Phase 1 — Project Initiation & Setup

Project Title

Predicting 30-Day Readmission for Diabetic Patients

Project Proposal

Problem Description:
Hospital readmissions within 30 days are a major driver of healthcare costs and indicate gaps in patient care. For diabetic patients, early readmission prediction enables hospitals to apply targeted interventions (education, medication adjustment, follow-up scheduling) to reduce readmission rates and costs. The goal: **build an interpretable predictive model** to flag patients at high risk of readmission within 30 days of discharge.

Why it matters:
Reducing avoidable readmissions improves patient outcomes and saves significant healthcare expenditure. Interpretability is essential because clinicians must understand why a patient is predicted to be at risk.

Dataset:
- **Name:** Diabetes 130-US hospitals for years 1999–2008.
- **Source:** Kaggle

Target: Predict whether a patient will be readmitted within 30 days (`readmitted == '<30'`) — binary classification: *readmit_30d = 1* if `<30`, else *0* (includes `>30` and `NO`).

** Methodology:** 1. Data collection: download dataset from Kaggle. 2. Preprocessing: handle missing values, remove/aggregate identifiers, encode categoricals, feature engineering (e.g., prior visits count, medication change flags). 3. EDA: distributions, correlations, missingness, class balance. 4. Modeling: baseline logistic regression + tree-based model(s) — Random Forest, XGBoost or GradientBoosting; evaluate with accuracy, precision, recall, F1, ROC-AUC, confusion matrix; use cross-validation. 5. Interpretability: feature importance, logistic coefficients, permutation importance; optionally SHAP if resources permit. 6. Deliverables: cleaned dataset snapshot, EDA plots, model scores, confusion matrix, feature importance plots, written actionable recommendations.

Tools & Libraries: Python 3.x, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.

Phase 2 — Data Collection & Pre-processing

Data Cleaning Plan

1. Inspect target variable and map to binary label (`readmit_30d`).
2. Drop clearly irrelevant identifiers (e.g., `encounter_id`, `patient_nbr`) for modeling — keep an anonymized ID mapping if needed for traceability.
3. Handle missing values:
   - Determine columns with many missing values — consider dropping if >50% missing.
   - For categorical missing values, fill with `'Unknown'` or `'Missing'`.
   - For numerical missing values, use median imputation.

4. Deduplicate rows if any duplicates exist.
5. Convert categorical variables:
   - For high-cardinality categorical columns consider label encoding or grouping infrequent categories into `'Other'`.
   - Use one-hot encoding for nominal features with manageable cardinality.
6. Feature engineering:
   - Transform `readmitted` to binary.
   - Aggregate medication change flags into features such as `num_med_changes` if available.
   - Create binary flags for `num_lab_procedures` being high/low, or `time_in_hospital` categories if helpful.
7. Split dataset into training and testing sets after processing.

Cleaning notes / rationale:
- We converted the multi-class `readmitted` into binary to focus on 30-day readmissions (clinical priority).
- Identifiers removed to prevent leakage.
- Missing values imputed conservatively (median or 'Missing') to retain rows.
- We reduced cardinality on diagnosis codes to avoid exploding one-hot features and overfitting.

Phase 3 — Exploratory Data Analysis (EDA) & Visualization

We'll produce key visualizations: class balance, distributions of numeric features, boxplots for outliers, correlation heatmap for numeric features, and categorical count plots for important categorical features.

EDA Observations:

- Class imbalance: Around *15%* of patients are readmitted within 30 days.
- `time_in_hospital` distribution: typically skewed toward short stays.
- Some numeric features (e.g., `num_medications`, `num_lab_procedures`) show outliers — consider robust scalers or capping for modeling.
- Correlation heatmap: identify pairs with moderate correlation; none should be extremely collinear but check `num_lab_procedures` vs `num_procedures`.
- Categorical variables like `admission_type_id` or `discharge_disposition_id` have clear dominant categories — these may carry predictive power.

Phase 4 — Model Building & Evaluation

Plan: 1. Prepare features and target. 2. Split into train/test. 3. Build two baseline models: - Logistic Regression (interpretable baseline) - Random Forest Classifier (nonlinear, robust) 4. Evaluate with accuracy, precision, recall, F1, and ROC-AUC. 5. Compare and choose best model; show confusion matrix and feature importances.

Model Selection Rationale:
- Logistic Regression provides interpretability (coefficients show directionality).
- Random Forest often yields better predictive performance on structured tabular data and can capture nonlinearities and interactions.
- Use evaluation metrics (precision, recall, F1, ROC-AUC) to choose model depending on

the clinical objective: if identifying all at-risk patients is vital (minimize false negatives), prioritize **recall**; if avoiding unnecessary interventions is important (minimize false positives), prioritize **precision**. For healthcare readmission, high recall often matters since missing a high-risk patient can have severe consequences, but this must be balanced with available resources.

Key Findings

- **Model performance:** Random Forest achieved *ROC-AUC = 65.82*, *Recall = 0.44*, *Precision = 71.42*, *F1 = 0.87* on the held-out test set. Logistic Regression achieved *ROC-AUC = 65.73*, *Recall = 56.53*, *Precision = 17.19*, *F1 = 26.36* .

Actionable Insights & Recommendations

1. **Targeted post-discharge interventions:** Flag patients with high predicted readmission risk for intensive post-discharge follow-up (telehealth calls within 7 days, appointment scheduling within 14 days, medication reconciliation).
2. **Medication review:** Patients with many medication changes during admission could benefit from medication counselling and early follow-up.
3. **Care pathways:** Develop care pathways for patients identified as high-risk (e.g., discharge case manager involvement).
4. **Monitoring & Re-training:** Periodically re-train model on newer data, and monitor fairness across demographic groups (age, gender, race) to avoid biased interventions.