# Confused student EEG brainwave data

...

PE

# Introduction


- Lecturers want immediate feedback for their MOOC(Massive open online course) videos.
- Usage of EEG (electro-encephalo-gram) is one solution for this.
- EEG signal is a voltage signal measured on the scalp.
- They used single channel EEG mindset (neurosky's) which is wireless to collect the data.

# Data Collection

- 10 videos
- Each video was 2 minutes long
- In each video of 2 minutes, the data of first 30 sec and last 30 sec are removed.

- Data is collected from 10 students, each watching 10 videos.
- Total of 100 data points.
- Each data point contains 120+ rows i.e; data is collected for every 0.5 seconds.
- After every session, student rates his confusion level from 1-7 .

# Columns in the dataset

| SubjectID | VideoID | Attention | Mediation | Raw | Delta | Theta | Alpha1 | Alpha2 | Beta1 | Beta2 | Gamma1 | Gamma2 | predefinedlabel | user-definedlabeln |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

| subject ID | age | ethnicity | gender |
|---|---|---|---|

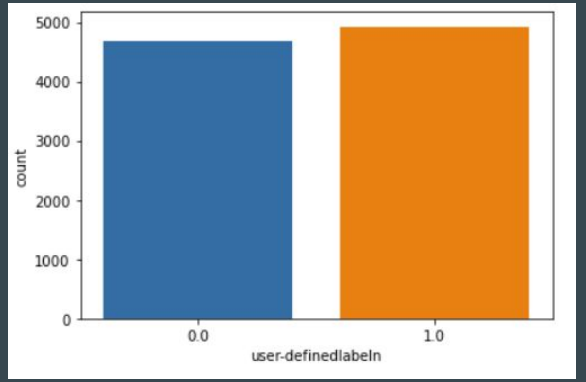
- Power spectrum
- delta (1-3Hz), theta (4-7 Hz)
- alpha-1 (lower 8-11 Hz), alpha-2 (higher 8-11 Hz)
- beta-1 (lower 12-29Hz), beta-2 (higher 12-29Hz)
- gamma-1 (lower 30-100 Hz), gamma-2 (higher 30-100 Hz)

# Preprocessing

- Used pandas dataframe for data storing input data.
- Dataframe contains 12811 rows.
- Did inner join between the two dataframes EEG_data and demographic_info on ‘SubjectID’.
- Performed one-hot encoding for categorical columns
  - ethnicity
  - gender
- Removed ‘SubjectId’, ‘VideoID’, ‘predefinedlabel’ columns from the dataframe as they are not useful in our prediction.

# Test-Train-Split

- Separated input and target labels into two dataframes x and y.
- Did test-train-split using parameters
  - test_size = 0.25
  - random_state = 42
  - stratify = y
- Visualized the target labels in the training dataset using sns.countplot to see whether the two classes present are in same proportion or not.
- Used standard scalar on input labels of train and test datasets.



# ML models

Models used are;

1. Naive bayes
2. Logistic regression
3. SVM
4. AdaBoost
5. XGBoost
6. Bagging
7. Decision Tree

# Features contribution

https://docs.google.com/document/d/1Mgo2V6A6LqLLNYmS-hS5v5uVHseDVqZI-F1DZHDnBV0/edit

Observations from the conducted experiments are;

1. The features Attention, Raw, Delta, Beta2, Gamma2, ethnicity, gender are very important for this classification task.
2. The feature Alpha2 contributes negatively as including it decreases the accuracy.
3. Remaining all features are also important for out task but not as important as above mentioned features.

# roc_auc_score for different models

| Model | ROC_AUC_SCORE |
|---|---|
| Naive Bayes | 0.6285047531135372 |
| Logistic Regression | 0.62581335085336 |
| SVM | 0.7334185275842884 |
| AdaBoost | 0.6659621982535635 |
| XGBoost | 0.7397647905204587 |
| Bagging | 0.6372960429344692 |
| Decision Tree | 0.6113497703227497 |

# References


1. https://www.kaggle.com/datasets/wanghaohan/confused-eeg
2. https://www.cs.cmu.edu/~kkchang/paper/WangEtAl.2013.AIED.EEG-MOOC.pdf
3. https://dl.acm.org/doi/pdf/10.1145/3107411.3107513