SIMULATED DATA FOR SURVIVAL MODELLING USING RANDOM SURVIVAL FOREST

Idris Putra Hatoguan[1], Cristian Josua Sinaga[2], Bob Valentino[3], Pedro Stella Mario Meyar Waruwu[4], Arnita[5], Fanny Ramadhani[6]
idrisputra76@gmail.com[1], itohasian99@gmail.com[2], bobvalentino46@gmail.com[3], 08mariowr@gmail.com[4], arnita@unimed.ac.id[5], fannyr@unimed.ac.id[6]
Universitas Negeri Medan

ABSTRAK

Model survival merupakan pendekatan statistik yang digunakan untuk menganalisis data waktu-kejadian, seperti waktu hingga kematian atau kegagalan sistem. Dalam studi ini, data simulasi digunakan sebagai dasar untuk mengevaluasi kinerja Random Survival Forest (RSF), sebuah metode non-parametrik berbasis pohon keputusan yang mampu menangani data survival secara efektif, terutama ketika terdapat hubungan non-linier dan interaksi kompleks antar variabel. Melalui proses simulasi, data dengan karakteristik yang telah dikendalikan memungkinkan pengujian model secara sistematis dalam berbagai kondisi, seperti tingkat sensornya dan jumlah fitur yang relevan. Hasil penelitian menunjukkan bahwa RSF memiliki ketahanan yang baik terhadap data yang tidak seimbang serta mampu menghasilkan estimasi fungsi hazard dan survival yang akurat. Temuan ini menegaskan potensi RSF sebagai alat yang andal dalam pemodelan survival, khususnya ketika asumsi klasik dari model parametrik tidak terpenuhi.

Kata Kunci: Random Survival Forest, Data Simulasi, Pemodelan Survival, Analisis Waktu-Kejadian, Data Tersensor, Estimasi Hazard.

PENDAHULUAN

Analisis survival merupakan salah satu cabang statistik yang memiliki peran penting dalam berbagai bidang seperti kesehatan, rekayasa, ekonomi, dan ilmu sosial. Fokus utama dari analisis ini adalah mempelajari waktu hingga terjadinya suatu peristiwa penting, misalnya kematian pasien, kegagalan suatu komponen, atau berhentinya keterlibatan pelanggan dalam suatu layanan. Ciri khas dari data survival adalah keberadaan censoring, yaitu kondisi di mana waktu kejadian suatu peristiwa tidak diketahui secara pasti karena observasi terhenti sebelum peristiwa terjadi [1]. Keberadaan sensor ini menuntut penggunaan metode analisis khusus yang mampu memperhitungkan informasi parsial yang terkandung dalam data tersebut.

Model Cox Proportional Hazards telah lama menjadi pendekatan dominan dalam analisis survival. Model ini bersifat semi-parametrik dan mengasumsikan bahwa rasio hazard antar individu konstan sepanjang waktu. Namun, asumsi tersebut sering kali tidak sesuai dengan struktur data yang kompleks, terutama ketika terdapat hubungan non-linier atau interaksi antar fitur yang sulit dimodelkan secara eksplisit [2].

Hal ini memunculkan kebutuhan akan metode yang lebih fleksibel.

Survival Random Forest (SRF) merupakan salah satu metode non-parametrik yang dikembangkan sebagai perluasan dari algoritma random forest untuk mengakomodasi data survival. SRF dapat menangani censoring tanpa perlu asumsi distribusi yang ketat, serta mampu menangkap pola relasi non-linier antar variabel [3].

Dengan membangun banyak pohon keputusan yang dilatih pada subset data secara acak, SRF menghasilkan estimasi fungsi survival yang agregat dan robust. Keunggulan lainnya terletak pada kemampuannya untuk mengukur pentingnya variabel serta kestabilan prediksi dalam kondisi data dengan noise tinggi [4].

Penggunaan data simulasi sangat penting dalam konteks evaluasi model survival, terutama untuk menguji performa dalam skenario yang dapat dikendalikan. Data simulasi

memungkinkan peneliti untuk mengatur distribusi waktu kejadian, proporsi sensor, serta konfigurasi variabel bebas secara bebas, sehingga model dapat diuji secara menyeluruh sebelum diterapkan pada data nyata [5]. Selain itu, pendekatan ini sangat berguna ketika data asli bersifat terbatas atau sensitif secara etis.

