Supplement to Example Manuscript Template for a Data Analysis Project

# 1. Overview

This supplementary material provides additional details, analyses, and outputs supporting the findings presented in the main manuscript. It includes extended methodological explanations, reproducibility instructions, and results from exploratory data analysis. Key supplementary tables and figures include variable summaries, survival modeling outputs, and machine learning performance metrics. Detailed results from Cox regression, Random Forest, and LASSO Cox models are presented along with confusion matrices, AUC values, and variable importance rankings. The structure and content of the project directory are outlined to facilitate full reproducibility. Together, these materials offer a comprehensive view of the analytical workflow and support the robustness of the study’s conclusions.

# 2. Code and file information

This project is organized in a modular and reproducible structure, separating raw data, processing scripts, analysis outputs, and final products. Below is a guide to the directory structure and instructions to reproduce all results.

# 3. Directory Structure Overview

### 3.0.1 📁 Directory Structure Overview

ZUBER-MADA-project/  
├── data/  
│ ├── raw-data/ # Original input datasets (raw files)  
│ └── processed-data/ # Cleaned and transformed datasets  
│  
├── code/  
│ ├── processing-code/ # Data cleaning scripts  
│ ├── eda-code/ # Exploratory data analysis scripts  
│ └── analysis-code/ # Statistical modeling and prediction  
│  
├── results/  
│ ├── edatables/ # Exploratory summary tables  
│ ├── edafigures/ # Exploratory plots  
│ ├── tables/ # Model results tables  
│ ├── figures/ # Model-related figures  
│ └── output/ # Rendered final outputs (PDF/HTML)  
│  
├── products/  
│ ├── manuscript/ # Manuscript and supplementary files  
│ ├── poster/ # Poster material (if applicable)  
│ └── slides/ # Presentation slides (if applicable)  
│  
├── assets/ # Bibliography and citation style files  
└── README.md # Project overview and general usage instructions

# 4. Additional result

## 4.1 Exploratory Analysis:Race Distribution

As part of the descriptive exploration of the dataset, we examined the distribution of patients by race. The majority of patients in the cohort identified as White, followed by Black, Asian or Pacific Islander, American Indian/Alaska Native, and a small proportion with unknown race information. [Table 1](#tbl-race-summary). Frequency distribution of patients across racial categories.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Distribution of patients by race in the analytic cohort.   | race | Count | | --- | --- | | American Indian/Alaska Native | 38 | | Asian or Pacific Islander | 506 | | Black | 1028 | | Unknown | 12 | | White | 4124 | |

## 4.2 Exploratory Summary: Year of Diagnosis

We also summarized the distribution of patients by year of cancer diagnosis to understand the temporal spread of the study cohort. This information helps contextualize potential changes in treatment patterns, diagnostic advancements, or survival outcomes over time. [Table 2](#tbl-year-summary). Frequency distribution of patients by year of diagnosis.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Distribution of patients by year of diagnosis.   | diagnosis\_year | Count | | --- | --- | | 2010 | 821 | | 2011 | 889 | | 2012 | 960 | | 2013 | 972 | | 2014 | 962 | | 2015 | 1104 | |

1. Bar plot displaying the distribution of patients by year of diagnosis. The figure shows a relatively even spread across the years, with some fluctuations that may reflect data entry trends or cohort recruitment variability.

|  |
| --- |
| Figure 1: Bar plot showing the distribution of patients by year of diagnosis. |

## 4.3 Exploratory Summary: Cause-Specific Death Classification

We summarized the distribution of outcomes according to the SEER cause-specific death classification. The majority of patients in the dataset (77.5%) were classified as having died due to the primary cancer diagnosis, while 22.5% were either alive or had died from other causes. This outcome classification was used as the primary event definition in survival models throughout the analysis.

1. Distribution of patients by cause-specific death classification.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Cause-specific death classification among patients in the cohort.   | cause\_specific\_death | Count | Percentage | | --- | --- | --- | | Alive or dead of other cause | 1287 | 22.5 | | Dead (attributable to this cancer dx) | 4421 | 77.5 | |

