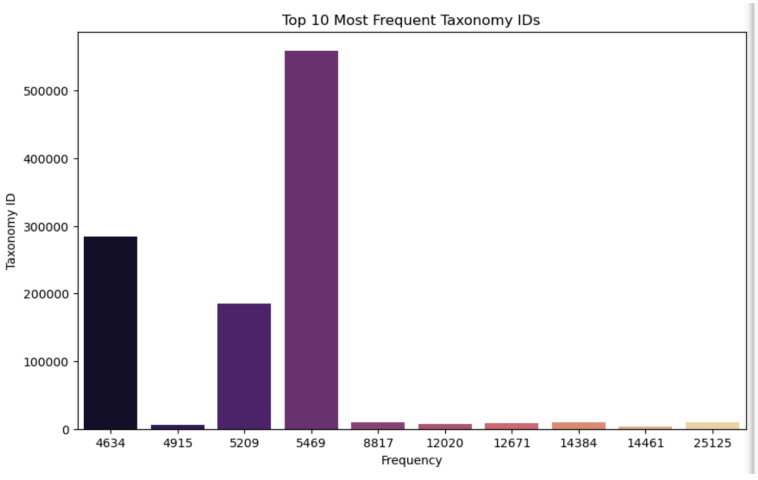
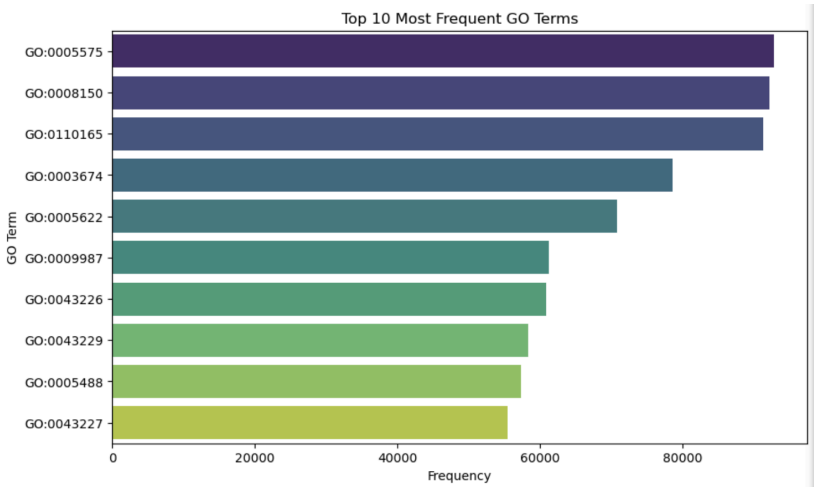
Exploratory Data Analysis

```
Train Taxonomy Data Summary:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142246 entries, 0 to 142245
Data columns (total 2 columns):
 #   Column      Non-Null Count   Dtype 
---  ------      --------------   ----- 
 0   EntryID     142246 non-null  object
 1   taxonomyID  142246 non-null  int64 
dtypes: int64(1), object(1)
memory usage: 2.2+ MB
None

Train Taxonomy Data Sample:
  EntryID  taxonomyID
0  Q8IXT2        9606
1  Q04418      559292
2  A8DYA3        7227
3  Q9UUI3      284812
4  Q57ZS4      185431

Train Terms Data Summary:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5363863 entries, 0 to 5363862
Data columns (total 3 columns):
 #   Column   Dtype 
---  ------   ----- 
 0   EntryID  object
 1   term     object
 2   aspect   object
dtypes: object(3)
memory usage: 122.8+ MB
None

Train Terms Data Sample:
      EntryID        term aspect
0  A0A009IHW8  GO:0008152    BPO
1  A0A009IHW8  GO:0034655    BPO
2  A0A009IHW8  GO:0072523    BPO
3  A0A009IHW8  GO:0044270    BPO
4  A0A009IHW8  GO:0006753    BPO
```

