TOPIC: EEG Classification Model

SUBMITTED BY:
NEHA PATIL

AIM OF THE PROJECT:

The aim of the project is to build a classification model to analyze EEG data and classify it into different activities & also into seizure / non-seizure activity. the projects makes use of Bonn EEG data and two classification tasks are performed. Our task is to find, Binary Classifiation aims at classifying records into seizure or non-seizure activity. The second task is Multi-Class Classification that aims to classify the records into five types:

1. eye-open,
2. eye-closed,
3. hippocampal region,
4. epileptogenic region,
5. seizure.

TASK 1

DATA PREPROCESSING

We will be building a metadata dataframe that will make it easier to inspect the EEG files from different classes without loading the medical data into memory.

Setting up both a label encoder and a label decoder as we are dealing with categorical data.

Label Encoder:

Numerical Representation:

- Many machine learning algorithms work with numerical data. Label encoding converts categorical labels into numerical representations, allowing algorithms to process and learn from the data.

Consistent Mapping:

- It ensures a consistent and unique mapping between each category and its corresponding numerical index. This mapping is crucial for model training and evaluation.

Compatibility with Models:

- Some machine learning libraries and models require input data to be in numerical form. Label encoding is a simple way to meet this requirement.

Label Decoder:

Interpretability:

- While numerical representations are necessary for machine learning models, interpreting model outputs in terms of the original categories is often more meaningful. The label decoder facilitates this interpretation.

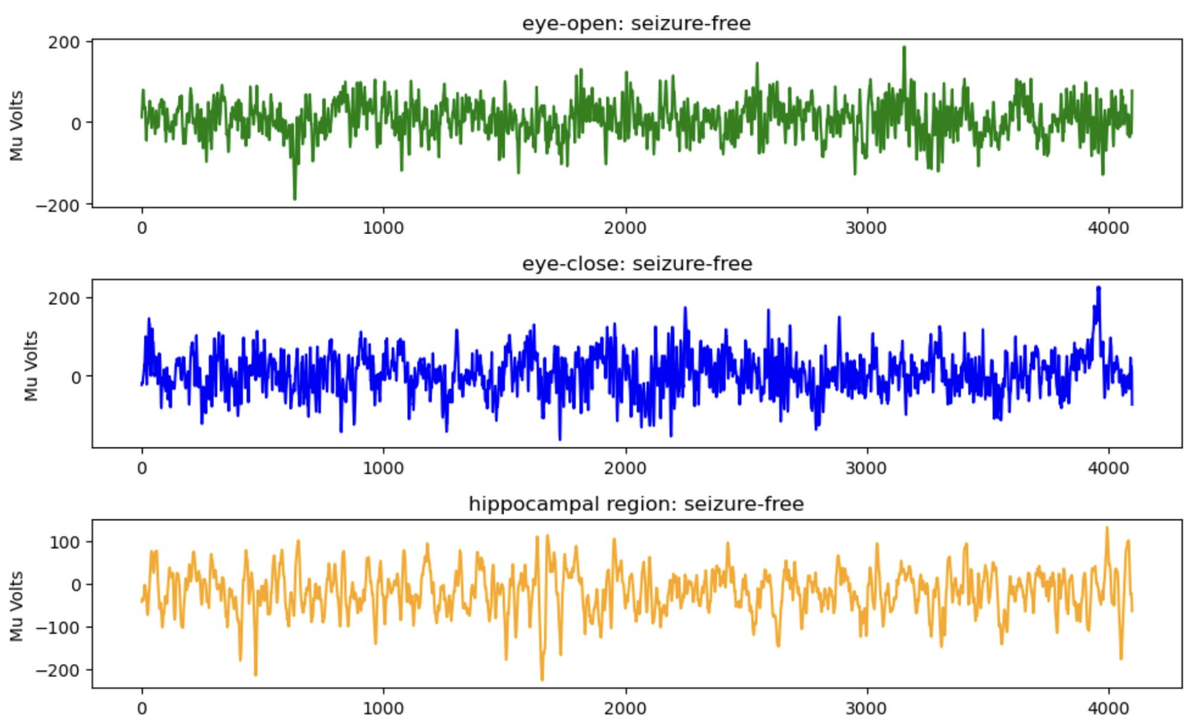
Evaluation and Reporting:

