Healthcare Analytics

Dataset Documentation Report

Comprehensive Data Dictionary & Preprocessing Guide

Project	Healthcare Predictive Analytics
Total Records	5,611 records across 4 datasets
Date Generated	December 2024
Purpose	ML-based Readmission Risk Prediction

Table of Contents

1. Executive Summary

This document provides comprehensive documentation for the Healthcare Analytics dataset used in the patient readmission risk prediction system. The dataset comprises four interconnected CSV files containing patient demographics, physician performance metrics, department statistics, and financial data spanning from 2022 to 2024.

Key Highlights:

- 5,611 total records across 4 primary datasets
- 3-year time span: January 2022 - December 2024
- 17 unique departments tracked
- 110 physicians with performance metrics
- 10 insurance providers represented
- Used for machine learning readmission prediction

2. Dataset Overview

Dataset	Records	Columns	Primary Use
patient_demographics.csv	1,001	7	Patient characteristics & outcomes
physician_performance.csv	3,960	10	Doctor performance metrics
department_metrics.csv	612	10	Department-level statistics
financial_performance.csv	36	10	Hospital financial data

All datasets are interconnected through common identifiers (department_id, time periods) enabling comprehensive cross-dataset analysis for healthcare insights.

3. Patient Demographics Dataset

File: patient_demographics.csv | Records: 1,001

The primary dataset used for machine learning predictions. Contains aggregated patient demographics with associated healthcare outcomes including length of stay, costs, and readmission rates.

Data Dictionary:

Column	Type	Description	Values/Range
age_group	Categorical	Patient age range	0-17, 18-34, 35-49, 50-64, 65+
gender	Categorical	Patient gender	M, F
insurance_type	Categorical	Insurance provider	10 providers (Aetna, Cigna, etc.)
patient_count	Integer	Number of patients in group	12 - 270
avg_length_of_stay	Float	Average hospital stay (days)	2.0 - 8.0
avg_cost	Float	Average treatment cost ($)	3,000 - 13,000
readmission_rate	Float	Rate of readmission	0.10 - 0.25 (10%-25%)

Insurance Providers:

Aetna, Anthem, Blue Cross Blue Shield, Cigna, Health Net, Humana, Kaiser Permanente, Medicare, Medicaid, UnitedHealthcare

Statistical Summary:

Metric	Min	Max	Mean	Std Dev
patient_count	12	270	156.8	62.4
avg_length_of_stay	2.0	8.0	4.9	1.7
avg_cost ($)	3,017	12,977	7,458	2,314
readmission_rate	0.10	0.25	0.175	0.044

4. Physician Performance Dataset

File: physician_performance.csv | Records: 3,960

Monthly performance metrics for 110 physicians across all departments. Tracks key quality indicators including patient satisfaction, complication rates, and revenue.

Data Dictionary:

Column	Type	Description
physician_id	String	Unique identifier (PHY000-PHY109)
physician_name	String	Full name with title
month	Integer	Month (1-12)
year	Integer	Year (2022-2024)
total_patients	Integer	Patients treated that month
avg_length_of_stay	Float	Average stay in days
avg_satisfaction_score	Float	Patient satisfaction (1-5)
complication_rate	Float	Rate of complications
readmission_rate	Float	Patient readmission rate
avg_revenue	Float	Average revenue per patient

Key Metrics Ranges:

Metric	Min	Max	Mean
total_patients	12	50	31
avg_satisfaction_score	3.6	5.0	4.3
complication_rate	5.0%	19.0%	11.0%
readmission_rate	10.0%	24.0%	17.0%
avg_revenue ($)	6,000	24,000	14,500

5. Department Metrics Dataset

File: department_metrics.csv | Records: 612

Monthly operational metrics for 17 hospital departments. Provides insights into resource utilization, capacity management, and departmental performance.

Data Dictionary:

Column	Type	Description
department_id	String	Unique identifier (DEPT000-DEPT016)
department_name	String	Full department name
month	Integer	Month (1-12)
year	Integer	Year (2022-2024)
total_admissions	Integer	Monthly admissions count
avg_length_of_stay	Float	Average patient stay (days)
avg_cost	Float	Average cost per admission
total_revenue	Float	Monthly department revenue
occupancy_rate	Float	Bed occupancy rate (0-1)
nurse_patient_ratio	Float	Nurses per patient

Departments List:

Emergency Medicine, Cardiology, Orthopedics, Neurology, Oncology, Pediatrics, Internal Medicine, Surgery, Psychiatry, Radiology, Anesthesiology, Obstetrics, Gastroenterology, Dermatology, Urology, Nephrology, Pulmonology

6. Financial Performance Dataset

File: financial_performance.csv | Records: 36

Monthly hospital-wide financial metrics spanning 3 years. Tracks revenue, expenses, profitability, and key financial health indicators.

