Epileptic Seizure Prediction

XIU ZHEN HUANG
22.09.2025

Overview

- Brief Description of Dataset
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Models Detail
  - Logistic Regression
  - Support Vector Classification (SVC)
  - ExtraTreesClassifier
- Results and Analysis
- Conclusion

The full project locates at the git repository.

Brief Description of Dataset

- The "Epileptic Seizure Classification" dataset available on BEED: Bangalore EEG Epilepsy Dataset, UCI Machine Learning Repository.
- The training dataset includes **8000 records** and contains 16,000 segments of 20-second EEG recordings labeled 0, 1, 2, 3, which means four epilepsy categories: **Healthy Subjects (0), Generalized Seizures (1), Focal Seizures (2), and Seizure Events (3)**.
- It includes **16 EEG channels (X1-X16)** which corresponding to different brain regions(16 features) and **1 target variable (y)** which refer to 4 categories

There are 8000 rows in the dataset.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	y
0	4	7	18	25	28	27	20	10	-10	-18	-20	-16	13	32	12	10	0
1	87	114	120	106	76	54	28	5	-19	-49	-85	-102	-100	-89	-61	-21	0
2	-131	-133	-140	-131	-123	-108	-58	-51	-70	-77	-76	-76	-73	-57	-40	-14	0
3	68	104	73	34	-12	-26	-38	-36	-67	-88	-25	31	18	-4	6	-29	0

Data Cleaning

- The dataset almost is a cleaning dataset that don’t need to clean it.
- It’s still necessary to check the data :
  - Dose it has null value ?
  - Are the labels all in [0, 1, 2, 3] ?
  - Are there any outer value ?
- The tables show that there is no null volue, outer value and all labels are in [1, 2, 3, 4].

```
y
0    2000
1    2000
2    2000
3    2000
Name: count, dtype: int64
```

```
Null value count:

X1     0
X2     0
X3     0
X4     0
X5     0
X6     0
X7     0
X8     0
X9     0
X10    0
X11    0
X12    0
X13    0
X14    0
X15    0
X16    0
y      0
dtype: int64
```

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
min	-281	-255	-255	-257	-264	-277	-277	-260	-290	-302	-276	-306	-288	-290	-323	-317
max	252	261	238	246	249	245	220	271	280	251	262	283	296	291	251	270

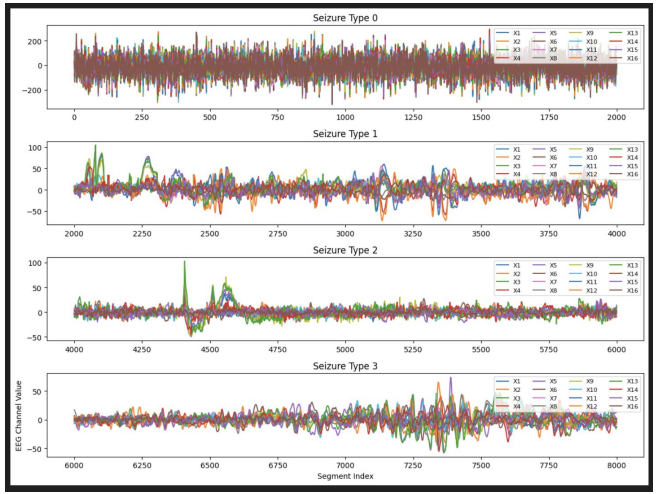
EDA - Channels visualizations

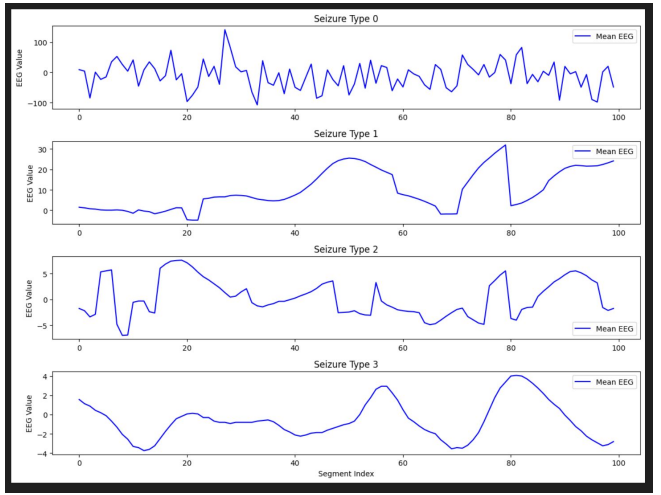
- Channels trends based on seizure types

  The trend of EEG channels varies across different seizure types. In particular, Type 0 (Healthy) shows patterns that are clearly distinct from the other types.