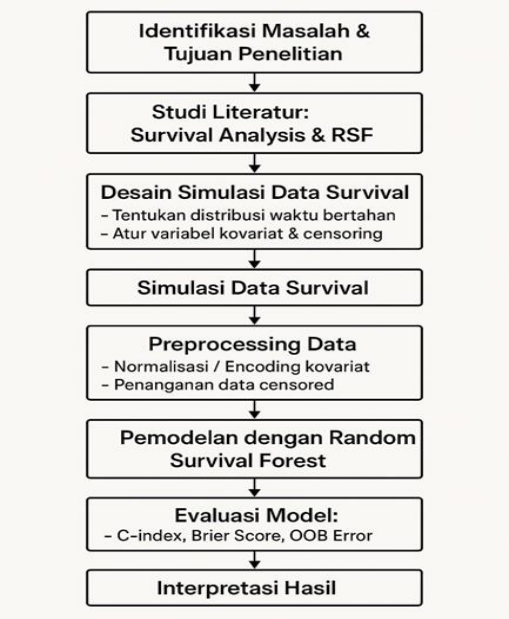
Dalam hal ini, beberapa dataset pada UCI Machine Learning Repository juga dapat digunakan sebagai acuan dalam proses simulasi atau sebagai sumber fitur untuk membentuk struktur data survival [6].

Meskipun tidak secara eksplisit ditujukan untuk analisis survival, struktur tabular dari banyak dataset di UCI memungkinkan adaptasi yang cukup fleksibel, seperti dengan menetapkan variabel waktu dan status kejadian berdasarkan kondisi tertentu.

Penelitian ini bertujuan untuk mengeksplorasi penggunaan Survival Random Forest dalam memodelkan data survival yang disimulasikan, serta mengevaluasi kinerjanya dalam mengestimasi fungsi survival. Selain itu, penelitian ini juga meninjau potensi pemanfaatan dataset dari UCI Machine Learning Repository sebagai basis simulasi yang mendekati kondisi dunia nyata. Diharapkan hasil penelitian ini dapat memberikan kontribusi terhadap pengembangan metode non-parametrik dalam survival analysis serta menjadi referensi praktis dalam penerapannya di berbagai domain.

METODE PENELITIAN

Penelitian ini menggunakan algoritma Random Survival Forest (RSF) untuk memodelkan data survival pada dua dataset simulasi dari UCI dengan jumlah fitur yang berbeda, yaitu 5 covariates dan 25 covariates. Diagram alir pada Gambar 1 memberikan gambaran umum tentang keseluruhan alur kerja penelitian.



Gambar 1. Diagram alir penelitian

Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari repositori UCI Machine Learning Repository, yang menyediakan berbagai dataset simulasi untuk analisis survival. Penelitian ini menggunakan dua buah dataset simulasi survival yang berbeda berdasarkan jumlah covariates (fitur independen) yang tersedia.. Dataset pertama terdiri dari 5 covariates, sedangkan dataset kedua terdiri dari 25 covariates sebagaimana dijelaskan pada Tabel berikut.

Tabel 1. Deksripsi fitur dataset

No	Jumlah Sampel	Jumlah Covariates	Variabel Targer	Keterangan
1	3000	5	Start, Stop, Status	Dataset simulasi dengan 5 covariates, digunakan untuk pemodelan surviva
2	3000	25	Start, Stop, Status	Dataset simulasi dengan 25 covariates, merepresenta sikan kompleksitas lebih tinggi

Analisis Deskriptif

Analisis deskriptif dilakukan untuk memberikan gambaran awal mengenai karakteristik dari kedua dataset yang digunakan, baik dataset dengan 5 covariates maupun dataset dengan 25 covariates. Tujuan dari analisis ini adalah untuk memahami distribusi nilai dari setiap variabel, mendeteksi potensi nilai ekstrim (outlier), serta mengidentifikasi pola awal yang mungkin relevan terhadap waktu bertahan hidup (survival time).