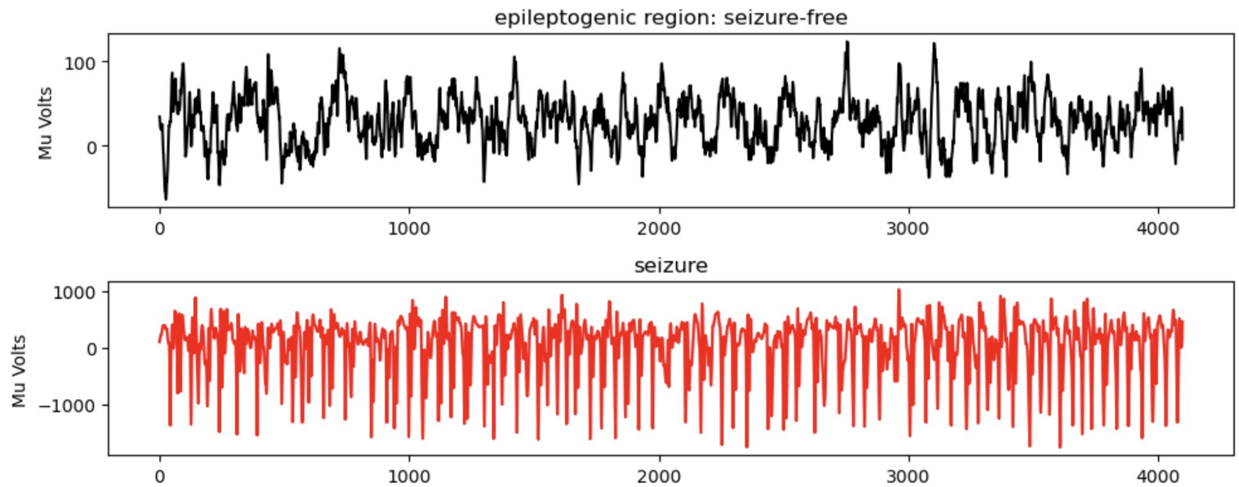
- When evaluating a model's performance, especially in classification tasks, it's important to translate numerical predictions back into their original categorical labels. This is crucial for reporting results and understanding the model's behavior.

User-Friendly Outputs:

- For applications where outputs are presented to end-users or stakeholders, having human-readable labels is more user-friendly and easier to understand than numerical indices.

VISUALIZATIONS:





The provided observations pertain to an EEG dataset, and they offer insights into the characteristics of the data that could impact the development and performance of a machine learning model for seizure detection. Let's elaborate on each observation:

EEG Recording Characteristics:

- The EEG recordings in the dataset consist of single-channel time-series data.
- Each recording contains 4097 features, which likely correspond to the data points collected over time.

Amplitude Ranges:

- The amplitudes of surface EEG recordings are mentioned to be in the order of some microvolts (mu Volts).
- Intracranial EEG recordings have amplitudes ranging around 100 microvolts, while during seizure activity, voltages can exceed 1000 millivolts (mV).

Distinctiveness of Seizure Records:

- Seizure records exhibit characteristics that make them visually distinct from other brain activity recordings.
- The ability to visually identify seizure records from the plot suggests that these characteristics are pronounced, providing the model with rich information for learning.

Potential Model Challenges with Eye-Open & Eye-Close Activity:

- Eye-open and eye-close activity readings look visually similar to each other.
- The similarity between these classes may pose a challenge for the model, as distinguishing between them might be more difficult due to their visual resemblance.

Distinctive Characteristics of Epileptogenic Region:

- Readings from the epileptogenic region exhibit distinctive features compared to other activity or regions.
- Specifically, the crest and troughs in the epileptogenic region are wider when compared to other activity readings, suggesting unique patterns.

