Advanced Data Analytics – Final Project Preliminary Report

Name: Aakash Tripathi

Data: Explain your dataset here.

The chosen dataset for this endeavor is sourced from The Cancer Genome Atlas (TCGA), a comprehensive resource housing multi-modal data encompassing gene expression data, clinical information such as age, tumor stage, treatments, and vital status, as well as medical images such as MRIs and CT scans. This dataset offers a rich tapestry of information, presenting a unique opportunity to delve into a multifaceted analysis of factors influencing cancer patient survival.

Analysis: Explain your planned analysis here.

The goal of this project lies in harnessing the power of Transformer models to develop a predictive framework for assessing the survival chances of cancer patients based on the multi-modal data available. The analysis will be structured as follows:

1. Data Preparation: Organizing the multi-modal data (genomic, clinical, and imaging data) into a structured format conducive for analysis. Next, conduct an exploratory data analysis (EDA) to understand the data's inherent structure, missing values, and potential correlations between different data modalities.
2. Model Architecture: Designing a custom Transformer architecture capable of ingesting and processing multi-modal data. Employing techniques such as multi-head attention to capture intricate relationships between different data modalities and features.
3. Model Training and Evaluation: Partitioning the data into training, validation, and test sets to ensure a rigorous evaluation of the model's performance. Employing appropriate loss functions and metrics, such as the concordance index, to gauge the model's efficacy in predicting survival probabilities.
4. Hyperparameter Tuning and Optimization: Conducting a systematic search for optimal hyperparameters to enhance the model's predictive accuracy. Exploring advanced optimization techniques to expedite the training process and improve generalization.