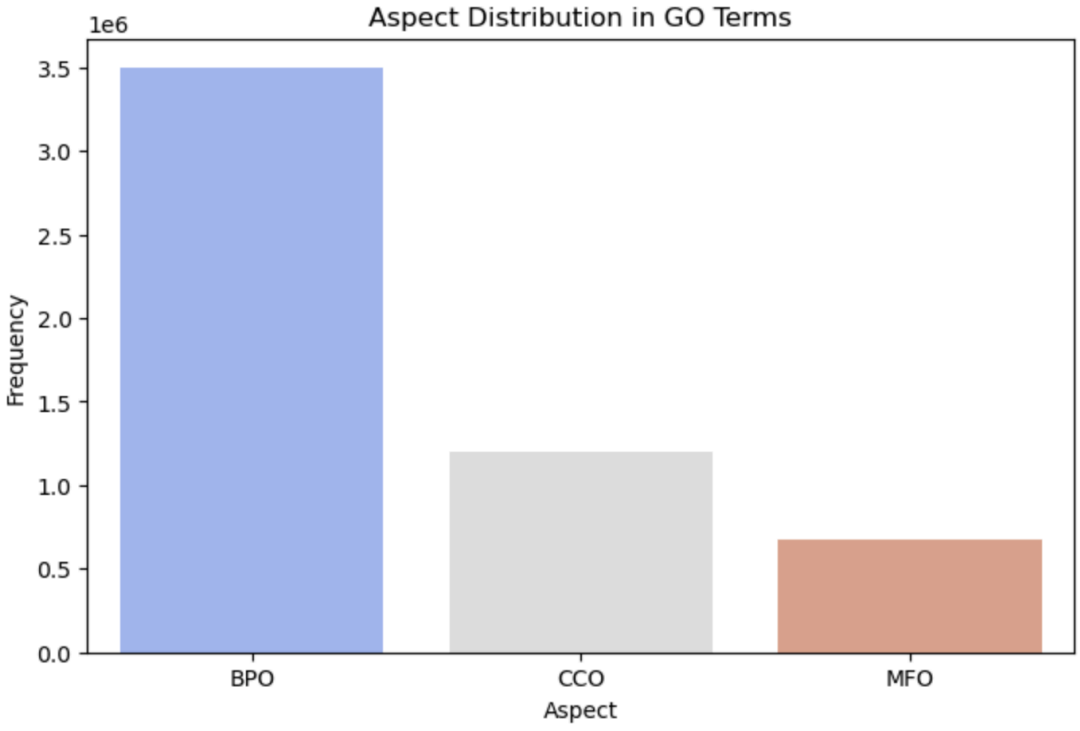
```
Missing Values in Train Taxonomy:
EntryID       0
taxonomyID    0
dtype: int64

Missing Values in Train Terms:
EntryID    0
term       0
aspect     0
dtype: int64

Unique taxonomy IDs in train_taxonomy.tsv:
3156

Unique GO terms in train_terms.tsv:
31466

Top 10 Most Frequent Taxonomy IDs:
taxonomyID
9606      25125
3702      14461
10090     14384
7955      12671
7227      12020
10116      8817
559292     5469
185431     5209
6239       4915
284812     4634
Name: count, dtype: int64
```



```
Top 10 Most Frequent GO Terms:
term
GO:0005575    92912
GO:0008150    92210
GO:0110165    91286
GO:0003674    78637
GO:0005622    70785
GO:0009987    61293
GO:0043226    60883
GO:0043229    58315
GO:0005488    57380
GO:0043227    55452
Name: count, dtype: int64
```