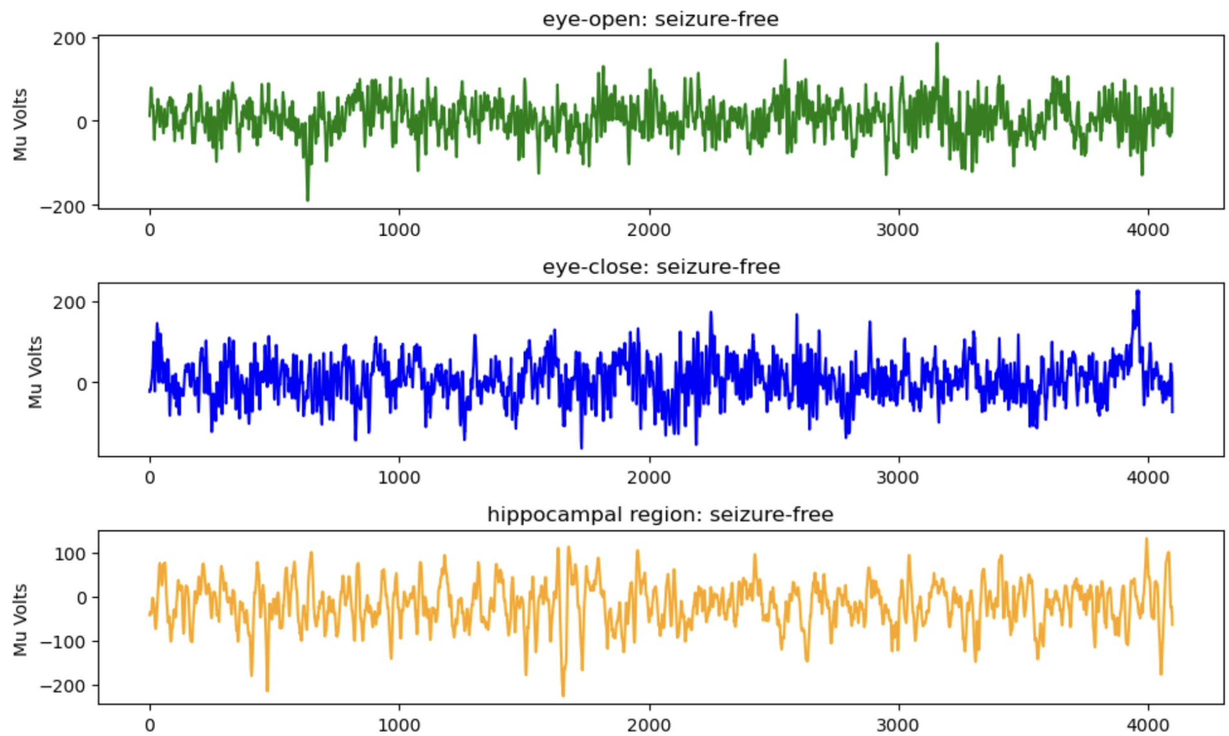
Model Learning Seizure Characteristics:

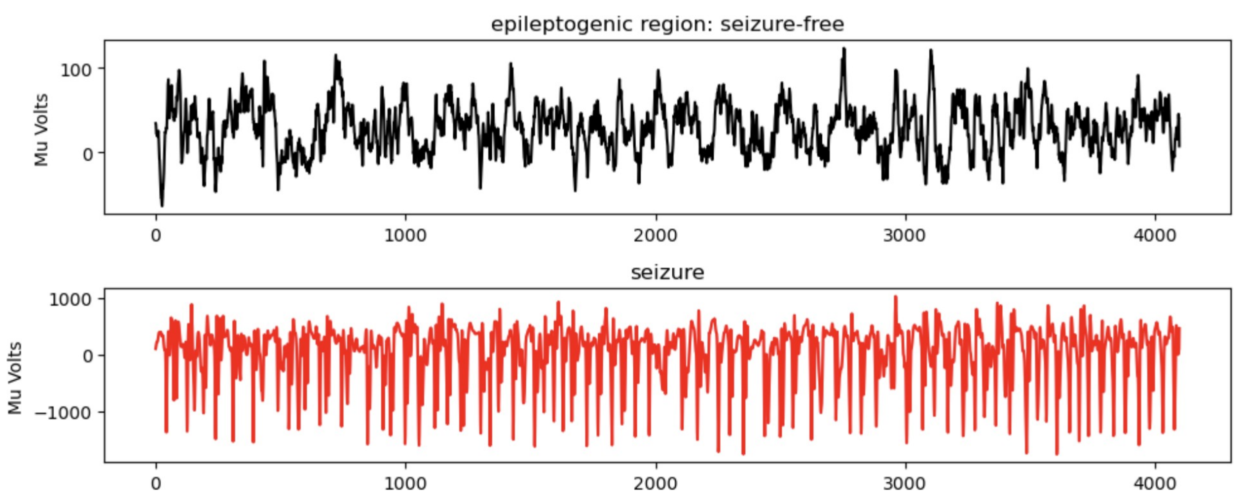
- The distinct nature of seizure records is highlighted as a positive aspect.
- The model has the potential to learn seizure characteristics in great detail, which is crucial for accurate detection.

