📊 Prostate Cancer Survival Analysis

📌 Project Overview

This project predicts 7-year survival outcomes for prostate cancer patients using logistic regression models.
The dataset contains patient demographics, tumor characteristics, PSA levels, Gleason scores, symptoms, therapies, and staging variables.

The workflow covers:

- Data preparation (cleaning, encoding, feature engineering, selection)
- Multiple logistic regression strategies (forward stepwise, full model, AIC stepwise)
- Interpretation via odds ratios
- Performance evaluation using confusion matrices and classification reports

---

📌 Key Takeaways

- High-risk predictors: N1 lymph nodes, higher Gleason scores, metastasis (M1b worst), smoking.
- Protective therapies: Radiotherapy, brachytherapy, and multi-therapy improve survival.
- Demographics: Patients ≤60 and 80–90 years had worse outcomes.
- Consistency across models: N1, Gleason, metastasis, radiotherapy, and smoking appear in all models.

📁 Dataset

Final split (post-cleaning):

- Train: (5929, 57)
- Test: (2541, 57)

Target variable: survival_7_years

- 1 = survived beyond 7 years
- 0 = did not survive

Predictors:

- Demographics -Age_Group, Race, BMI_Grouped
- Clinical Markers - PSA_diagnosis, tumor_1_year, stage
- Pathology -gleason_score_grouped, t_score, n_score, m_score,
- Therapies- rd_thrpy, brch_thrpy, multi_thrpy, cry_thrpy, chm_thrpy, h_thrpy
- Symptoms - U05, P01, P02, P03, S10, O09, O10, O08, O01

⚙️ Methodology

1. Data Preparation

- Cleaning
    - Records with missing values were dropped after careful analysis.
    - To minimize patient loss, uninformative variables were first removed using correlation filtering and chi-square tests.
    - Since this is clinical data, imputing values could distort medically significant patterns.
    - Final dataset reduced from ~15k to ~8.5k patients.
- Type Conversion
    - Converted numeric variables into categorical groups (e.g., Age bins, Gleason groups).
    - Encoded categorical variables into binary indicators (e.g., therapy received = 1/0).
- Feature Engineering
    - Age Categories: <60 = 1, 60–70 = 2, 70–80 = 3, 80–90 = 4, >90 = 5
    - BMI Calculation:

BMI=0.45455×weight(0.0254×height)2BMI = 0.45455 \times \frac{\text{weight}}{(0.0254 \times \text{height})^2}BMI=0.45455×(0.0254×height)2weight

Classified as Normal (<25), Overweight (25–30), Obese (>30).

  - High-Risk Indicator: Stage III/IV + Gleason ≥7 + PSA ≥10.
  - Symptom Encoding: Converted each code into binary flags (U05, P01, P02, P03, S10, O09).

- Feature Selection
  - Correlation filtering dropped collinear tumor/PSA features.
  - Chi-square tests removed weak predictors (e.g., previous_cancer, h_thrpy).
  - Stepwise AIC selection reduced predictors from 57 to 13 final variables:
  - n_score, gleason_score_grouped, tumor_1_year, m_score, rd_thrpy,
  - Age_Group, U05, brch_thrpy, S10, multi_thrpy, O09, smoker, race

---

2. Exploratory Analysis

- Univariate Analysis: Distributions of age, PSA, Gleason, tumor stages.
- Bivariate Analysis: Compared therapies, race, smoking vs. survival outcome.
- Correlation Analysis: Checked collinearity among continuous features.
- Chi-square Tests: Dropped non-significant categorical predictors.

---

3. Model Development

- Forward Stepwise Logistic Regression – incremental variable selection.
- Full Logistic Regression (57 predictors) – broad baseline model.
- Final Stepwise AIC Model (13 predictors) – optimized balance of fit and simplicity.

---

4. Model Evaluation

- Confusion Matrices and Classification Reports used for performance checks.
- Accuracy: ~64% (train and test) ➔ stable and balanced model.

---

5. Interpretation

- Regression coefficients converted to odds ratios.
- Identified key risk factors (e.g., Gleason, N1, metastasis, smoking) and protective factors (e.g., radiotherapy, brachytherapy).
- Stable predictors across models confirm clinical reliability.

📊 Model Insights

🔹 Forward Stepwise Logistic Regression

- Prostate removal surgery ➔ 21% higher survival odds.
- Multiple therapies ➔ 23% higher survival odds.
- Age 60–70 ➔ 34% lower survival odds.
- Age >90 ➔ 25% lower survival odds.
- Non-obese patients ➔ 12% higher survival odds.
- Not high-risk patients ➔ 33.5% higher survival odds.

🔹 Full Logistic Regression (57 predictors)

- Symptom counts (1–6) ➔ 50–110% higher survival odds.
- Advanced T-stages (T3–T4) ➔ 40–60% worse outcomes.
- N1 lymph nodes ➔ 59% lower survival odds.
- Gleason 7–10 ➔ 28–50% lower survival odds.
- Metastasis (M1a, M1b, M1c) ➔ strongly adverse, worst for M1b (56% lower odds).
- Radiotherapy ➔ 28% better survival odds.

- Brachytherapy ➔ 14% better survival odds.
- Age ≤60 ➔ 39% lower survival odds; Age 80–90 ➔ 22% lower.
- Smokers ➔ 32% higher odds of death.
- Race 1.0 ➔ 25% lower survival odds.

◆ Final Stepwise AIC Model (13 predictors)

- N1 lymph nodes ➔ 63% lower survival odds.
- Gleason scores: Group 2 ↓16%, Group 3 ↓29%, Group 4 ↓49%.
- Metastasis: M1a ↓55%, M1b ↓76%, M1c ↓58%.
- Tumor size at 1 year ➔ each unit ↑ ➔ 0.7% lower odds.
- Radiotherapy ➔ 28% better survival odds.
- Brachytherapy ➔ 17% better survival odds.
- Multi-therapy ➔ 16% better survival odds.
- U05 symptom ➔ 30% lower survival odds.
- S10 symptom ➔ 31% lower survival odds.
- O09 symptom ➔ 60% lower survival odds.
- Age ≤60 ➔ 42% lower survival odds.
- Age 80–90 ➔ 22% lower survival odds.
- Smokers ➔ 32% higher odds of death.
- Race 1.0 ➔ 25% lower survival odds.

📈 Model Performance

Dataset	Accuracy	Precision (0/1)	Recall (0/1)	F1-score (0/1)
Train	64.2%	0.66 / 0.63	0.64 / 0.65	0.65 / 0.64

Dataset	Accuracy	Precision (0/1)	Recall (0/1)	F1-score (0/1)
Test	63.9%	0.66 / 0.62	0.63 / 0.65	0.64 / 0.64

- Balanced results across both classes.
- Consistent train–test ➔ no major overfitting.