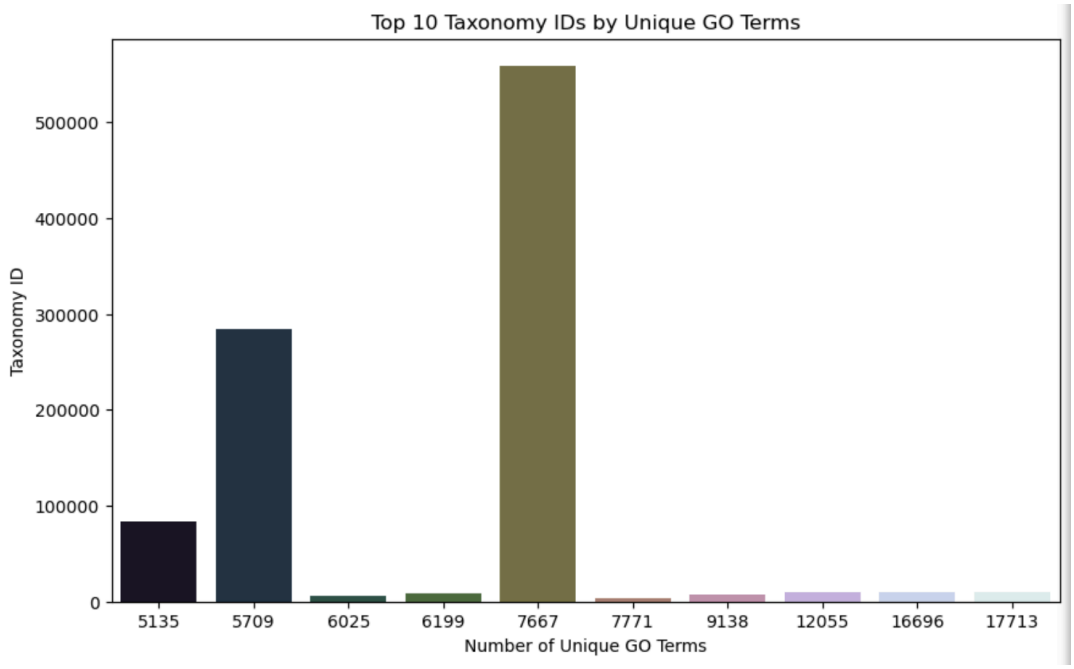



```
Aspect Distribution in GO Terms:
aspect
BPO    3497732
CCO    1196017
MFO     670114
Name: count, dtype: int64
```



```
Merging train_taxonomy and train_terms datasets...
Merged Data Summary:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5363863 entries, 0 to 5363862
Data columns (total 4 columns):
 #   Column      Dtype
---  ------      -----
 0   EntryID     object
 1   taxonomyID  int64
 2   term        object
 3   aspect      object
dtypes: int64(1), object(3)
memory usage: 163.7+ MB
None
```

```
Merged Data Sample:
  EntryID  taxonomyID        term aspect
0  Q8IXT2        9606  GO:0003677    MFO
1  Q8IXT2        9606  GO:1990837    MFO
2  Q8IXT2        9606  GO:0003676    MFO
3  Q8IXT2        9606  GO:0005488    MFO
4  Q8IXT2        9606  GO:0003690    MFO

Top 10 Taxonomy IDs by Number of GO Terms:
taxonomyID
9606      17713
10090     16696
10116     12055
7227       9138
3702       7771
559292     7667
7955       6199
6239       6025
284812     5709
83333      5135
Name: term, dtype: int64
```



Data Preprocessing

```
Removing duplicates...
train_taxonomy shape after removing duplicates: (142246, 2)
train_terms shape after removing duplicates: (5363863, 3)

Handling missing values...
Missing values in train_taxonomy:
EntryID       0
taxonomyID    0
dtype: int64
```

```
Missing values in train_terms:
EntryID    0
term       0
aspect     0
dtype: int64
train_taxonomy shape after dropping missing values: (142246, 2)
train_terms shape after dropping missing values: (5363863, 3)

Filtering invalid entries...
train_taxonomy shape after filtering invalid taxonomyID: (142246, 2)
train_terms shape after filtering invalid GO terms: (5363863, 3)

Encoding categorical features...
Aspect encoding mapping: {'BPO': 0, 'CCO': 1, 'MFO': 2}

Normalizing taxonomyID...

Merging processed datasets...
Processed data shape: (5363863, 6)

Processed data saved to 'processed_data.csv'.
```