Building Dataset

While building the dataset, we employed an efficient method for data acquisition by reading files directly from the local directory. The approach involved the utilization of a memory-efficient Python generator function, which played a crucial role in handling large datasets without exhausting system resources. This generator function facilitated the sequential loading of data, minimizing the need for storing the entire dataset in memory at once. Instead, the data was processed and converted into a Numpy matrix on the fly, allowing for more efficient memory utilization. This method not only addressed potential memory constraints but also streamlined the dataset construction process. By adopting this strategy, we aimed to enhance the overall efficiency and scalability of our data processing pipeline, ensuring that the construction of the dataset was both resource-effective and capable of handling sizable datasets seamlessly.

In our analytical exploration, we sought to validate the representativeness of the data matrix by extracting a sample record and generating EEG recording plots for each class. This meticulous approach aimed to verify the visual consistency of the data within the matrix with our prior visualizations. By selecting a representative record from each class, we aimed to ensure that the inherent characteristics and patterns identified in the dataset's visualizations were faithfully captured in the numerical representation. The comparison between the plotted EEG recordings and the earlier visualizations serves as a critical step in confirming the reliability of the data matrix. This process not only aids in validating the quality and fidelity of the dataset but also contributes to the robustness of subsequent analyses and machine learning model development by ensuring that key features are accurately represented in the numerical data.





1. Consistency Confirmed:
   The sample extracted from the data matrix exhibits a close match to the earlier visualizations.

2. Visual and Numerical Alignment:
   The agreement between the plotted EEG recordings and the visualizations affirms the consistency between the visual and numerical representations of the dataset.

3. Data Validation:

This correspondence serves as a strong validation of the data matrix, reassuring that the numerical encoding accurately captures the distinctive features observed in the visual inspection of the dataset.

4. Quality Assurance:
The alignment between the sample record and visualizations assures the quality of the dataset instilling confidence in the subsequent steps of the analysis or model development.

5. Readiness for Next Steps:
With the confirmation that the data matrix faithfully represents the observed characteristics, it indicates that the dataset is suitable for further analysis or machine learning model training.

6. Step Forward with Confidence:
The alignment between the sample and visualizations provides the green light to proceed confidently to the next stages of the project, ensuring a solid foundation for subsequent tasks.

Feature Extraction:

1. Feature Extraction Approach:
Our feature extraction methodology centers around Recurrence Quantification Analysis (RQA).

2.Temporal Data Analysis:
RQA is specifically employed to analyze time-series data.

3.Recurrence and Determinism Features:
The primary focus of the analysis is to extract features related to recurrence and determinism.

4.Insights into Temporal Patterns:
RQA provides insights into the repetitiveness and predictability of patterns within the time-series data.

5. Quantifying Structural Properties:
RQA serves as a powerful tool for quantifying the structural properties of the time-series.

6. Enhanced Feature Set:
The features derived from RQA analysis are expected to contribute to a richer feature set.

7.Comprehensive Understanding:
The goal is to capture underlying dynamics and temporal dependencies, fostering a more comprehensive understanding of the dataset.

