba492f77f2524382be7e86dd7ba26318 |
|---|---|---|---|---|
| Run Name: | SVM | Gradient Boosting | Random Forest | Logistic Regression |
| Start Time: | 2025-02-20 13:49:47 | 2025-02-20 13:48:45 | 2025-02-20 13:48:41 | 2025-02-20 13:48:38 |
| End Time: | 2025-02-20 13:50:07 | 2025-02-20 13:49:47 | 2025-02-20 13:48:45 | 2025-02-20 13:48:41 |
| Duration: | 20.7s | 1.0min | 3.4s | 3.2s |

*Figure 3. Evaluation of Machine Learning Models using MLFlow with accuracy*

- Logistic Regression is model is selected and saved for inferencing and exhibiting excellent performance and would able to predict the text related to the 4 medical conditions.



```
def test_predict():
    """
    Test the predict route with test data
    """
    test_description = {"description": "Dementia refers to a group of symptoms affecting memory, thinking, and social abilities, significa
            #"Obsessive Compulsive Disorder (OCD) is a mental health condition characterized by persistent, intrusive thoughts (obsessions
            #"Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease that affects nerve cells in the brain and spi
```

```
(F:\task\venv) F:\task>python test.py
Calling API with test description:
Obsessive Compulsive Disorder (OCD) is a mental health condition characterized by persistent, intrusive thoughts (obsessions) and repetitive behaviors or m
ental acts (compulsions) performed to alleviate anxiety. Individuals with OCD often feel driven to perform these rituals, even when they realize they are e
xcessive. OCD can significantly impact daily functioning and cause distress, but it is treatable with therapy and medication.
Response:
200
{'prediction': 'Obsessive Compulsive Disorder'}

(F:\task\venv) F:\task>python test.py
Calling API with test description:
Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease that affects nerve cells in the brain and spinal cord, leading to muscle wea
kness, paralysis, and eventually respiratory failure. There is no cure for ALS, but treatment can help manage symptoms and improve quality of life.
Response:
200
{'prediction': 'ALS'}

(F:\task\venv) F:\task>python test.py
Calling API with test description:
Parkinson's disease is a progressive neurological disorder that primarily affects movement, leading to tremors, rigidity, and bradykinesia (slowness of mov
ement). It results from the degeneration of dopamine-producing neurons in the brain.
Response:
200
{'prediction': 'Parkinson’s Disease'}

(F:\task\venv) F:\task>python test.py
Calling API with test description:
Dementia refers to a group of symptoms affecting memory, thinking, and social abilities, significantly interfering with a person's daily life. Alzheimer's
disease is the most common cause of dementia.
Response:
200
{'prediction': 'Dementia'}

(F:\task\venv) F:\task>
```

*Figure 4. Model Predictions*