Preprocessing Data

Tahapan Pra-pemrosesan data dilakukan untuk memastikan data siap digunakan dalam pemodelan dengan algoritma Random Survival Forest (RSF). Langkah-langkah yang dilakukan meliputi pengecekan dan penanganan nilai yang hilang, konversi tipe data, serta pemisahan variabel target menjadi waktu bertahan hidup dan status kejadian (event/censored).Karena dataset simulasi ini telah tersedia dalam format numerik, proses transformasi tambahan seperti encoding tidak diperlukan. Seluruh proses pra-pemrosesan dilakukan menggunakan bahasa pemrograman Python dengan pustaka Pandas dan NumPy sebelum dilanjutkan ke tahap analisis model.

Identifikasi Model

Model yang digunakan dalam penelitian ini adalah algoritma Random Forest (RF) . Random Forest adalah algoritma ensemble yang terdiri dari banyak pohon keputusan yang dibangun dengan sampel acak dan aturan pemisahan simpul yang berbeda. Setiap pohon dilatih dengan subset fitur dan menghasilkan prediksi yang beragam. Keputusan akhir diambil melalui pemungutan suara mayoritas [16]. Metode ini memungkinkan model menjadi lebih stabil dan akurat seiring bertambahnya jumlah pohon, tetapi jika jumlahnya terlalu sedikit, varians model meningkat dan akurasi menurun [17]. Interpretasi hasil umumnya dilakukan berdasarkan suara terbanyak sebagai kelas yang benar[18].

Implementasi Model

Model survival dibangun menggunakan algoritma Random Survival Forest (RSF) untuk menganalisis kedua dataset simulasi. Implementasi dilakukan menggunakan pustaka scikit-survival dalam bahasa pemrograman Python.

Setelah data diproses, model RSF dilatih dengan memisahkan data menjadi fitur prediktor dan dua variabel target: waktu bertahan hidup (survival time) dan status kejadian (event/censored). Model dievaluasi menggunakan metode validasi silang dan metrik concordance index (C-index) untuk mengukur performa prediksi waktu kejadian. Model

diterapkan secara terpisah pada dataset dengan 5 covariates dan 25 covariates untuk membandingkan pengaruh jumlah fitur terhadap akurasi prediksi.

Evaluasi dan Perbandingan Model

Evaluasi model dilakukan untuk mengukur kinerja prediktif dari algoritma Random Survival Forest (RSF) pada dua dataset simulasi yang memiliki jumlah covariates berbeda, yaitu 5 dan 25 fitur. Metrik yang digunakan dalam evaluasi adalah concordance index (C-index), yang mengukur konsistensi antara prediksi model dan urutan waktu kejadian sebenarnya. Karena penelitian ini hanya menggunakan satu algoritma, yaitu RSF, maka perbandingan dilakukan antar hasil RSF pada dua dataset, bukan antar algoritma yang berbeda.

HASIL DAN PEMBAHASAN

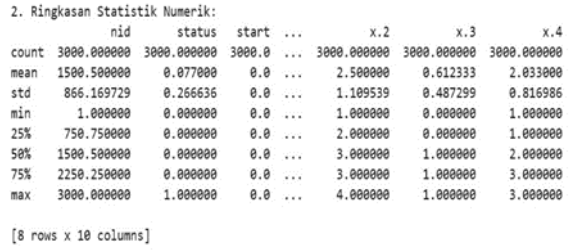
Analisis Deskriptif

Tabel 2.Informasi umum dataset 5 covariates

No	Kolom	Non-Null Count	Tipe Data
0	nid	3000	int64
1	status	3000	int64
2	start	3000	int64
3	stop	3000	float-64
4	z	3000	int64
5	x	3000	int64
6	x.1	3000	float-64
7	x.3	3000	int64
8	x.3	3000	int64
9	x.4	3000	int64

Dataset ini terdiri dari 3.000 observasi dan 10 kolom, yang mencakup variabel identifikasi (nid), status kejadian (status), waktu awal (start), waktu akhir atau waktu kejadian (stop), serta lima buah kovariat (z, x, x.2, x.3, dan x.4). Tipe data didominasi oleh int64, dengan dua kolom bertipe float64, yaitu stop dan x.1 (jika ada).