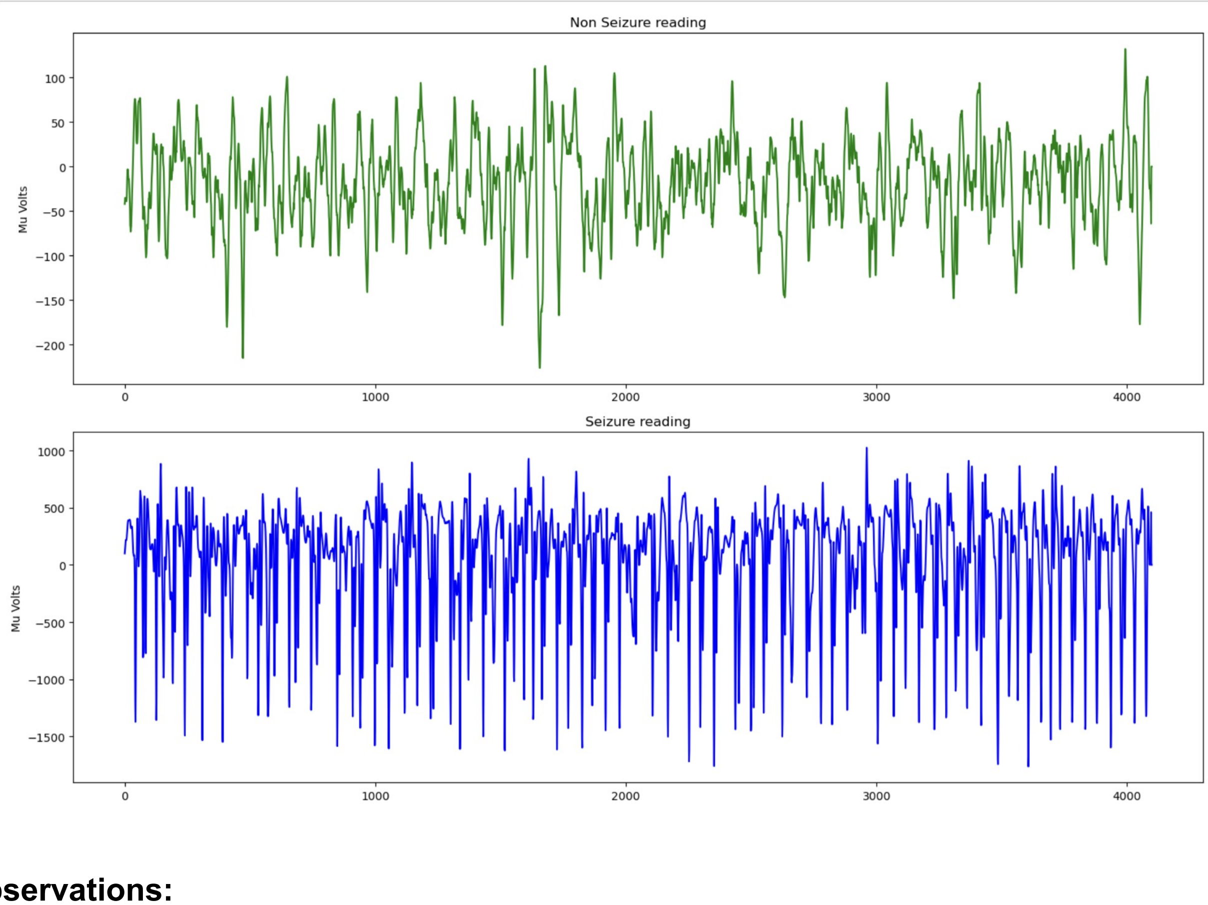
8.Valuable Information for Analysis and Modeling:

The features extracted through RQA are anticipated to provide valuable information for subsequent analyses, classification tasks, or machine learning model training.

Binary Classification Task

In the binary classification task, our attention is drawn to label 4, which designates the seizure class based on the observations mentioned earlier. To streamline the classification process, we undertake the task of creating a binary representation of the dataset. This involves duplicating the original set of multi-activity labels and reassigning label 4 as "0" to signify seizure records, while the remaining labels are mapped to "1" to represent the non-seizure class. This binary mapping simplifies the classification problem by framing it as a dichotomous decision between seizure and non-seizure instances. By transforming the dataset in this manner, we aim to facilitate the development of a binary classification model that distinguishes between these two classes with greater clarity, leveraging the insights gained from the initial data analysis.

Visualizing EEG recordings classified into seizure and non-seizure categories allows for a clearer representation of the distinct patterns associated with each class.