Data Dictionary:

Column	Type	Description
month	Integer	Month (1-12)
year	Integer	Year (2022-2024)
total_revenue	Float	Monthly revenue ($)
total_expenses	Float	Monthly expenses ($)
net_income	Float	Revenue minus expenses ($)
operating_margin	Float	Profitability ratio
bad_debt	Float	Uncollectable debt ($)
charity_care	Float	Free care provided ($)
insurance_contractual	Float	Insurance adjustments ($)
cash_on_hand	Float	Available cash ($)

Financial Summary (3-Year Period):

Metric	Total/Average	Range
Total Revenue	$709M (3 years)	$14M - $25M/month
Total Expenses	$575M (3 years)	$12M - $21M/month
Net Income	$134M (3 years)	$1.4M - $5.6M/month
Operating Margin	18.8% avg	7.6% - 27.8%
Cash on Hand	$5.8M avg	$3.3M - $8.2M

7. Data Preprocessing Pipeline

The following preprocessing steps transform raw data into ML-ready features:

Step 1: Data Loading & Validation

- Load CSV files using pandas.read_csv()
- Validate column names and data types
- Check for missing values (confirmed: 0 missing)
- Verify data integrity and consistency

Step 2: Target Variable Creation

- Binary classification: High vs Low readmission risk
- Threshold: readmission_rate > 0.20 → High Risk (1)
- Distribution: 33% High Risk, 67% Low Risk
- Creates class imbalance requiring SMOTE

Step 3: Categorical Encoding

- Label Encoding using sklearn.preprocessing.LabelEncoder
- age_group: 5 categories → 0-4
- gender: 2 categories → 0-1
- insurance_type: 10 categories → 0-9
- Encoders stored for inverse transformation

Step 4: Train-Test Split

- Split ratio: 75% training, 25% testing
- Stratified split preserving class distribution
- Random state: 42 for reproducibility
- Training: 750 samples, Testing: 251 samples

8. Feature Engineering

Final Feature Set (6 Features):

Feature	Type	Encoding	Description
age_group_encoded	Categorical	Label (0-4)	Patient age category
gender_encoded	Categorical	Label (0-1)	Patient gender
insurance_type_encoded	Categorical	Label (0-9)	Insurance provider
patient_count	Numerical	None	Group size
avg_length_of_stay	Numerical	None	Hospital stay duration
avg_cost	Numerical	None	Treatment cost

SMOTE Class Balancing:

- Problem: Class imbalance (67% Low Risk, 33% High Risk)
- Solution: Synthetic Minority Over-sampling Technique (SMOTE)
- Method: Generate synthetic samples for minority class
- Parameters: k_neighbors=5, random_state=42
- Result: Training set balanced to 50%-50% (503 samples each)
- Original training: 750 → Balanced training: 1,006 samples

Feature Importance (from trained models):

Rank	Feature	RF Importance	XGB Importance
1	avg_cost	25.8%	28.1%
2	patient_count	22.4%	21.7%
3	avg_length_of_stay	19.2%	18.9%
4	insurance_type_encoded	14.1%	13.8%
5	age_group_encoded	10.8%	10.2%
6	gender_encoded	7.7%	7.3%

9. Data Quality & Validation

Quality Checks Performed:

- Missing Values: 0 across all datasets
- Duplicate Records: None detected
- Data Type Consistency: All columns correctly typed
- Range Validation: All values within expected bounds
- Referential Integrity: IDs properly linked across datasets

Data Validation Rules:

Column	Validation Rule	Status
readmission_rate	0 ≤ value ≤ 1	✓ Pass
occupancy_rate	0 ≤ value ≤ 1	✓ Pass
satisfaction_score	1 ≤ value ≤ 5	✓ Pass
operating_margin	-1 ≤ value ≤ 1	✓ Pass
age_group	Valid categories only	✓ Pass
gender	M or F only	✓ Pass

Data Completeness:

Dataset	Expected Fields	Populated	Completeness
patient_demographics	7	7	100%
physician_performance	10	10	100%
department_metrics	10	10	100%
financial_performance	10	10	100%

10. Conclusions

Dataset Strengths:

- Comprehensive: Covers patients, physicians, departments, and finances
- Time Series: 36 months of historical data for trend analysis
- Clean: No missing values or data quality issues
- Balanced Features: Mix of categorical and numerical variables
- ML-Ready: Preprocessed and encoded for immediate model training

Preprocessing Summary:

- Raw data: 5,611 records across 4 CSV files
- Primary ML dataset: 1,001 patient demographic records
- Features used: 6 (3 categorical, 3 numerical)
- Target: Binary readmission risk (threshold: 20%)
- Class balancing: SMOTE (750 → 1,006 training samples)
- Final split: 1,006 training, 251 testing samples

Recommendations for Future Work:

- Add clinical features: diagnoses, procedures, medications
- Include temporal features: seasonality, trends
- Integrate physician-patient linkage for deeper analysis
- Consider time-series models for trend prediction
- Expand to multi-class risk stratification (Low/Medium/High)

— End of Dataset Documentation Report —