Clinical Data Science & LLM Assessment – Final Report

1. Approach: Predictive Modelling:

- Explored and cleaned the data, including class balance checks and summary statistics.
- Selected XGBoost for binary classification, performed feature engineering by binning age, label-encoding categorical variables, and addressing class imbalance using SMOTE.

  Evaluated model performance with confusion matrix, ROC AUC, F1-score, and additional metrics (precision, recall, Cohen's Kappa, Matthews Correlation Coefficient).
- Identified feature importances to highlight the most influential predictors.

NLP / LLM Entity Extraction:

- Applied both a standard Hugging Face NER pipeline (dslim/bert-base-NER) and a prompt-based LLM approach (Flan-T5) to extract clinical entities from free-text discharge notes.
- Used custom prompt engineering to encourage Flan-T5 to extract and categorize relevant clinical information (diagnosis, treatment, symptom, medication, follow-up action).

2. Key Results: Predictive Model: Model: XGBoost classifier after SMOTE oversampling.

- **Metrics on Test Set: Accuracy:** 0.55, **F1-score:** 0.47, **ROC AUC:** 0.57, **Precision:** 0.38, **Recall (Sensitivity):** 0.62, **Matthews Correlation Coefficient:** 0.13, **Cohen's Kappa:** 0.12
- **Confusion Matrix:** The model captures some true positives but also makes frequent false positives and false negatives, reflecting the small and challenging dataset.
- **Feature Importance:** age, length of stay & diagnosis code were most influential predictors

LLM/NLP Entity Extraction:

- **Standard NER Model:** Did not extract meaningful clinical entities from discharge notes, as expected (model is trained on general English, not medical data).
- **Prompt-Based Flan-T5:** Produced structured but generic output. With prompt engineering, it could extract entity categories, but produced placeholder text than specific clinical terms

3. Practical Implications

- The current predictive model offers only limited accuracy and reliability, reflecting the small dataset and the need for richer features (e.g., more granular clinical history, lab results)
- General-purpose LLMs and NER models require further adaptation (domain-specific training)

4. Next Steps with more Time/Data

- **Increase Data Volume:** Acquire more historical patient records and richer features, such as detailed clinical events, comorbidities, or medication changes over time.
- **Advanced Feature Engineering:** Integrate information from discharge notes using clinical NLP or embedding methods (e.g.BioBERT) & experiment with additional feature combinations
- **Model Optimization:** Perform systematic hyperparameter tuning and test alternative models (Random Forest, ensemble learning, deep learning).
- **Domain-specific LLM:** Use clinical language models for prediction, entity extraction
- **End-to-End Pipeline:** For automated prediction and clinical information extraction.

  Overall, this project demonstrates the end-to-end data science workflow for clinical applications, highlights both technical successes and limitations, and outlines clear paths for future enhancement.