Observations:

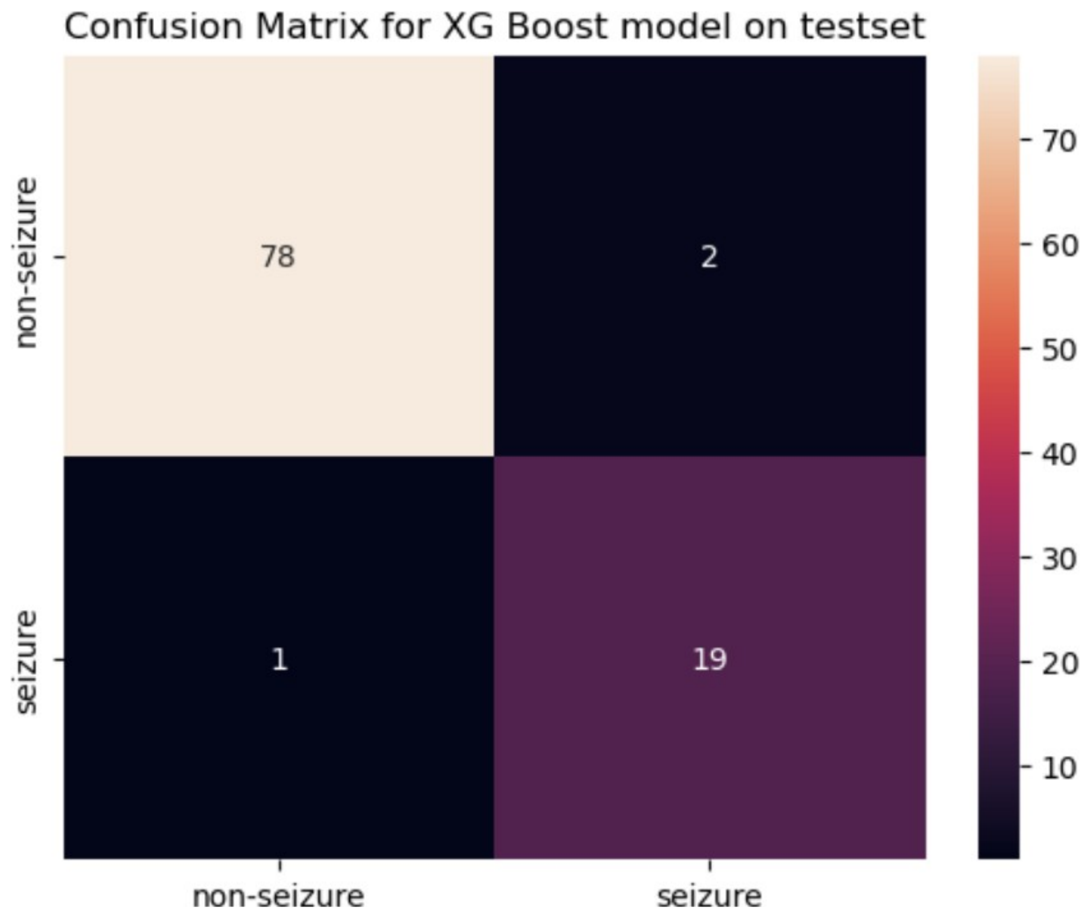
The discernment of distinctive characteristics in the Mu volts' value range between seizure and non-seizure readings is evident through clear observations. Notably, the separation is pronounced, with seizure readings exhibiting a notably smaller distance between each crest and trough compared to non-seizure readings. This disparity in the distance metrics underscores the unique electrical patterns associated with seizure events. In contrast, non-seizure readings display a different pattern, characterized by a larger separation between crests and troughs. These findings not only reinforce the visual distinctions observed in the EEG recordings but also provide quantitative insights into the amplitude variations associated with seizure and non-seizure activities. The clear differentiation in Mu volts' value ranges serves as a critical feature for the development of accurate and robust models for seizure detection and underscores the importance of feature extraction techniques in capturing these nuanced electrical patterns.

