**Spring 2023**

**"Tabular Data In The Wild: A Study In Health Informatics"**

**Objective/Problem Statement:** The primary objective was to analyze a vast dataset of patient records, including demographic information and various data sources, to identify relevant features associated with Long COVID symptoms. The project aimed to create interpretable algorithms for predicting outcomes at the time of COVID-19 diagnosis and hospitalization.

**Experiments (Spring 2023):** The project leveraged data from the N3C enclave and employed supervised machine learning techniques. While the detailed results of experiments were not provided, the focus was on creating models that could predict Long COVID outcomes and identifying significant features associated with the condition.

**Key Learnings (Spring 2023):**

1. **Relevance of Data:** The project discovered that not every file or column in the N3C dataset was relevant for Long COVID prediction. This highlights the importance of feature selection and data preprocessing.
2. **Label Change:** The dataset no longer included the 'Long COVID' label but did have information related to patient deaths, which was used to create a 'deceased' label.
3. **Multiple Records per Patient:** One patient could have multiple conditions, observations, and drugs per visit (occurrence or era). This complexity in the data required careful handling.
4. **Encoding:** Encoding the fields as binary or numerical had no significant impact on modeling performance across three data frames, indicating the robustness of the chosen encoding methods.
5. **Aggregation Decisions:** Early aggregation decisions had a noticeable impact on the outcome, emphasizing the importance of data preprocessing and aggregation strategies.
6. **Age Prediction:** The project successfully used a Logistic Regression model to predict missing age values, contributing to data completeness.
7. **Concept Selection:** Concepts with more than 1 million patients per concept were selected for the 'sample,' representing a substantial portion of the patient population.
8. **Total Features:** The number of concepts and represented features were crucial factors considered in the analysis, as they determined the scope of the study.
9. **Model Performance:** The Logistic Regression model was employed to predict negative and positive labels, while Random Forest models primarily predicted low probabilities of being a positive case (deceased). Gradient Boosting models excelled in predicting both classes. The use of the 'deceased' label achieved higher scores than using the Long COVID label.
10. **Enclave Optimization:** In terms of the N3C Enclave, overall runtime was reduced, and transformations ran smoothly without interrupting or corrupting other processes.

**Summer 2023**

**"Long COVID Prediction with Custom Loss Functions and Boosting Models for Imbalanced Datasets"**

In Summer 2023, the project shifted its focus to predicting severe Long COVID symptoms in females following the COVID-19 pandemic. The work conducted during this semester includes:

**Objective/Problem Statement:** The primary objective was to predict severe Long COVID symptoms in a specific demographic group: females aged 15-55. The challenge addressed was dealing with imbalanced datasets, where the target class was significantly imbalanced.

**Experiments:** The experiments in Summer 2023 were extensive and focused on addressing class imbalance. Key experiments included:

* Under sampling and oversampling techniques to mitigate class imbalance.
* The creation of a balanced sample dataset using a combination of under sampling and stratified sampling.
* Model comparisons between XGBoost, LightGBM, CatBoost, and Gradient Boosting.
* Introduction of custom loss functions to influence sensitivity-specificity balance.

**Main Takeaways:** The Summer 2023 project demonstrated the effectiveness of addressing class imbalance through sampling techniques and custom loss functions. It also highlighted that different models exhibited distinct prediction capabilities, emphasizing the importance of model selection in addressing specific challenges.

In summary, the work conducted in the past semesters has progressed from understanding Long COVID and identifying features in Spring 2023 to addressing class imbalance and improving predictive models, particularly for a specific demographic, in Summer 2023. These projects have built upon each other, with the Summer project extending and applying the knowledge gained from the Spring project to tackle a more specific problem and implement various experiments to address class imbalance effectively.

**Experiments (Summer 2023):**

1. **Gradient Boosting Models**: The project utilized gradient boosting models, including XGBoost, LightGBM, and CatBoost, to predict severe Long COVID symptoms. These models were chosen for their effectiveness in handling complex datasets.
2. **Custom Loss Functions**: Custom loss functions were explored to address the challenge of class imbalance in the dataset. The use of custom loss functions is indicative of efforts to improve model performance, especially on the minority class (severe Long COVID).
3. **Sampling Techniques**: Various sampling techniques were experimented with, including under sampling and oversampling. Under sampling involved randomly removing instances from the majority class, while oversampling involved generating synthetic instances of the minority class.
4. **Evaluation Metrics**: To assess the performance of the models, the project employed a set of standard evaluation metrics, including Precision, Recall, F1 Score, and AUC-ROC. These metrics were used to evaluate how well the models predicted severe Long COVID in females.

