

Healthcare Dataset Report

This report summarizes the synthetic MSK-IMPACT-style dataset generated for encrypted cancer prediction experiments. The dataset mimics genomic feature patterns described in the original 2017MSK-IMPACT study but contains no real patient data.

Total Patients: 3500

Patients Labeled as Cancer: 250

Patients Labeled Non-Cancer: 3250

Patients were labeled as cancer or non-cancer using a probability model designed to simulate oncogenic genomic patterns: High CNA extremes (large positive/negative alterations) High mutation burden Detection of gene fusion events Elevated structural variation count Each factor increases the probability that a sample is classified as cancer-positive. This allows meaningful testing of encrypted logistic regression models.

Financial Dataset Report

This report describes the synthetic financial transaction dataset used for encrypted fraud detection testing.

Total Transactions: 499

Fraudulent Transactions: 76

Legitimate Transactions: 424

Fraud Logic Overview:

Fraud was intentionally injected using a probability model that considers: High transaction amount (>\$300) Large geographical distance (> 50 km) Low device trust score (< 0.3) Multiple chargebacks (≥ 2) Each factor contributes to the likelihood of a transaction being marked as fraudulent.

Academic Dataset Report

This report describes the synthetic academic performance dataset used for encrypted prediction testing.

Total Students: 100

Students Labeled 'At Risk': 36

Students in Good Standing: 64

Risk Assessment Logic:

Academic risk was generated using a weighted probability model based on realistic performance indicators: Low GPA (< 2.5) Low weekly study hours (< 8) Poor attendance (attendance rate < 0.80) High number of missing assignments (≥ 4) Insufficient sleep (< 6 hours per night) Each factor increases the probability of a student being labeled as academically at risk.