Previsione di outcome clinici tramite modelli di regressione

Ferrarini Marco

Esame UFS08 - Machine Learning Supervised

22/05/2025

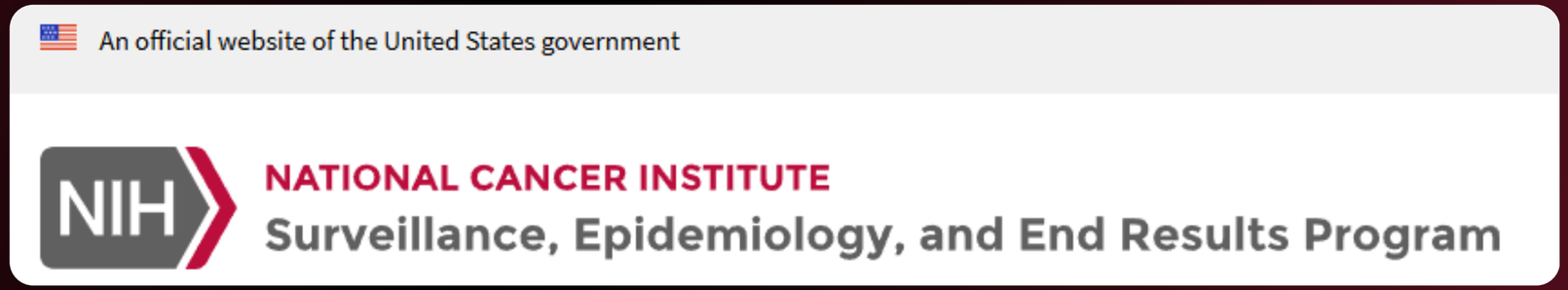
Processo

1. RACCOLTA DATI
2. PULIZIA
3. REGRESSIONE LINEARE
4. RANDOM FOREST
5. XGBOOST

Obiettivo:

predire con un regressore il numero di mesi di vita rimanendti al paziente dal momento della diagnosi

Dataset



The Surveillance, Epidemiology, and End Results (SEER) Program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. SEER is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS).

Variabili selezionate

Shape in origine: 131974 x 53

- **# Biomarcatori**
  - 'E_R_binary', 'pr_binary', 'her2_binary',
- **# Trattamento**
  - 'days_from_diagnosis_to_treatment',
  - 'rx_summ_surg_prim_site',
  - 'rx_summ_scope_reg_ln_sur',
  - 'rx_summ_surg_oth_reg_dis',
  - 'rx_summ_surg_rad_seq', 'reason_no_surgery',
  - 'radiation', 'chemo_yes_no', 'rx_summ_systemic_sur_seq',
- **# Fonte**
  - 'report_source'
- **# Dati demografici e socioeconomici**
  - 'age',
  - 'sex',
  - 'marital_status',
  - 'Race recode (White, Black, Other)',
  - 'Race recode (with detailed Asian and Native Hawaiian other PI)',
  - 'Origin recode NHIA (Hispanic, Non-Hisp)',
  - 'median_household_income_adj_2023',
  - 'rural_urban_continuum',
- **# Caratteristiche del tumore**
  - 'primary_site', 'Schema ID (2018+)', 'ICD-O-3 Hist/behav',
  - 'clinical_grade', 'diagnostic_confirmation',
  - 'tumor_size_summary',
- **# Stadio**
  - 'eod_t', 'eod_n', 'eod_m', 'eod_stage_group',
  - 'eod_primary_tumor', 'eod_regional_nodes', 'eod_mets',
- **#Storia clinica**
  - 'n_sentinel_lymph_nodes',
  - 'n_benign_borderline_tumors',
  - 'n_in_situ_malignant_tumors',
  - 'survival_months',

Metrica di valutazione: Mae

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{actual_i - predicted_i}{actual_i}\right|$$

$$n = number\ of\ considered\ points$$

$$actual_i = actual\ value$$

$$predicted_i = the\ predicted\ value$$

Problemi:

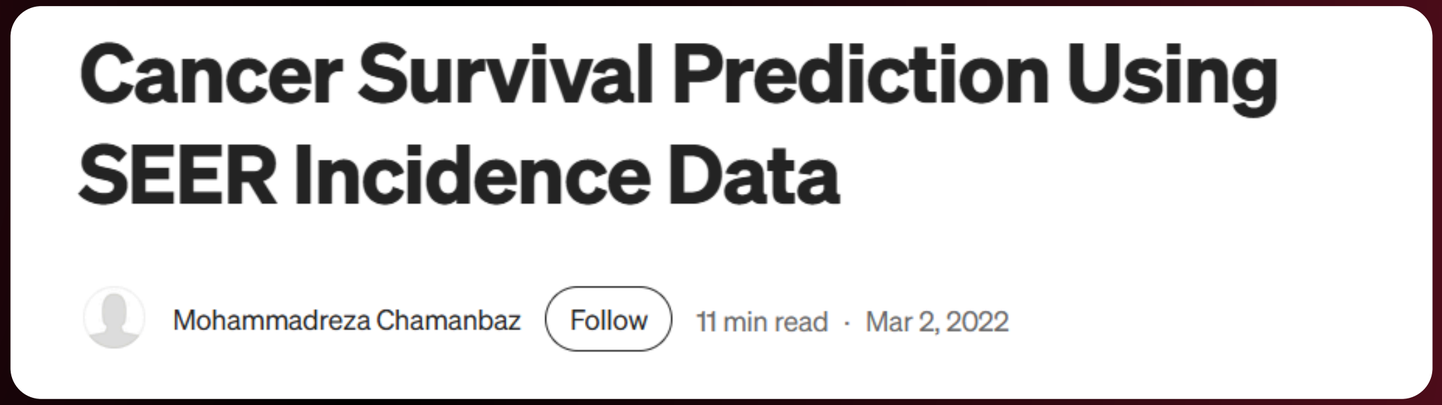
- pesantezza del dataset
- valori mancanti
- codifiche specifiche e complesse

Strumenti:

- ColumnTransformer, Pipeline, optuna, GridSearchCV, Kfold
- Regressori (LinearRegression, XGBRegressor, RandomForestRegressor)

Baseline:
*The baseline model achieves a mean absolute error of **11.9**. We then tune the network's hyperparameters [...]*

*Among all the generated networks, the best achieved mean absolute error is **11.5**, [...] the embedding layer is doing a great job in improving the performance of the regression algorithm.*



MAE sul test set

Regressione Lineare con Target Encoding: **6.35276298286698**

Random Forest senza model selection: **6.258560916284831**

XGBoost con model selection: **5.764026641845703**

Doppio algoritmo con correzione degli errori: **5.749392509460449**

Spostiamoci su VScode

Pulizia

Modelli

Grafici

Analisi dei residui

Fonti

- https://seer.cancer.gov/
- https://medium.com/@m.chamanbaz/cancer-survival-prediction-using-seer-incidence-data-e04503d2d92d
- https://proceedings.mlr.press/v85/hegselmann18a/hegselmann18a.pdf