- Channels mean across individuals

  By averaging the values of all 16 channels (X1–X16) for each individual and plotting them sequentially, we can also observe that the mean channel trends differ across seizure types.





EDA - Statistical tests

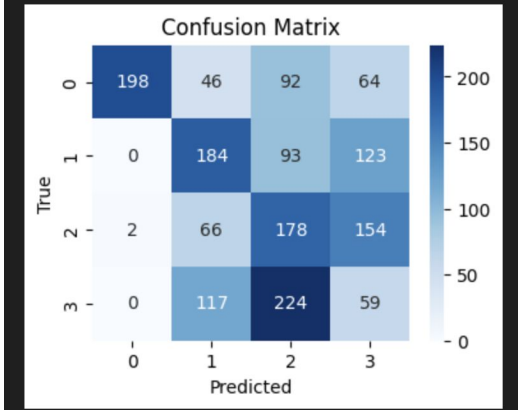
- The F-statistics for channels **X1, X9, X13, and X14** are greater than 100, indicating that these **channels show particularly strong distinctions across seizure types**.
- Moreover, the p-values for all channels are extremely small, confirming that **every channel has significant differences** between groups.
- Based on these statistical results, we should **avoid removing any channel during feature engineering**, as all of them carry meaningful discriminatory information.

```
=== Statistical Testing for Group Differences ===
X1: F-statistic = 105.942, p-value = 2.932e-67 (SIGNIFICANT)
X2: F-statistic = 76.877, p-value = 5.089e-49 (SIGNIFICANT)
X3: F-statistic = 57.400, p-value = 1.052e-36 (SIGNIFICANT)
X4: F-statistic = 50.306, p-value = 3.345e-32 (SIGNIFICANT)
X5: F-statistic = 87.202, p-value = 1.615e-55 (SIGNIFICANT)
X6: F-statistic = 70.327, p-value = 6.930e-45 (SIGNIFICANT)
X7: F-statistic = 54.223, p-value = 1.090e-34 (SIGNIFICANT)
X8: F-statistic = 56.994, p-value = 1.902e-36 (SIGNIFICANT)
X9: F-statistic = 100.722, p-value = 5.376e-64 (SIGNIFICANT)
X10: F-statistic = 84.886, p-value = 4.612e-54 (SIGNIFICANT)
X11: F-statistic = 67.741, p-value = 2.983e-43 (SIGNIFICANT)
X12: F-statistic = 60.427, p-value = 1.270e-38 (SIGNIFICANT)
X13: F-statistic = 104.087, p-value = 4.227e-66 (SIGNIFICANT)
X14: F-statistic = 101.372, p-value = 2.108e-64 (SIGNIFICANT)
X15: F-statistic = 75.528, p-value = 3.612e-48 (SIGNIFICANT)
X16: F-statistic = 82.385, p-value = 1.726e-52 (SIGNIFICANT)
```

Models Detail - Logistic Regression

- Logistic Regression is a linear model, it struggle to capture the complex patterns in EEG data, which can lead to lower performance.
- We can see that the confusion matrix indicates that the model has difficulty distinguishing between seizure type 1, 2, 3, leading to misprediction. This could be due to the feature between type 1, 2, 3 being quite similar, making it challenging for a linear model to separate them effectively.
- It achieved a relatively low accuracy of 47%.

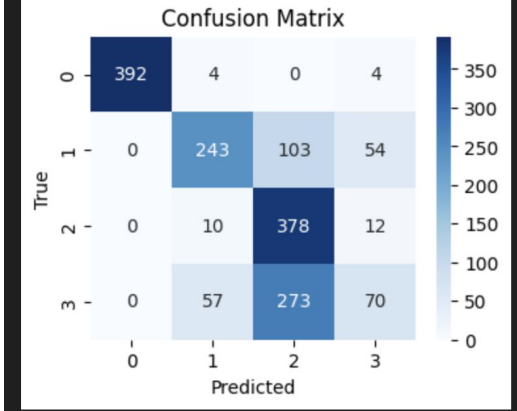
	precision	recall	f1-score	support
0	0.99	0.49	0.66	400
1	0.45	0.46	0.45	400
2	0.30	0.45	0.36	400
3	0.15	0.15	0.15	400
accuracy			0.39	1600
macro avg	0.47	0.39	0.41	1600
weighted avg	0.47	0.39	0.41	1600



Models Detail - SVC

- SVC performed better than Logistic Regression, likely due to its ability to handle non-linear decision boundaries.
- The confusion matrix shows that while SVC improves prediction accuracy, there are still misprediction between seizure types 1, 2 and 3.
- It only has accuracy with 69%.