Feature Engineering

```
Starting Feature Engineering...
Extracting features from GO terms...
Encoding GO terms...
Number of unique GO terms: 31466
Scaling numeric features...
Adding derived features...
One-hot encoding aspect column...
Creating interaction features...
Extracting text-based features from aspect (if applicable)...

Feature Engineering Completed.
Engineered data saved to 'engineered_data.csv'.
```

Model Selection

```
Splitting data into training and testing sets...
Training data size: (4291090, 12)
Testing data size: (1072773, 12)
Scaling numeric features...
```

```
Training NaiveBayes...
NaiveBayes training completed.
Evaluating NaiveBayes on the test set...
NaiveBayes Accuracy: 0.9681
NaiveBayes Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.98    954213
           1       0.60      0.99      0.75     50026
           2       0.77      0.94      0.85      2701
           3       1.00      0.99      0.99     62888
           4       0.38      0.95      0.55       446
           5       0.64      0.91      0.75       279
           6       0.99      0.92      0.96       752
           7       0.85      0.84      0.84       433
           8       0.75      0.85      0.80       255
           9       0.75      0.46      0.57        83
          10       0.43      1.00      0.60        86
          11       0.94      0.30      0.46       105
          12       0.96      0.96      0.96        52
          13       1.00      1.00      1.00        12
          15       1.00      0.97      0.99        79
          16       0.97      0.93      0.95       245
          17       0.88      0.97      0.92       118

    accuracy                           0.97   1072773
   macro avg       0.82      0.88      0.82   1072773
weighted avg       0.98      0.97      0.97   1072773

NaiveBayes model saved as 'NaiveBayes.pth' (compressed).

Training DecisionTree...
DecisionTree training completed.
Evaluating DecisionTree on the test set...
DecisionTree Accuracy: 1.0000
DecisionTree Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    954213
           1       1.00      1.00      1.00     50026
           2       1.00      1.00      1.00      2701
           3       1.00      1.00      1.00     62888
           4       1.00      1.00      1.00       446
           5       1.00      1.00      1.00       279
           6       1.00      1.00      1.00       752
           7       1.00      1.00      1.00       433
           8       1.00      1.00      1.00       255
           9       1.00      1.00      1.00        83
          10       1.00      1.00      1.00        86
          11       1.00      1.00      1.00       105
          12       1.00      1.00      1.00        52
          13       1.00      1.00      1.00        12
          15       1.00      1.00      1.00        79
          16       1.00      1.00      1.00       245
          17       1.00      1.00      1.00       118

    accuracy                           1.00   1072773
   macro avg       1.00      1.00      1.00   1072773
weighted avg       1.00      1.00      1.00   1072773

DecisionTree model saved as 'DecisionTree.pth' (compressed).
```

```
Training RandomForest...
RandomForest training completed.
Evaluating RandomForest on the test set...
RandomForest Accuracy: 1.0000
RandomForest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    954213
           1       1.00      1.00      1.00     50026
           2       1.00      1.00      1.00      2701
           3       1.00      1.00      1.00     62888
           4       1.00      1.00      1.00       446
           5       1.00      1.00      1.00       279
           6       1.00      1.00      1.00       752
           7       1.00      1.00      1.00       433
           8       1.00      1.00      1.00       255
           9       1.00      1.00      1.00        83
          10       1.00      0.99      0.99        86
          11       0.97      1.00      0.99       105
          12       1.00      0.92      0.96        52
          13       0.86      1.00      0.92        12
          15       1.00      0.99      0.99        79
          16       0.99      1.00      0.99       245
          17       1.00      0.98      0.99       118

    accuracy                           1.00   1072773
   macro avg       0.99      0.99      0.99   1072773
weighted avg       1.00      1.00      1.00   1072773

RandomForest model saved as 'RandomForest.pth' (compressed).
```

```
Training SGDClassifier (Approximate SVM)...
SGDClassifier (Approximate SVM) training completed.
Evaluating SGDClassifier (Approximate SVM) on the test set...
SGDClassifier (Approximate SVM) Accuracy: 0.9728
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetr
icWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetr
icWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetr
icWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
SGDClassifier (Approximate SVM) Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00    954213
           1       1.00      0.53      0.69     50026
           2       0.00      0.00      0.00      2701
           3       0.81      1.00      0.90     62888
           4       0.00      0.00      0.00       446
           5       0.00      0.00      0.00       279
           6       0.00      0.00      0.00       752
           7       0.00      0.00      0.00       433
           8       0.00      0.00      0.00       255
           9       0.00      0.00      0.00        83
          10       0.00      0.00      0.00        86
          11       0.00      0.00      0.00       105
          12       0.00      0.00      0.00        52
          13       0.00      0.00      0.00        12
          15       0.00      0.00      0.00        79
          16       0.00      0.00      0.00       245
          17       0.00      0.00      0.00       118

    accuracy                           0.97   1072773
   macro avg       0.16      0.15      0.15   1072773
weighted avg       0.98      0.97      0.97   1072773
```

```
SGDClassifier (Approximate SVM) model saved as 'SGDClassifier_(Approximate_SVM).pth' (compressed).

Training SVM on a subset of the data...
SVM training completed on subset.
SVM Accuracy: 0.9946
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetr
icWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
SVM Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    954213
           1       0.92      0.99      0.95     50026
           2       0.97      0.88      0.92      2701
           3       1.00      1.00      1.00     62888
           4       0.88      0.83      0.85       446
           5       0.69      0.90      0.78       279
           6       0.99      0.91      0.95       752
           7       0.84      0.89      0.86       433
           8       0.77      0.80      0.78       255
           9       0.00      0.00      0.00        83
          10       0.53      0.94      0.68        86
          11       0.67      0.53      0.60       105
          12       0.33      0.48      0.39        52
          13       0.00      0.00      0.00        12
          15       0.16      0.72      0.26        79
          16       0.62      0.18      0.28       245
          17       0.53      0.07      0.12       118

    accuracy                           0.99   1072773
   macro avg       0.64      0.65      0.61   1072773
weighted avg       0.99      0.99      0.99   1072773

SVM model saved as 'SVM_Subset.pth' (compressed).

 All models trained and saved successfully.
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetr
icWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetr
icWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

Model Comparison