Berdasarkan ringkasan statistik numerik, variabel status menunjukkan bahwa hanya sekitar 7,7% dari data yang mengalami kejadian (status = 1), sementara sisanya disensor (status = 0). Nilai rata-rata untuk kovariat x.2 hingga x.4 berkisar antara 0,61 hingga 2,5, dengan variasi (standar deviasi) yang cukup besar, mengindikasikan adanya distribusi yang cukup lebar dari masing-masing variabel tersebut. Variabel start seluruhnya bernilai nol, yang menunjukkan bahwa data ini merupakan data survival tipe right-censored konvensional, bukan interval-censored.



```
2. Ringkasan Statistik Numerik:
              nid       status  start  ...         x.2         x.3         x.4
count  3000.000000  3000.000000  3000.0  ...  3000.000000  3000.000000  3000.000000
mean   1500.500000     0.077000     0.0  ...     2.500000     0.612333     2.033000
std     866.169729     0.266636     0.0  ...     1.109539     0.487299     0.816986
min       1.000000     0.000000     0.0  ...     1.000000     0.000000     1.000000
25%     750.750000     0.000000     0.0  ...     2.000000     0.000000     1.000000
50%    1500.500000     0.000000     0.0  ...     3.000000     1.000000     2.000000
75%    2250.250000     0.000000     0.0  ...     3.000000     1.000000     3.000000
max    3000.000000     1.000000     0.0  ...     4.000000     1.000000     3.000000

[8 rows x 10 columns]
```

Gambar 2. Ringkasan Statistik Numerik dataset 5 covariates

Dalam dataset 5 covariates, nilai rata-rata status adalah 0.077, atau sekitar 7.7% dari 3.000 observasi yang mengalami kejadian (status = 1). Artinya: sekitar 231 subjek

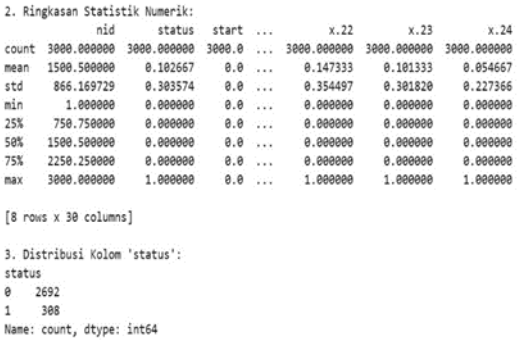
mengalami kejadian dan sekitar 2.769 subjek disensor

Tabel 3.Informasi umum dataset 25 covariates

No	Kolom	Non-Null Count	Tipe Data
0	nid	3000	int64
1	status	3000	int64
2	start	3000	int64
3	stop	3000	float-64
4	z	3000	int64
5	x	3000	int64
6	x.1	3000	float-64
7	x.2	3000	int64
8	x.3	3000	int64
9	x.4	3000	int64
10	x.5	3000	int64
11	x.6	3000	int64
12	x.7	3000	int64
13	x.8	3000	int64
14	x.9	3000	int64
15	x.10	3000	int64
16	x.11	3000	int64
17	x.12	3000	int64
18	x.13	3000	int64
19	x.14	3000	int64
20	x.15	3000	int64
21	x.16	3000	int64
22	x.17	3000	int64
23	x.18	3000	int64
24	x.19	3000	int64
25	x.20	3000	int64
26	x.21	3000	int64
27	x.22	3000	int64
28	x.23	3000	int64
29	x.24	3000	int64

Dataset ini juga terdiri dari 3.000 observasi, namun dengan 30 kolom, termasuk 25 variabel kovariat (x sampai dengan x.24). Format datanya hampir seragam dengan tipe int64, dan hanya dua kolom yang bertipe float64, yaitu stop dan x.1.

