Data Science Assessment

Objective: The main objective of this assessment was to develop an AI assistant to support clinical decision-making by: (1) predicting whether a patient will be readmitted to the hospital within 30 days, and (2) extracting and categorizing key information from free-text discharge notes. The goal was to combine structured data modeling and NLP techniques to provide actionable insights while ensuring interpretability and clinical relevance.

Task 1 – Predicting 30-Day Readmissions: For Task 1, we leveraged structured patient data to build a binary classification model. Initial exploratory data analysis revealed trends such as higher readmissions for younger patients and males, and longer average stays per admission correlating with readmission risk. We engineered two new features (age_bin and stay_per_admission) to capture clinically meaningful patterns and applied one-hot encoding for categorical variables. Due to the imbalance in the target variable, SMOTE was used to synthesize minority class samples in the training set. We evaluated 4 models which are Logistic Regression, Decision Tree, Random Forest, and XGBoost using ROC-AUC, F1 score, and confusion matrices. The Decision Tree was selected as the best-performing model due to its balance of predictive performance and interpretability with ROC_AUC - 0.621 F1 Score - 0.5 and accuracy at 65%. Key predictive features included stay_per_admission, length_of_stay, num_previous_admissions, age_bin_senior, and gender_male. These results suggest that patients with longer or frequent hospitalizations are at higher risk of 30-day readmission.

Task 2 – Extracting and Categorizing Discharge Note Information: For Task 2, we implemented a two-step NLP pipeline. First, Flan-T5 was used for named entity extraction to identify clinical entities such as diagnoses, symptoms, treatments, medications, and follow-up instructions. Second, GPT-4 categorized these extracted entities into structured fields. The two-step approach improves modularity, interpretability, and allows the model to capture subtle distinctions between entity types. The pipeline successfully converted unstructured notes into structured data, facilitating downstream analysis and reporting. Key risks and limitations include potential hallucinations, ambiguity in entity categorization, sensitivity to negations (e.g., “no complications”), and reliance on general-purpose LLMs that are not clinically certified. Despite these limitations, the approach provides a rapid and practical method for structuring discharge notes for clinical use.

Practical Implications and Next Steps: The models and NLP pipeline developed here can help clinical teams identify high-risk patients for readmission and structure discharge information for easier review. This can inform interventions, improve patient follow-up, and reduce readmission rates. With more time or data, the predictive models could benefit from hyperparameter tuning, feature selection, and experimentation with ensemble methods. For the NLP component, we could explore domain-specific models (e.g., BioClinicalBERT) to improve extraction accuracy, better handle negations, and reduce hallucinations. Additionally, integrating this pipeline with hospital EHR systems could allow for real-time risk assessment and automated discharge note structuring.