## 4.4 Univariate Analysis

To evaluate whether survival outcomes varied over time, we conducted a bivariate Cox proportional hazards analysis using year of diagnosis as the predictor. The year 2010 was used as the reference category. Hazard ratios (HRs) for each subsequent year (2011 to 2015) were calculated to assess whether patients diagnosed in later years had different risks of death compared to those diagnosed in 2010. As shown in [Table 4](#tbl-year-diagnosis), all HRs were close to 1, and none of the comparisons reached statistical significance (all p > 0.05), indicating that there was no significant year-to-year variation in survival over the study period. These findings suggest a relative consistency in survival outcomes for patients diagnosed between 2010 and 2015 within this dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Hazard ratios from Cox regression evaluating year of diagnosis and survival.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | diagnosis\_year2011 vs. 2010 | 0.981 | 0.054 | -0.368 | 0.713 | | diagnosis\_year2012 vs. 2010 | 0.984 | 0.052 | -0.316 | 0.752 | | diagnosis\_year2013 vs. 2010 | 1.000 | 0.053 | -0.001 | 1.000 | | diagnosis\_year2014 vs. 2010 | 0.978 | 0.053 | -0.429 | 0.668 | | diagnosis\_year2015 vs. 2010 | 0.982 | 0.052 | -0.349 | 0.727 | |

## 4.5 Random Forest Model: Predicting Cause-Specific Death

We implemented a random forest classifier to predict whether a patient’s death was attributable to cancer using SEER registry variables. The dataset was preprocessed with downsampling to address class imbalance. We used 5-fold cross-validation to tune hyperparameters and selected the best model based on ROC AUC. The model was trained on 80% of the data and evaluated on a held-out 20% test set.

The top predictors, based on variable importance, included lung and liver metastases, age group, and year of diagnosis.

[Figure 2](#fig-vip-rf) .Variable importance plot from the final tuned random forest model. The top 10 predictors are displayed, with mets\_lung and mets\_liver being the most influential variables for predicting cause-specific death.

|  |
| --- |
| Figure 2: Top 10 most important variables from the random forest model predicting cause-specific death. |

1. Confusion matrix summarizing the random forest model’s classification performance on the test set. The model shows better performance in identifying cancer-attributable deaths than non-cancer deaths, although overall accuracy was limited.

|  |
| --- |
| Table 5: Confusion matrix from the random forest model on the test set.  $table  Truth Prediction No Yes  No 152 385  Yes 92 447  attr(,"class") [1] "conf\_mat" |

1. Area under the receiver operating characteristic curve (ROC AUC) from the random forest model. The model achieved an AUC of approximately 0.39, indicating limited discriminative performance, possibly due to overlapping features and class imbalance.

|  |
| --- |
| Table 6: ROC AUC score of the final random forest model.  # A tibble: 1 × 3  .metric .estimator .estimate  <chr> <chr> <dbl> 1 roc\_auc binary 0.394 |

## 4.6 LASSO Cox Regression for Variable Selection.

To identify the most influential predictors of cancer-specific survival, we applied LASSO (Least Absolute Shrinkage and Selection Operator) penalized Cox regression using the glmnet package. This approach helps in variable selection by shrinking less informative coefficients to zero while retaining strong predictors. We performed 10-fold cross-validation to select the optimal penalty parameter (λ), minimizing the partial likelihood deviance.

1. Selected variables from the LASSO model using the optimal penalty parameter (λ.min). These are the predictors that remained in the final model with non-zero coefficients.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 7: Non-zero coefficients from the LASSO Cox regression model.   | variable | coefficient | | --- | --- | | mets\_lungYes | 0.3323 | | mets\_liverYes | 0.3862 | | mets\_boneYes | 0.2833 | | mets\_brainYes | 0.5570 | | age\_group25-29 years | -0.1979 | | age\_group30-34 years | -0.1054 | | age\_group35-39 years | -0.2170 | | age\_group40-44 years | -0.3530 | | age\_group45-49 years | -0.1630 | | age\_group50-54 years | -0.0678 | | age\_group55-59 years | -0.0226 | | age\_group60-64 years | 0.0159 | | age\_group70-74 years | 0.2417 | | age\_group75-79 years | 0.3073 | | age\_group80-84 years | 0.4241 | | age\_group85+ years | 0.6496 | | raceBlack | 0.3345 | | raceAsian or Pacific Islander | -0.0173 | | raceAmerican Indian/Alaska Native | 0.1343 | | median\_income$40,000 - $44,999 | 0.0456 | | median\_income$50,000 - $54,999 | 0.0571 | | median\_income$65,000 - $69,999 | 0.0468 | | median\_income$70,000 - $74,999 | 0.0038 | | diagnosis\_year2013 | 0.0283 | | diagnosis\_year2014 | -0.0097 | |