```
Evaluating saved models...

Loading model: NaiveBayes
```

```
/opt/anaconda3/lib/python3.12/site-packages/sklearn/base.py:486: UserWarning: X has feature names, but GaussianNB was fitted without featu
re names
  warnings.warn(
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/base.py:486: UserWarning: X has feature names, but DecisionTreeClassifier was fitted w
ithout feature names
  warnings.warn(
```

```
NaiveBayes Accuracy: 0.0610
NaiveBayes Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.02      0.03   4771062
           1       0.05      0.98      0.09    250129
           2       0.00      0.00      0.00     13504
           3       0.81      0.00      0.00    314440
           4       0.00      0.00      0.00      2227
           5       0.44      0.00      0.01      1397
           6       0.00      0.00      0.00      3760
           7       0.00      0.00      0.00      2166
           8       0.00      0.00      0.00      1275
           9       0.00      0.00      0.00       413
          10       0.00      0.00      0.00       432
          11       0.00      0.00      0.00       525
          12       0.00      0.00      0.00       260
          13       0.00      0.00      0.00        62
          15       0.00      0.00      0.00       393
          16       0.38      0.00      0.00      1226
          17       0.00      0.00      0.00       592

    accuracy                           0.06   5363863
   macro avg       0.15      0.06      0.01   5363863
weighted avg       0.81      0.06      0.03   5363863
```

