Supplementary material - Machine learning survival models for colorectal cancer: Advantages of Random Survival Forests over classification algorithms

Rafael Sant'Ana Herzog

2025-11-11

This document aims to present supplementary results that are referenced throughout the main text of the article "Machine learning survival models for colorectal cancer: Advantages of Random Survival Forests over classification algorithms". Additionaly, all scripts used and the resulting outputs can be found at this GitHub repository.

1 Excluded covariates

As stated in the main text of the article, ten of the 25 covariates used by Buk Cardoso et al. (2023) were excluded from the analysis. Specifically, the following variables were removed:

- `IBGE` and `IBGEATEN`, which represented, respectively, the "IBGE code of the patient's city of residence" and the "IBGE code of the institution";
- `TRATHOSP`, which represented the "combination code of treatments carried out at the hospital", but was a combination of other treatment variables already involved in the analysis;
- `NONE`, which represented "treatment received in hospital = none (0 = no and 1 = yes)", but had very few observations in category 1 (only 0.3%);
- `TMO`, which represented "treatment received in hospital = tmo (0 = no and 1 = yes)", but had only one observation in category 1;
- `IMUNO`, which represented "treatment received in hospital = immunotherapy (0 = no and 1 = yes)", but had very few observations in category 1 (only 0.1%);
- `NENHUMANT`, which represented "treatment received outside hospital and before admission = none (0 = no and 1 = yes)", but had very few observations in category 0 (less than 0.1%);
- `CONSDIAG`, which represented the "difference in days between the dates of consultation and diagnosis", but had negative values;
- and `RRAS` and `DRS`, which represented, respectively, the "Regional Health Care Network code" and the "Regional Health Department code".

2 Characteristics of the population in the total, training, and test datasets

The proportion of events was approximately the same across all total, train (70%) and test (30%) datasets, as shown in **Table 1**.

Event status	Proportion		
	Total dataset	Train	Test
0	0.588	0.589	0.587
1	0.412	0.411	0.413

Table 1: Proportion of events (failures) in the total, train, and test datasets.

The nonparametric estimated median survival time and the survival probabilities at 1, 3, and 5 years, presented in **Table 2**, were also roughly the same across the three datasets.

Time	Estimate		
	Total dataset	Train	Test
Median	6.55 (95% CI: 6.22–7.04)	6.65 (95% CI: 6.22–7.18)	6.41 (95% CI: 5.83–7.4)
1 year	0.815 (95% CI: 0.81–0.819)	0.816 (95% CI: 0.811–0.821)	0.813 (95% CI: 0.805–0.821)
3 years	0.622 (95% CI: 0.616–0.628)	0.622 (95% CI: 0.615–0.629)	0.622 (95% CI: 0.611–0.633)
5 years	0.535 (95% CI: 0.529–0.542)	0.537 (95% CI: 0.529–0.544)	0.533 (95% CI: 0.521–0.545)

Table 2: Estimated median survival and survival probabilities at 1, 3, and 5 years for the total, train, and test datasets.

Frequency distributions and log-rank test results for each covariate and dataset are summarized in **Table 3**. As shown, the percentages of each variable's categories are roughly the same across the three datasets. The only differences in significance in the log-rank test were observed for the radiotherapy and hormone therapy variables.

