This report documents the preprocessing of dataset to make a binary classification model To detect the pancreatic cancer early using routine blood and urine tests

Dataset: 600 patients (256 healthy, 344 cancer)
Features: 7 routine clinical parameters
Output: Clean, normalized, balanced data ready for machine learning

Original Diagnosis:

- Diagnosis 1 → Healthy / No cancer (256 patients)
- Diagnosis 2 → Early-stage cancer (214 patients)
- Diagnosis 3 → Advanced-stage cancer (130 patients)

Binary Target Created:

- **Class 0 (Healthy):** Diagnosis 1 → 256 patients (42.7%)
- **Class 1 (Cancer):** Diagnosis 2 & 3 → 344 patients (57.3%)

Preprocessing Steps

1-2: Data Loading and Target Creation

- **Loaded 600 patients successfully**
- **Created binary target has cancer (0=Healthy, 1=Cancer)**
- **No missing values or duplicates found**

Step 3: Feature Selection

- **Selected 7 routine clinical parameters**
- **Age, sex, creatinine, bilirubin, glucose, urine volume, urine**
- **Rationale: Affordable, accessible, non-invasive**

Step 4: Categorical Encoding

- **Encoded gender: Male=1, Female=0**
- **New feature: sex_encoded**

Step 5: Train-Test Split

- **Split ratio: 80% training, 20% testing**
- **Stratification: Yes (maintains class distribution)**
- **Random state: 42 (reproducible)**

Results:

- **Training: 480 samples (205 healthy, 275 cancer)**
- **Testing: 120 samples (51 healthy, 69 cancer)**

Step 6: Feature Normalization

- **Method: StandardScaler (Z-score normalization)**
- **Formula: z = (x - μ) / σ (μ = mean of feature , σ = standard deviation of feature both from training data)**
- **Result: All features scaled to mean≈0, std≈1**
- **Critical: Scaler fitted on training data only (prevents data leakage)**

Step 7: Class Balancing with SMOTE

- Imbalanced training set (42.7% vs 57.3%)
- **Applied to:** Training data only
- **Result:** Balanced training set (50:50)

After SMOTE:

Training Set:

Healthy: 275 (50. 0%)

Cancer: 275 (50.0%)

Total: 550 (+70 synthetic samples)

Results

1-preproccesed dataset

dataset	Samples	features	Distribution	status
Training	550	7	275 healthy, 275 cancer (50:50)	Balanced and Normalized
Testing	120	7	51 healthy, 69 cancer (42.5:57.5)	Original and Normalized

2- Training statistics (features)

Feature	**Mean**	**Std**	**Min**	**Max**
age	0.00	1.00	-2.45	2.38
sex_encoded	0.00	1.00	-1.23	0.81
creatinine	0.00	1.00	-2.12	2.56
bilirubin	0.00	1.00	-1. 89	2.74
glucose	0.00	1.00	-2.34	2.45
urine_volume	0.00	1. 00	-2.18	2.31
urine_pH	0.00	1.00	-2.06	2.15