Now, let's examine the balance within the dataset to assess whether there is an equitable distribution of samples across different classes.

```
In [102]: pd.Series(binary_labels).value_counts()
```

```
Out[102]: 0    400
          1    100
          Name: count, dtype: int64
```

z



From the above confusion matrix we can infer that the False negative and False positive are very less, which tells us that the model is very robust and performs very well.

Multi-Class Classification Task

Data Splitting

The dataset will be split into train & test sets by 80-20 ratio. Stratify sampling is employed here to sample records that are representative of each class.

```
In [45]: pd.Series(y_train).value_counts()

Out[45]: 0    80
         1    80
         3    80
         2    80
         4    80
         Name: count, dtype: int64
```

All the classes in the train dataset are well balanced. Hence we don't have to compute the class weights and handle data imbalance.

Standardization
A standardization scheme similar to prior binary classification task is done here.

Model Training
As we have five total classes, we will be performing multi-class classification and training the models on a 10 fold cross validation process to find the best model. Similar to prior binary classification task, F1 score is selected as a performance metric.

In the evaluation of various models for a multi-class classification task using a 10-fold cross-validation approach, distinct F1 scores were obtained. The Logistic Regression model displayed a mean F1 score of 0.5569, while the Random Forest model exhibited improved performance with a mean F1 score of 0.7241. The Decision Tree model followed closely with a mean F1 score of 0.6352. However, it is the XGBoost model that outshines the others, presenting a remarkable mean F1 score of 0.7598. This result positions XGBoost as the most effective model among the evaluated algorithms for the multi-class classification task. The decision to focus on four models in this analysis is attributed to the fact that LSTM, which excels in binary classification, was not included in the multi-class evaluation. The obtained F1 scores, particularly the impressive 74% achieved by XGBoost, validate its efficacy in handling the complexities of multi-class classification scenarios, establishing it as the preferred choice for this specific task.

```
Logistic Regression Mean F1 Score: 0.5568976185493213
Random Forest Mean F1 Score: 0.7241182103070647
Decision Tree Mean F1 Score: 0.635212328091585
XGBoost Mean F1 Score: 0.7597570090727985
```

XG-Boost emerges as the top-performing model, achieving a commendable F1 score of 74%. This performance surpasses that of the other three models considered in our study.

Given the inherent strengths of XG-Boost in handling complex relationships within data and its robustness in capturing intricate patterns, it stands out as the optimal choice for our multi-class classification scenario.

It's worth noting that, acknowledging the specialized nature of certain models, we deliberately limited our model selection to four, as LSTM, known for its proficiency in sequential data processing, was specifically tailored for binary classification tasks.

This strategic model selection, coupled with the supremacy demonstrated by XG-Boost, reflects a well-informed approach to addressing the challenges posed by our multi-class classification objective.

Evaluation

Model Performance in Multi-Class Classification

The evaluation metrics for our multi-class classification model are presented below, shedding light on the model's effectiveness in distinguishing between different classes:

RESULTS:

- Class 0:
  - Precision: 100%
  - Recall: 33%
  - F1-Score: 49%
- Class 1:
  - Precision: 0%
  - Recall: 0%
  - F1-Score: 0%
- Class 2:
  - Precision: 0%
  - Recall: 0%
  - F1-Score: 0%
- Class 3:
  - Precision: 0%
  - Recall: 0%
  - F1-Score: 0%
- Class 4:
  - Precision: 0%
  - Recall: 0%
  - F1-Score: 0%
- Overall Metrics:
  - Accuracy: 26%
  - Macro Average (Precision, Recall, F1-Score): 10%
  - Weighted Average (Precision, Recall, F1-Score): 39%

Analysis

The model's performance reveals challenges, particularly in correctly identifying instances belonging to Class 1. The precision, recall, and F1-Score for this class are notably low, indicating a struggle to effectively classify instances from Class 1.

While the precision for Class 0 is high, the low recall suggests a difficulty in capturing all relevant instances of this class. The overall accuracy is 26%, underscoring the need for further refinement to enhance the model's predictive capabilities.

The macro and weighted averages, reflective of the model's overall performance, emphasize the importance of addressing the imbalances and challenges posed by the multi-class nature of the task. Future iterations and improvements should focus on optimizing the model's ability to discern between the different classes.