Data Science & LLM Technical Assessment

1. Approach

Part 1: Predictive Modelling
We developed a binary classification model to predict whether a patient would be readmitted to the hospital within 30 days

Key steps included:

- Data cleaning and preprocessing (eg: handling missing values, encoding categorical variables)
- Exploratory analysis and feature engineering (Bivariate analysis with target variable , correlation matrix, new feature creation like high risk flag)
- Training multiple models (Random Forest,Logistic Regression)
- Evaluation using AUC-ROC, F1-score, and confusion matrix

Part 2: Named Entity Recognition (NER)

- Used the **d4data/biomedical-ner-all** model to extract clinical entities from free-text discharge notes.
- Extracted entity types like Disease_disorder, Lab_value, Clinical_event, Sign_symptom, etc.
- Grouped and split outputs into separate structured columns for interpretability (e.g: Diagnosis, Symptoms, Treatment. etc).

---

2. Key Results

Part 1: Model Performance

	Model	Accuracy	F1 Score	Recall	Precision	ROC AUC
0	Random Forest	0.550	0.182	0.154	0.222	0.467
1	Random Forest + SMOTE	0.500	0.231	0.231	0.231	0.442
2	Logistic Regression	0.375	0.359	0.538	0.269	0.376

- **Logistic Regression** performed best in identifying readmitted patients, with the highest recall and F1 score.
- This is important because **minimizing false negatives** (missed readmissions) is critical in a healthcare setting.
- Random Forest had slightly higher AUC but worse recall.

Part 2 – NER Output

Mild reaction to medication. Switched to alternative treatment.	[('mild', 'Severity')]
Patient discharged with minor discomfort. Advised rest and hydration.	[('discharged', 'Clinical_event'), ('minor', 'Severity'), ('rest', 'Therapeutic_procedure'), ('hydra'
Patient showed improvement. Prescribed antibiotics for 5 days.	[('antibiotics', 'Medication'), ('5', 'Duration')]
Patient showed improvement. Prescribed antibiotics for 5 days.	[('antibiotics', 'Medication'), ('5', 'Duration')]
Patient discharged with minor discomfort. Advised rest and hydration.	[('discharged', 'Clinical_event'), ('minor', 'Severity'), ('rest', 'Therapeutic_procedure'), ('hydra'
No further signs of infection. Resume normal diet and activity.	[('normal', 'Lab_value'), ('diet', 'Diagnostic_procedure')]
Patient showed improvement. Prescribed antibiotics for 5 days.	[('antibiotics', 'Medication'), ('5', 'Duration')]
Discharge after recovery from pneumonia. No complications observed.	[('discharge', 'Clinical_event'), ('complications', 'Disease_disorder')]
Good recovery trajectory. Follow-up scan scheduled next month.	[]
Symptoms controlled. Monitoring for relapse advised.	[('symptoms', 'Sign_symptom'), ('re', 'Therapeutic_procedure'), ('##lapse', 'Therapeutic_proc
Patient discharged in stable condition. Recommend follow-up in 2 weeks.	[('discharged', 'Clinical_event'), ('stable', 'Lab_value'), ('follow', 'Clinical_event'), ('2', 'Date')]
Patient showed improvement. Prescribed antibiotics for 5 days.	[('antibiotics', 'Medication'), ('5', 'Duration')]
Patient discharged with minor discomfort. Advised rest and hydration.	[('discharged', 'Clinical_event'), ('minor', 'Severity'), ('rest', 'Therapeutic_procedure'), ('hydra'

- These were successfully split into structured fields for analysis.

---

3. Practical Implications

- The readmission model helps prioritize **at-risk patients at discharge** for proactive care.
- NER transforms unstructured discharge summaries into structured data that can be used in dashboards or clinical summaries.
- Together, these tools support **better care planning and resource allocation**.

---

4. More time on Data

- Engineer richer features using visit history, multiple existing health conditions, and medications over time(Time since last admission, Number of admissions in the past 6 or 12 months),demographics
- Use a transformer model like BERT to convert discharge notes into features the model can learn from
- Try other language models like BioBERT or GPT-style models to improve how the notes are understood
- Clean and improve the extracted entities using simple post-processing rules.
- The small dataset (~200 patients) made it difficult for both the prediction and text extraction models to perform reliably, with more data, both parts could improve significantly.