UAE Cancer Outcomes: Data Analysis & Machine Learning

CDS 2413 - CRN: 14350 - Group: 3 - Members: Aisha Alshamsi (H00535685), Sara Almulla (H00532724)

1. Introduction

We analyze a real UAE cancer cohort (N=10,000) to profile patients and build baseline models for outcomes. Targets: Deceased (Yes/No) and Survival_Months (diagnosis→death; deceased only). We use clean, reproducible pipelines to report descriptives, run core statistical tests, and train simple predictive models.

2. Objectives

01 Describe cohort and key variables (age, stage, type, treatment).

02 Test associations (numeric–numeric; categorical–categorical).

03 Classify Deceased (Yes/No) with LR, KNN, NB, DT.

04 Forecast Survival_Months via Multiple Linear Regression.

05 Report clear metrics and practical takeaways.

3. Data & Pipeline

- Outcome balance: Deceased 992 (9.92%) vs Not Deceased 9,008 (90.08%).
- Derivation: Survival_Months = (Death_Date − Diagnosis_Date)/30.44 (deceased only).
- Preprocessing: One-Hot (categoricals), StandardScaler (numerics), pairwise deletion for missing.
- Splits: 80/20 (stratified for classification), random_state=42; 5-fold CV on train.

4. Descriptive & Statistical Tests

- Survival_Months (n=992): mean 17.51, median 17.71, sd 9.07, IQR 15.36; near-symmetric (skew −0.079, kurt −1.156).
- Representativeness: one-sample t-test (sample n=150): $t \approx 0.50$, $p \approx 0.618$ → sample ≈ population.
- Age ↔ Survival_Months: Pearson 0.006 (p=0.859); Spearman 0.005 (p=0.8776) → no meaningful correlation.
- Stage × Deceased: $\chi^2(3)=7.027$, p=0.071; Cramér's V=0.027 → very small.

5. Modeling Results (Classification + Regression)

- Classification (Deceased): Pipeline (OHE+Scaling, stratified CV, class_weight="balanced").
- Best overall: Logistic Regression (highest ROC-AUC with balanced precision/recall). DT interpretable; KNN/NB trailed.
- Simple Regression: $\hat{y} = \beta_0 + \beta_1 \cdot \text{Age}$; $\beta_1 \approx 0$ → age alone not predictive.
- Multiple Linear Regression (Survival_Months):
- CV: MAE 7.89±0.33, RMSE 9.33±0.35, R^2 −0.074±0.064.
- Test: MAE 7.83, RMSE 9.12, R^2 0.045 vs baseline MAE 8.12, RMSE 9.36.
- Top |coef|: Cancer_Type (Lung/Liver/Leukemia), Gender_Other, Hypertension.

6. Conclusion & QR

Cohort well-characterized; no strong age–survival link; stage–death association is weak.

LR is the most reliable baseline classifier; MLR lowers error slightly but R^2 remains low.

Next: more clinical features; use KM/Cox for time-to-event.

QR: code & results (EDA, classifiers, MLR).

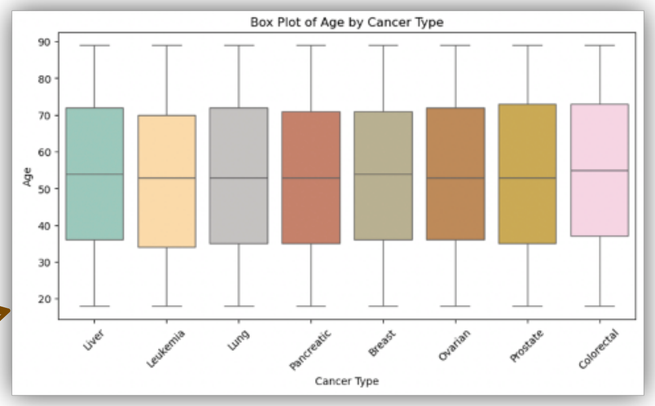

Figure 1 — Box Plot: Age by Cancer Type
Medians are similar and IQRs overlap across types → age alone does not separate cancer types or explain outcome differences.

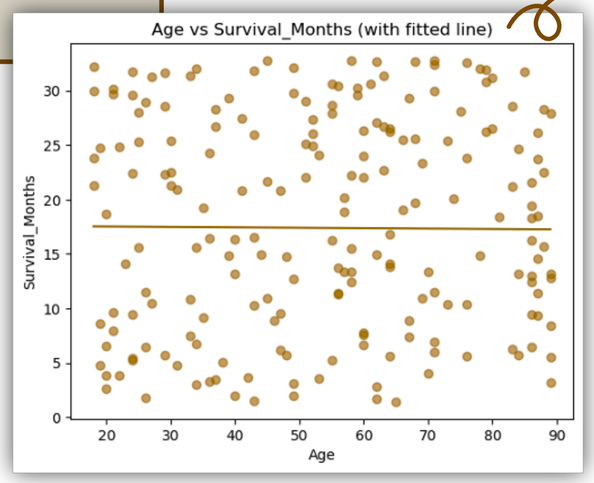

Figure 2 — Age vs Survival_Months (with fitted line)
The OLS line is almost flat ($\beta_1 \approx 0$) → age has a negligible effect on survival months; multivariable models are needed.
Implication. Basic demographics by themselves are weak predictors; add clinical features (stage, type, treatment, comorbidities).