Variable	Categories	%			P-value (Log-rank)		
		Total	**Train**	**Test**	**Total**	**Train**	**Test**
Education level	Illiterate	4.1%	3.9%	4.3%	<0.0001	<0.0001	<0.0001
	Incomplete Primary Education	35.4%	35.4%	35.3%			
	Complete Primary Education	36.6%	36.7%	36.2%			
	High School	17.4%	17.4%	17.3%			
	Higher Education	6.6%	6.6%	6.9%			
Age group	0 to 49 years	17.2%	17%	17.6%	<0.0001	<0.0001	<0.0001
	50 to 74 years	65.6%	65.9%	65%			
	$\geq$ 75 years	17.2%	17.1%	17.4%			
Sex	Male	51.8%	52%	51.2%	<0.0001	<0.0001	0.0002
	Female	48.2%	48%	48.8%			
Diagnostic care category	Agreement or Private Healthcare	8.2%	8%	8.5%	<0.0001	<0.0001	<0.0001
	Public Healthcare	57.1%	57.2%	57%			
	No information	34.7%	34.8%	34.5%			
Previous diagnosis and treatment	No diagnosis / No treatment	41.9%	41.9%	41.8%	<0.0001	<0.0001	0.0003
	With diagnosis / No treatment	58.1%	58.1%	58.2%			
Clinical staging group	I	12%	12.1%	11.9%	<0.0001	<0.0001	<0.0001
	II	28.4%	28.1%	29%			
	III	32.9%	33%	32.8%			
	IV	26.6%	26.8%	26.3%			
Treatment received in hospital: surgery	No	28.5%	28.7%	28.1%	<0.0001	<0.0001	<0.0001
	Yes	71.5%	71.3%	71.9%			
Treatment received in hospital: radiotherapy	No	72.1%	71.9%	72.6%	0.0415	0.5895	0.0035
	Yes	27.9%	28.1%	27.4%			
Treatment received in hospital: chemotherapy	No	29.5%	29.1%	30.3%	<0.0001	0.0077	0.0007
	Yes	70.5%	70.9%	69.7%			
Treatment received in hospital: hormone therapy	No	99.3%	99.4%	99.2%	0.0010	0.0065	0.0579
	Yes	0.7%	0.6%	0.8%			
Treatment received in hospital: others	No	92.6%	92.8%	92.2%	<0.0001	<0.0001	<0.0001
	Yes	7.4%	7.2%	7.8%			
Time between consultation and treatment	$\leq$ 60 days	75.1%	75%	75.4%	<0.0001	<0.0001	0.0367
	>60 days	24.9%	25%	24.6%			
Time between diagnosis and treatment	$\leq$ 81 days	73.2%	73%	73.6%	<0.0001	<0.0001	0.0010
	>81 days	26.8%	27%	26.4%			
Year of diagnosis	$\leq$ 2006	21.2%	21.2%	21%	<0.0001	<0.0001	<0.0001
	>2006	78.8%	78.8%	79%			
Recurrence	No	87.5%	87.4%	87.8%	<0.0001	<0.0001	<0.0001
	Yes	12.5%	12.6%	12.2%			

Table 3: Percentages and log-rank test results for each variable in total, train, and test datasets.

3 Additional context on the simulation study

To incorporate random censoring in the simulation study performed, censoring times were drawn from a uniform distribution over the interval $(0, \theta_h)$, for $h = 0, 1$. The value of θ_h was determined to achieve a desired right-censoring proportion p_c, by solving $P(T_h > C_h) = p_c$, where T_h and C_h denote failure and censoring times, respectively.

The survival behaviors of clinical stages I and II were specifically used as a basis because they ensured that all resulting θ_h values exceeded 5 — the longest time of interest in this study — across all evaluated scenarios. This guaranteed that censoring could occur throughout the entire follow-up period, allowing for a realistic distribution of censored observations.

The θ_h values obtained for each censoring probability and predictor are presented in **Table 4**.

Table 4: Values of theta for each censoring probability and predictor.

Theta / Censoring probability	0.25	0.40	0.50	0.60	0.70
Related to predictor x = 0	908.873	354.887	192.396	99.668	46.437
Related to predictor x = 1	242.288	97.841	54.306	28.878	13.880

References

Buk Cardoso, Lucas, Vanderlei Cunha Parro, Stela Verzinhasse Peres, Maria Paula Curado, Gisele Aparecida Fernandes, Victor Wünsch Filho, and Tatiana Natasha Toporcov. 2023. "Machine Learning for Predicting Survival of Colorectal Cancer Patients." *Scientific Reports* 13 (1): 8874.