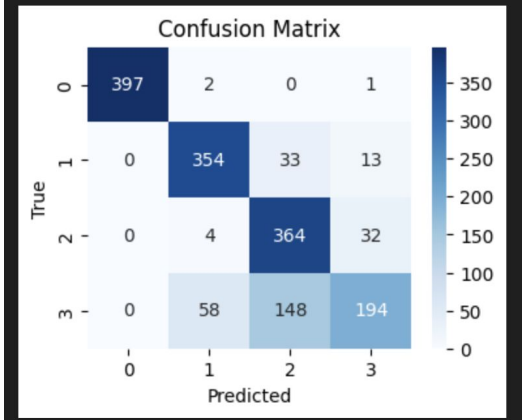
	precision	recall	f1-score	support
0	1.00	0.98	0.99	400
1	0.77	0.61	0.68	400
2	0.50	0.94	0.66	400
3	0.50	0.17	0.26	400
accuracy			0.68	1600
macro avg	0.69	0.68	0.65	1600
weighted avg	0.69	0.68	0.65	1600



Models Detail - Extra Trees Classifier

- ExtraTreesClassifier achieved the highest performance, likely due to its tree-based nature and ability to capture complex interactions between features.
- The confusion matrix indicates that ExtraTreesClassifier excels in accurately predicting all seizure types, with minimal mispredictions.
- It achieved 83% accuracy and 81% f1-score.

	precision	recall	f1-score	support
0	1.00	0.99	1.00	400
1	0.85	0.89	0.87	400
2	0.67	0.91	0.77	400
3	0.81	0.48	0.61	400
accuracy			0.82	1600
macro avg	0.83	0.82	0.81	1600
weighted avg	0.83	0.82	0.81	1600



Models Detail - Hyperparameter Tuning

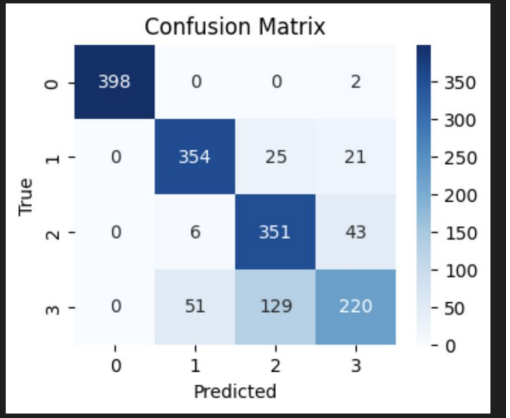
Because **ExtraTreesClassifier** has the highest accuracy, I want to perform hyperparameter tuning using GridSearchCV to optimize model performance.

- With Best Params:

  {'max_depth': None, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300}

- We get a higher f1-score with 82%.

	precision	recall	f1-score	support
0	1.00	0.99	1.00	400
1	0.86	0.89	0.87	400
2	0.70	0.88	0.78	400
3	0.77	0.55	0.64	400
accuracy			0.83	1600
macro avg	0.83	0.83	0.82	1600
weighted avg	0.83	0.83	0.82	1600



Results and Analysis

Model	Precision	Recall	F1-Score
Logistic Regression	0.47	0.39	0.41
SVC	0.69	0.68	0.65
ExtraTreesClassifier Hyper Tuning	0.83	0.83	0.82

According to this project, we have the following findings:

1. Statistical tests indicate that all EEG channels contribute significantly to class discrimination, suggesting that feature selection by simple removal may not be appropriate.
2. Linear model, like Logistic Regression, may not fully capture the nonlinear and irregular patterns of EEG signals, resulting in limited performance
3. Tree-based and ensemble methods achieved better performance, as they can better capture the complex and heterogeneous relationships among EEG features.

Conclusion

In this project, I successfully developed machine learning models to predict epileptic seizures using EEG data from the BEED dataset.

The ExtraTreesClassifier emerged as the best-performing model, achieving **83% precision, recall, and 82% F1-score**. This indicates that ensemble methods can effectively capture the complex patterns dataset, such as EEG signals, which are associated with different seizure types.

There are several areas for potential improvement and future work. We can use **deep learning approaches** which could potentially give better performance by capturing temporal and spatial patterns in EEG data more effectively.

Reference

- BEED: Bangalore EEG Epilepsy Dataset, UCI Machine Learning Repository
- Epileptic Seizure Detection and Analysis Using Machine Learning
- Scikit - ExtraTreesClassifier