# Data Collection And Preprocessing Phase

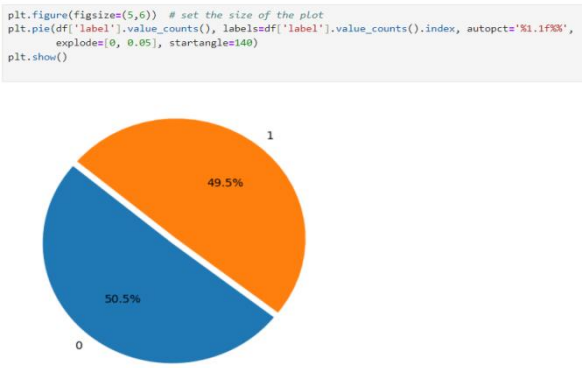
| | |
|---|---|
| Date | 15 November 2024 |
| TeamID | 739909 |
| ProjectName | Unlocking the Minds: Analyzing Mental Health with NLP |
| Maximum Marks | 6 Marks |

**Data Exploration and Preprocessing Report:**

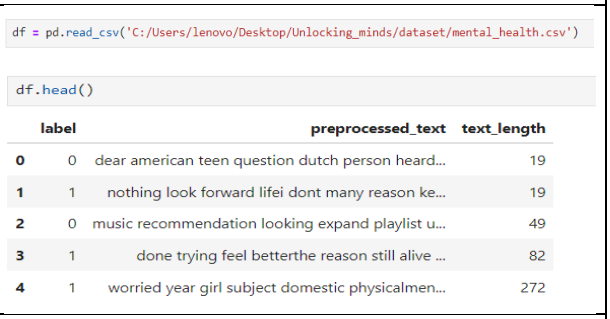
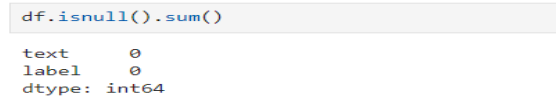
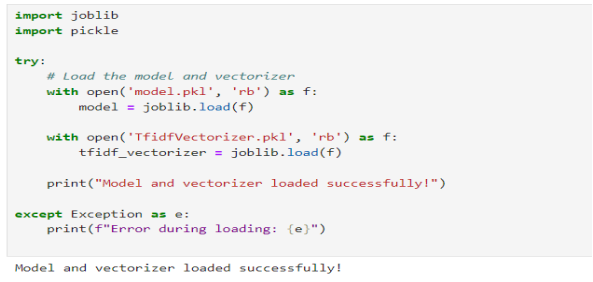
Dataset variables will undergo statistical analysis to identify patterns and outliers. Python will be used for preprocessing tasks such as normalization and feature engineering. Data cleaning will address missing values and outliers to ensure data quality for subsequent analysis and modeling, forming a robust foundation for insights and predictions.

| Section | Description |
|---|---|
| Data Overview | `df.describe(include='all')` |

| | text | label |
|---|---|---|
| count | 27972 | 27972.000000 |
| unique | 27972 | NaN |
| top | dear american teens question dutch person hear... | NaN |
| freq | 1 | NaN |
| mean | NaN | 0.494709 |
| std | NaN | 0.499981 |
| min | NaN | 0.000000 |
| 25% | NaN | 0.000000 |
| 50% | NaN | 0.000000 |
| 75% | NaN | 1.000000 |
| max | NaN | 1.000000 |

| | |
|---|---|
| Univariate Analysis |  |
| Outliers and Anomalies | - |

## Data Preprocessing Code Screenshots

| | |
|---|---|
| Loading Data |  |
| Handling Missing Data |  |
| Data Transformation | - |
| Feature Engineering | Attached the codes in final submission. |
| Save Processed Data |  |



```
plt.figure(figsize=(5,6))  # set the size of the plot
plt.pie(df['label'].value_counts(), labels=df['label'].value_counts().index, autopct='%1.1f%%',
        explode=[0, 0.05], startangle=140)
plt.show()
```

```
df = pd.read_csv('C:/Users/lenovo/Desktop/Unlocking_minds/dataset/mental_health.csv')
```

```
df.head()
```

| | label | preprocessed_text | text_length |
|---|---|---|---|
| 0 | 0 | dear american teen question dutch person heard... | 19 |
| 1 | 1 | nothing look forward lifei dont many reason ke... | 19 |
| 2 | 0 | music recommendation looking expand playlist u... | 49 |
| 3 | 1 | done trying feel betterthe reason still alive ... | 82 |
| 4 | 1 | worried year girl subject domestic physicalmen... | 272 |

```
df.isnull().sum()
```

```
text      0
label     0
dtype: int64
```

```
import joblib
import pickle

try:
    # Load the model and vectorizer
    with open('model.pkl', 'rb') as f:
        model = joblib.load(f)

    with open('TfidfVectorizer.pkl', 'rb') as f:
        tfidf_vectorizer = joblib.load(f)

    print("Model and vectorizer loaded successfully!")

except Exception as e:
    print(f"Error during loading: {e}")
```

```
Model and vectorizer loaded successfully!
```