**Main Takeaways (Spring and Summer 2023):**

* In Spring 2023, the focus was on data analysis and the development of predictive models for Long COVID. Detailed results were not provided, but it laid the groundwork for subsequent experiments.
* In Summer 2023, the project continued its work by specializing in predicting severe Long COVID in females. It addressed class imbalance issues through custom loss functions and explored various sampling techniques. Evaluation metrics were used to assess model performance.
* The experiments in Summer 2023 built upon the foundation of Spring 2023, with a more targeted objective and a focus on addressing imbalanced datasets to improve predictive accuracy. Together, these two phases contributed to a comprehensive study of Long COVID prediction and model refinement.

**Fall 2023: "Comparative Analysis of Deep Learning and Gradient Boosting Models on Tabular Data"**

**Objective/Problem Statement:** The objective was to conduct a comparative analysis of deep learning and gradient boosting models' performance on tabular data.

**Experiments:**

* Conducted a literature review on deep learning and GBDT models in tabular data analysis.
* Utilized real-world healthcare data from the N3C Enclave.
* Explored data pre-processing, feature selection, and model implementation.
* Performed a comparative analysis between deep learning and GBDT models.
* Investigated interpretability techniques using SHAP and LIME.

**Main Takeaway:** This project built on the previous two by providing a systematic comparison of deep learning and GBDT models on tabular data. It emphasized the importance of selecting the right model for the given task and highlighted the role of interpretability techniques. Additionally, it showcased the integration of real-world healthcare data, further enhancing the project's practical applicability.

**Graph-Based Methods in Health Informatics: A Journey from Long COVID Prediction to Model Comparison**

In the realm of health informatics, graph-based methods offer a promising avenue for exploring complex relationships within patient data and clinical records. These methods leverage the inherent structure of healthcare information to extract valuable insights, improve predictive models, and enhance patient care. Here, we outline some potential graph-based methods and applications within the context of Long COVID prediction and beyond.

1. Patient Interaction Graphs:

- Create a graph where patients are nodes, and edges represent interactions or shared comorbidities.

- Analyze the graph to identify clusters of patients with similar healthcare journeys.

- Identify potential influencers or super-spreaders of medical concepts.

2. Symptom Co-occurrence Network:

- Build a network where symptoms are nodes, and edges represent the co-occurrence of symptoms in patients.

- Identify patterns of symptom clusters and their associations with Long COVID.

3. Medication Interaction Graphs:

- Create graphs that connect patients who have taken similar medications.

- Study the influence of medication usage on Long COVID outcomes.

- Detect potential drug interactions and side effects.

4. Network Propagation Models:

- Implement network propagation algorithms to model the spread of information or diseases.

- Simulate the impact of interventions or policy changes on Long COVID prevalence.

- Optimize resource allocation in healthcare systems.

5. Knowledge Graphs:

- Create a knowledge graph that connects medical concepts, treatments, and patient outcomes.

- Use the graph to support clinical decision-making by providing evidence-based recommendations.

- Improve medical literature mining and knowledge discovery.

6. Graph Neural Networks (GNNs):

- Utilize GNNs to capture complex dependencies within patient data.

- Train GNNs to predict Long COVID risk and severity.

- Leverage GNNs for feature extraction and representation learning.

7. Predictive Modeling with Graph Features:

- Enhance predictive models by incorporating graph-derived features.

- Utilize features like node centrality, clustering coefficients, and graph embeddings to improve model accuracy.

- Combine graph-based features with traditional patient data for holistic predictions.

Graph Neural Networks (GNNs):

Capturing Complex Dependencies:

* + GNNs excel in capturing the intricate web of dependencies and interactions between patients, healthcare providers, and medical events.
  + By modeling patients as nodes and their interactions as edges in a graph, GNNs can unveil hidden patterns and connections that traditional methods may miss.

Predicting Long COVID Risk and Severity:

* + One of the primary applications of GNNs in health informatics is their ability to predict Long COVID risk and assess its severity.
  + GNNs can learn from historical patient data, uncovering subtle relationships between pre-existing conditions, treatment strategies, and Long COVID outcomes.

Feature Extraction and Representation Learning:

* + GNNs provide a unique advantage in feature extraction and representation learning.
  + They can automatically derive meaningful features from the patient interaction graph, potentially revealing critical factors contributing to Long COVID risk.