Details

� � Project Title: Hospital Patient Records Cleanup Domain: Healthcare Level: Final Year B.Tech / Data Science Professionals Difficulty: High Use Case: Prepare hospital patient data for epidemiological research and predictive analytics � � Project Objective: To clean, curate, and prepare hospital data (patients, admissions, discharges, diagnoses, medications) for accurate insights and modeling, while ensuring data quality, regulatory compliance (e.g., HIPAA), and readiness for analytics workflows. � � Provided Dataset: A multi-sheet Excel or CSV-based mock dataset simulating: 1. patients.csv – Patient demographic details 2. visits.csv – Admission/discharge data 3. diagnoses.csv – ICD-10 diagnosis codes 4. medications.csv – Prescribed drugs per visit 5. staff.csv – Doctor/nurse assignments 6. hospital_info.csv – Hospital unit metadata (Downloadable from sources like CMS Inpatient Dataset) � � Exercises and Tasks by Phase � � 1. Understand the Business Context ● Identify the objective: Predict readmission risks, analyze comorbidities ● Define KPIs: Avg Length of Stay (LOS), Readmission rate, Mortality rate ● Document data use-cases: Resource optimization, care quality improvement ● Consider constraints: HIPAA compliance, missing discharge notes � � 2. Data Discovery and Ingestion ● Load CSVs or Excel sheets using pandas.read_csv() or Power BI import ● Validate sources, file sizes, data types ● Record metadata: file name, rows, columns, date range ● Store a version-controlled raw copy � � 3. Data Profiling (Exploration & Assessment) Use pandas-profiling or Sweetviz on each dataset. ✓ Identify: ● Nulls (e.g., missing gender, DOB) ● Invalid entries (e.g., age > 120, DOB > admission) ● Duplicate patient IDs or visit IDs ● ICD-10 code format mismatches (regex: [A-Z][0-9][0-9A-Z]) ● High-cardinality fields (e.g., Notes, Diagnoses) � � Deliverable: Data Profiling Report for all 6 sheets. � � 4. Schema Alignment & Standardization ● Standardize column names to snake_case ● Normalize date formats (admission_date, discharge_date) ● Recode gender: M, F, O → Male, Female, Other ● Standardize drug names (e.g., brand → generic) ● Use dictionaries to convert ICD codes to disease names � � 5. Data Cleaning ● Impute missing age using DOB ● Remove duplicate patient records based on name + DOB + hospital ● Drop invalid entries: negative LOS, discharge before admission ● Flag logic issues: medication prescribed before admission ● Clean special characters in free-text notes ● Remove extreme outliers in billing or stay duration � � 6. Data Integration and Merging ● Merge patients + visits on patient_id ● Merge diagnoses and medications on visit_id ● Merge staff using attending_physician_id or unit_id ● Handle conflicting visit IDs (use suffixes _x, _y) ● Resolve entity duplicates using fuzzy matching on names + DOB � � 7. Data Transformation ● Derive length_of_stay = discharge_date - admission_date ● Generate binary flags: is_readmitted, is_high_risk ● Bucket ages: 0–18, 19–35, 36–60, 60+ ● One-hot encode admission types (Emergency, Scheduled, Transfer) ● Normalize lab result fields (if present) � � 8. Data Validation and Quality Checks ● Assert: No negative LOS, DOB ≤ admission ● Compare admission counts before and after cleaning ● Create validation tests using pytest or Great Expectations ● Run sanity checks: unique patient_id per row in patients.csv ● Log failed validations to a validation_report.txt � � 9. Documentation and Data Dictionary Create a Data Dictionary with: Column Name Type patient_id string Description Unique patient identifier admission_date date icd_code string Hospital entry date ICD-10 diagnostic code medication_name string Drug prescribed is_readmitted Example P12345 2021-06-23 E11.9 Metformin boolean True if readmitted within 30 days True � � 10. Export, Deployment, and Handoff ● Export cleaned, curated data to: o clean_patients.csv o curated_visits.csv ● Deploy to PostgreSQL or BigQuery (optional) ● Create a ZIP of: o Final CSVs o Data dictionary (Excel or Markdown) o Validation report ● Provide summary in a handover document (PDF or Notion) � � Tools to Use Task Ingestion Profiling Cleaning Integration Validation Export Tool pandas, Excel, SQL pandas-profiling, Sweetviz pandas, OpenRefine pandas.merge, fuzzywuzzy Great Expectations, pytest to_csv(), Power BI, Excel Documentation Markdown, Notion, Excel � � Final Deliverables 1. ✅ Cleaned and Merged Dataset (CSV or Excel) 2. ✅ Data Profiling Report 3. ✅ Data Dictionary 4. ✅ Validation and Cleaning Summary Report 5. ✅ README.md or Handover PDF � � 11. Predictive Analytics Tasks (Regression, Association & Decision Tree) In this final phase, machine learning techniques are applied on the curated hospital data to generate insights for risk prediction and operational optimization. � � Regression Analysis: - Objective: Predict Length of Stay (LOS) based on patient demographics, diagnosis, and admission type. - Model: Linear Regression, Ridge, or Lasso Regression - Features: age, diagnosis_code, admission_type, medications_count - Target: length_of_stay - Metrics: RMSE, MAE, R^2 � � Association Rule Mining: - Objective: Discover frequent co-occurrence patterns of diagnoses and medications. - Technique: Apriori algorithm or FP-Growth - Use Case: Generate rules like “IF diagnosis = E11 AND medication = Metformin THEN likely readmission.” � � Decision Tree Classification: - Objective: Classify whether a patient is likely to be readmitted within 30 days. - Model: Decision Tree, Random Forest, Gradient Boosted Trees - Features: age_group, length_of_stay, medication_count, admission_type - Target: is_readmitted - Metrics: Accuracy, Precision, Recall, F1-Score, AUC Deliverables: - Clean dataset with engineered features for modeling - Jupyter notebooks or Python scripts for each method - Confusion matrix, ROC curve, and feature importance analysis - Interpretation of findings with healthcare implications