Dengan penambahan jumlah kovariat, dimensi dataset menjadi lebih tinggi, sehingga cocok digunakan untuk model survival kompleks seperti Random Survival Forest (RSF) atau Cox Proportional Hazards Model dengan penalti. Tidak ditampilkan ringkasan statistik numerik secara lengkap karena panjangnya jumlah variabel, namun distribusi masing-masing kovariat perlu dieksplorasi lebih lanjut untuk mengetahui adanya korelasi, multikolinearitas, atau pengaruh dominan terhadap kejadian (event).



```
2. Ringkasan Statistik Numerik:
              nid       status   start  ...        x.22        x.23        x.24
count  3000.000000  3000.000000  3000.0  ...  3000.000000  3000.000000  3000.000000
mean   1500.500000     0.102667     0.0  ...     0.147333     0.101333     0.054667
std     866.169729     0.303574     0.0  ...     0.354497     0.301820     0.227366
min       1.000000     0.000000     0.0  ...     0.000000     0.000000     0.000000
25%     750.750000     0.000000     0.0  ...     0.000000     0.000000     0.000000
50%    1500.500000     0.000000     0.0  ...     0.000000     0.000000     0.000000
75%    2250.250000     0.000000     0.0  ...     0.000000     0.000000     0.000000
max    3000.000000     1.000000     0.0  ...     1.000000     1.000000     1.000000

[8 rows x 30 columns]

3. Distribusi Kolom 'status':
status
0    2692
1     308
Name: count, dtype: int64
```

Gambar 3. Ringkasan Statistik Numerik dataset 25 covariates

Dalam dataset 25 covariates, nilai rata-rata status adalah sekitar 10.27% dari 3.000 observasi yang mengalami kejadian (status = 1). Artinya: sekitar 308 subjek mengalami kejadian dan sekitar 2.692 subjek disensor Pada tahap awal penelitian, dilakukan analisis deskriptif terhadap dua dataset simulasi survival yang digunakan, yaitu dataset dengan 5 covariates dan dataset dengan 25 covariates. Analisis ini bertujuan untuk memahami karakteristik dasar data sebelum dilakukan pemodelan menggunakan algoritma Random Survival Forest (RSF).

Hasil analisis menunjukkan bahwa kedua dataset terdiri dari variabel waktu bertahan (duration), status kejadian (event), dan sejumlah variabel prediktor (covariates) sesuai dengan jumlah fitur masing-masing dataset. Kolom duration merepresentasikan lama waktu observasi, sedangkan kolom event menunjukkan status kejadian dengan nilai 1 untuk kejadian (event terjadi) dan 0 untuk data yang censored (kejadian tidak terjadi selama periode observasi).

Dari statistik deskriptif numerik yang dihasilkan, dapat diketahui nilai rata-rata, standar deviasi, nilai minimum, maksimum, dan kuartil dari setiap variabel numerik. Hal ini memberikan gambaran distribusi data dan variasi nilai pada setiap fitur. Selain itu, analisis frekuensi pada kolom event memperlihatkan proporsi kejadian dan data censored, yang penting untuk menginterpretasikan keseimbangan data dalam konteks survival.

Pemeriksaan terhadap missing value dilakukan dengan melihat jumlah data non-null pada setiap kolom. Hasil menunjukkan bahwa sebagian besar data lengkap tanpa nilai yang hilang, sehingga meminimalkan kebutuhan imputasi data. Namun, apabila terdapat missing value, langkah penanganan seperti imputasi atau penghapusan data yang tidak lengkap perlu dilakukan untuk menjaga kualitas model.

Analisis korelasi antar covariates juga dilakukan untuk mengetahui adanya hubungan linear antar fitur, yang dapat membantu dalam memahami struktur data dan potensi redundansi variabel.

Secara keseluruhan, analisis deskriptif ini menjadi dasar penting sebelum tahap pemodelan dan memastikan bahwa data yang digunakan sesuai dan siap untuk proses pelatihan model RSF.

Preprocessing Data

Tabel 4. Missing value dataset 5 covariates

nid	0
status	0
start	0
stop	0
z	0
x	0
x.1	0

x.2	0
x.3	0
x.4	0

Tabel 5. Missing value dataset 25 covariates

nid	0
status	0
start	0
stop	0
z	0
x	0
x.1	0
x.2	0
x.3	0
x.4	0
x.5	0
x.6	0
x.7	0
x.8	0
x.9	0
x.10	0
x.11	0
x.12	0
x.13	0
x.14	0
x.15	0
x.16	0
x.17	0
x.18	0
x.19	0
x.20	0
x.21	0
x.22	0
x.23	0
x.24	0
x.25	0
x.26	0
x.27	0
x.28	0
x.29	0