```
Loading model: DecisionTree
DecisionTree Accuracy: 1.0000
DecisionTree Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   4771062
           1       1.00      1.00      1.00    250129
           2       1.00      1.00      1.00     13504
           3       1.00      1.00      1.00    314440
           4       1.00      1.00      1.00      2227
           5       1.00      1.00      1.00      1397
           6       1.00      1.00      1.00      3760
           7       1.00      1.00      1.00      2166
           8       1.00      1.00      1.00      1275
           9       1.00      1.00      1.00       413
          10       1.00      1.00      1.00       432
          11       1.00      1.00      1.00       525
          12       1.00      1.00      1.00       260
          13       1.00      1.00      1.00        62
          15       1.00      1.00      1.00       393
          16       1.00      1.00      1.00      1226
          17       1.00      1.00      1.00       592

    accuracy                           1.00   5363863
   macro avg       1.00      1.00      1.00   5363863
weighted avg       1.00      1.00      1.00   5363863
```

```
Loading model: RandomForest
/opt/anaconda3/lib/python3.12/site-packages/sklearn/base.py:486: UserWarning: X has feature names, but RandomForestClassifier was fitted w
ithout feature names
  warnings.warn(
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/base.py:486: UserWarning: X has feature names, but SGDClassifier was fitted without fe
ature names
  warnings.warn(
RandomForest Accuracy: 0.9993
RandomForest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   4771062
           1       1.00      1.00      1.00    250129
           2       1.00      0.99      0.99     13504
           3       1.00      1.00      1.00    314440
           4       1.00      0.83      0.91      2227
           5       0.94      1.00      0.97      1397
           6       1.00      0.97      0.98      3760
           7       0.56      1.00      0.72      2166
           8       0.00      0.00      0.00      1275
           9       0.00      0.00      0.00       413
          10       0.00      0.00      0.00       432
          11       0.46      0.94      0.62       525
          12       0.00      0.00      0.00       260
          13       0.00      0.00      0.00        62
          15       0.00      0.00      0.00       393
          16       0.63      1.00      0.77      1226
          17       1.00      0.81      0.89       592

    accuracy                           1.00   5363863
   macro avg       0.56      0.62      0.58   5363863
weighted avg       1.00      1.00      1.00   5363863

Loading model: SGDClassifier (Approximate SVM)
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
SGDClassifier (Approximate SVM) Accuracy: 0.0368
SGDClassifier (Approximate SVM) Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.04      0.07   4771062
           1       0.53      0.00      0.00    250129
           2       0.01      0.00      0.00     13504
           3       0.20      0.03      0.06    314440
           4       0.00      0.00      0.00      2227
           5       0.00      0.99      0.00      1397
           6       0.00      0.00      0.00      3760
           7       0.00      0.00      0.00      2166
           8       0.00      0.00      0.00      1275
           9       0.00      0.00      0.00       413
          10       0.00      0.00      0.00       432
          11       0.00      0.00      0.00       525
          12       0.00      0.00      0.00       260
          13       0.00      0.00      0.00        62
          15       0.00      0.00      0.00       393
          16       0.00      0.00      0.00      1226
          17       0.00      0.00      0.00       592

    accuracy                           0.04   5363863
   macro avg       0.10      0.06      0.01   5363863
weighted avg       0.93      0.04      0.07   5363863
```

```
Loading model: SVM Subset
Error evaluating SVM Subset: name 'PCA' is not defined

Summary of Model Performance:
                            Model  Accuracy
0                      NaiveBayes  0.060986
1                    DecisionTree  1.000000
2                    RandomForest  0.999324
3  SGDClassifier (Approximate SVM)  0.036825
```

```
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```