Data Challenge: Predictive Modeling for Mortality Risk

Welcome to the data challenge! This task involves developing predictive models to assess the risk of in-hospital mortality based on a patient's clinical and demographic data. Your goal is to train a variety of models, from traditional machine learning to state-of-the-art transformer architectures, and compare their performance.

---

1. Data Description

The dataset provided is a synthetic representation of patient data, inspired by the **MIMIC-III** and **MIMIC-IV** datasets. It contains information for **5,000** patients spread across three key tables. All tables can be joined using `subject_id` and `hadm_id`.

Table : patients_table

This table contains basic patient demographic information.

Column	Description
`subject_id`	Unique identifier for each patient.
`gender`	Patient's gender.
`anchor_age`	Age of the patient at a specific reference point.
`anchor_year`	The calendar year corresponding to the anchor age.
`anchor_year_group`	A group of years used for data anonymization.
`dod`	Date of death (if applicable).

Table : patient admissions

This table provides detailed information about hospital stays.

Column	Description
`subject_id`	Unique patient identifier.
`hadm_id`	Unique identifier for each hospital admission.

`admittime`	Date and time of hospital admission.
`dischtime`	Date and time of hospital discharge.
`deathtime`	Date and time of death during the hospital stay. This column is crucial for the target variable.
`admission_type`	Type of admission (e.g., Emergency, Urgent).
`admission_location`	Location from which the patient was admitted.
`discharge_location`	Location to which the patient was discharged.
`insurance`	Patient's insurance provider.
`marital_status`	Patient's marital status.
`race`	Patient's self-reported race.
`hospital_expire_flag`	**The target variable:** A binary flag (0 or 1) indicating if the patient died during their hospital stay. This will be your primary outcome to predict.
Other columns	Additional demographic and visit-specific details (e.g., provider, language).

`Table : disease_diagnosis_code`

This table contains the diagnoses associated with each hospital admission, coded using the **International Classification of Diseases (ICD)**.

Column	Description
`subject_id`	Unique patient identifier.
`hadm_id`	Unique hospital admission identifier.
`seq_num`	Sequence number for the diagnoses within an admission.
`icd_code`	The ICD code for a diagnosis.
`icd_version`	The version of the ICD code (e.g., 9 or 10). **Note:** These codes need to be harmonized to a single version (e.g., ICD-9) for consistency.

2. Task Description 📈

Your primary task is to build a predictive model to predict two outcomes

1. the `hospital_expire_flag` for each patient, signifying in-hospital mortality. This is a **binary classification problem**.
2. Predict the length of `hospital stay`.
3. Is there a difference in prediction performance between different ethnic groups?
4. **Final Presentation:** Using the result prepare a 10 mins short presentation summarising your data analysis approach, choice, including a description of your data preprocessing steps, model architectures, and key findings.

To demonstrate your expertise, you must perform the following:

1. Data Preparation and Feature Engineering

- **Data Integration:** Merge the three tables (`patients_subset`, `admissions_subset`, and `diagnosis_icd_subset`) into a unified dataset.
- **ICD Code Unification:** The `diagnosis_icd_subset` table contains a mix of **ICD-9** and **ICD-10** codes. You must harmonize these codes to a single version (preferably **ICD-10**) to ensure consistency.

- **Feature Engineering:** Create meaningful features from the raw data.
    - **Temporal Features:** Extract features from timestamps (e.g., length of stay from `admittime` and `dischtime`).
    - **Categorical Features:** Encode categorical variables like `gender`, `race`, and `admission_type`.
    - **Diagnosis Features:** Group ICD codes into clinical categories (e.g., using **Clinical Classifications Software (CCS)**) or use them to create a unique patient journey (sequence of events).
    - **Target Variable:** Use `hospital_expire_flag` from the `admissions_subset` table as your prediction target.

2. Model Development & Comparison

You are required to build and evaluate **at least three distinct types of models**:

- **Traditional Machine Learning Model:** Train an **XGBoost** or **Decision Tree** model.
- **Sequential/Recurrent Model:** Train a model capable of handling longitudinal data, such as a **Long Short-Term Memory (LSTM)** network.

- **Transformer Model:** Develop a **transformer-based model** to capture the long-range dependencies within the sequence of patient diagnoses and other events. This model should be specifically designed to handle the sequential nature of a patient's clinical history.
- **Fine-tuned Large Language Model (LLM):** If you can, fine-tune an existing LLM to predict the risk of mortality.

3. Evaluation & Reporting

- **Metrics:** Evaluate all models using appropriate classification metrics such as **AUC-ROC**, **precision**, **recall**, and **F1-score**.
- **Performance Comparison:** Present a clear comparison of the models' performance. Discuss the strengths and weaknesses of each model type in the context of this problem. Explain why one model might outperform another.