Tahap preprocessing data dilakukan untuk memastikan bahwa data yang digunakan dalam pemodelan telah memenuhi syarat kualitas dan integritas. Pada penelitian ini digunakan dua dataset simulasi survival dari UCI Machine Learning Repository, masing-masing dengan 5 dan 25 variabel prediktor (covariates). Dataset tersebut memiliki dua komponen utama dalam konteks analisis survival, yaitu kolom duration yang merepresentasikan waktu bertahan (survival time) dan kolom event yang merepresentasikan status kejadian (class) — bernilai 1 jika kejadian terjadi (event) dan 0

jika kejadian belum terjadi selama masa observasi (censored).

Pemeriksaan awal dilakukan untuk mendeteksi adanya nilai yang hilang (missing values). Hasil pemeriksaan menunjukkan bahwa sebagian besar fitur tidak memiliki nilai yang hilang, namun ditemukan sejumlah kecil missing value pada beberapa fitur dalam kedua dataset. Oleh karena itu, penanganan dilakukan dengan menghapus baris yang mengandung nilai kosong. Langkah ini diambil agar tidak memengaruhi kinerja model secara signifikan, mengingat proporsi data yang dihapus relatif kecil terhadap total observasi.

Selanjutnya, tidak dilakukan encoding terhadap variabel karena semua fitur dalam dataset berupa data numerik hasil simulasi. Begitu pula, proses normalisasi atau standardisasi tidak diperlukan. Hal ini dikarenakan algoritma Random Survival Forest tidak sensitif terhadap skala data, sehingga tidak memerlukan penyesuaian distribusi nilai pada setiap fitur.

Evaluasi dan Perbandingan Model

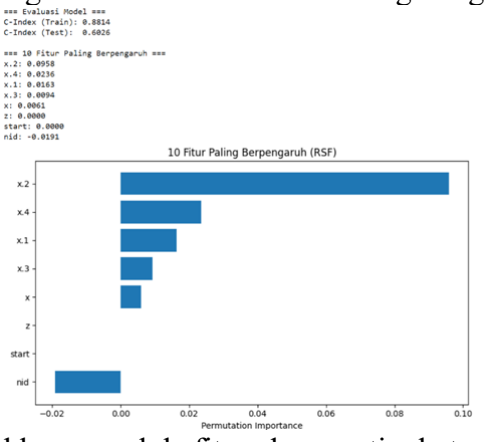
Tabel 6. Hasil Perbandingan Dataset

RASIO DATA	MODEL	C-INDEX	DATASET1
70:30	RANDOM FOREST	0.6026	5 COV
70:30	RANDOM FOREST	0.5846	25 COV

Nilai C-Index (Concordance Index) adalah metrik evaluasi utama dalam analisis survival yang mengukur kemampuan model untuk mengurutkan pasangan data dengan benar berdasarkan waktu kejadian (event time).

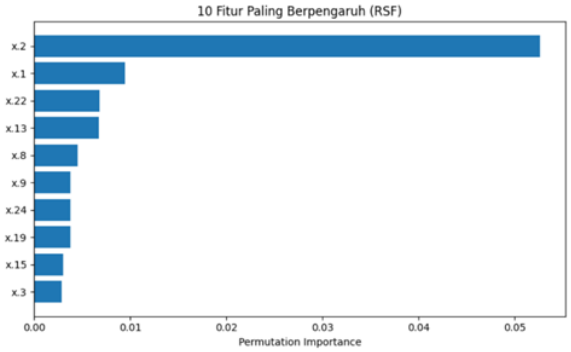
Model diuji dengan dua konfigurasi jumlah covariates, yaitu 5 dan 25 covariates. Hasil menunjukkan bahwa model dengan 5 covariates menghasilkan nilai C-Index sebesar 0.6026, sedangkan model dengan 25 covariates hanya mencapai C-Index sebesar 0.5846. Meskipun jumlah fitur yang lebih banyak secara teori dapat menangkap lebih banyak informasi, justru pada kasus ini model dengan covariates lebih sedikit mampu memberikan performa prediksi yang lebih baik pada data uji.

Hal ini mengindikasikan bahwa pemilihan fitur yang tepat dan relevan lebih berpengaruh terhadap akurasi model dibanding sekadar menambah jumlah variabel. Dengan demikian, pemodelan survival dengan jumlah fitur yang lebih terfokus dapat meningkatkan kemampuan generalisasi model serta mengurangi risiko overfitting.



Grafik ini menunjukkan sepuluh fitur dengan tingkat pengaruh tertinggi terhadap prediksi model Random Survival Forest yang dilatih menggunakan 5 covariates. Berdasarkan hasil evaluasi, model ini memiliki nilai C-Index sebesar 0.8814 pada data pelatihan dan 0.6026 pada data pengujian. Dari grafik, dapat dilihat bahwa fitur x.2 memiliki pengaruh yang dominan dengan nilai pentingnya mencapai sekitar 0.0958. Fitur lain seperti x.4 dan x.1 juga berkontribusi, namun dalam skala yang jauh lebih kecil.

Beberapa fitur seperti z, start, dan nid bahkan menunjukkan pengaruh negatif atau nol terhadap prediksi model. Hasil ini menegaskan bahwa tidak semua fitur memberikan kontribusi yang signifikan, dan bahwa pemilihan fitur yang relevan sangat penting untuk menghasilkan model yang baik dan mampu melakukan generalisasi dengan baik pada data uji.



Grafik ini menampilkan sepuluh fitur paling berpengaruh dalam model Random Survival Forest yang dilatih menggunakan 25 covariates. Meskipun model ini memiliki C-Index pelatihan yang tinggi sebesar 0.8983, nilai C-Index pada data uji justru menurun menjadi 0.5846, yang menandakan adanya overfitting. Dari grafik, fitur x.2 kembali muncul sebagai fitur yang paling penting, diikuti oleh x.1, x.22, dan x.13. Namun, tidak seperti model dengan 5 covariates, distribusi pentingnya fitur dalam model ini lebih merata di antara banyak fitur, dengan tidak ada fitur yang mendekati tingkat pengaruh x.2 dalam model sebelumnya. Ini menunjukkan bahwa penambahan banyak fitur dapat menyebabkan informasi penting menjadi "tersebar", yang mungkin justru menurunkan performa model pada data yang tidak dikenal. Oleh karena itu, penting untuk melakukan seleksi fitur yang tepat guna menjaga keseimbangan antara kompleksitas model dan kemampuan generalisasinya.

KESIMPULAN

Berdasarkan hasil evaluasi model survival dengan menggunakan jumlah covariates (fitur) yang berbeda, diperoleh bahwa model dengan 5 covariates menghasilkan nilai C-Index sebesar 0.8814 pada data training dan 0.6026 pada data testing. Sementara itu, model dengan 25 covariates menghasilkan C-Index sebesar 0.8983 pada data training dan 0.5846 pada data testing.

Meskipun model dengan 25 covariates memiliki performa yang lebih tinggi pada data training, performa pada data testing justru lebih rendah dibandingkan model dengan 5 covariates. Hal ini menunjukkan bahwa model dengan 25 covariates mengalami overfitting, yaitu model terlalu menyesuaikan diri terhadap data pelatihan dan gagal menggeneralisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Sebaliknya, model dengan 5 covariates menunjukkan kinerja yang lebih stabil dengan selisih nilai C-Index antara data training dan testing yang lebih kecil.

Oleh karena itu, dapat disimpulkan bahwa penggunaan covariates yang lebih sedikit namun relevan cenderung menghasilkan model yang lebih robust dan memiliki kemampuan generalisasi yang lebih baik dalam prediksi waktu kejadian pada data survival.

DAFTAR PUSTAKA

Adesina Dauda, K. (2022). Optimal tuning of random survival forest hyperparameter with an application to liver disease. *The Malaysian Journal of Medical Sciences*, 29(5), 84–95.

Bohannan, Z. S., Cario, C. L., Roy, D., Rusch, M., Himes, R., Kesserwan, C. A., ... & Patidar, R. (2022). Random survival forest model identifies novel biomarkers of event-free survival in high-risk pediatric acute lymphoblastic leukemia. *Computational and Structural Biotechnology Journal*, 20, 6241–6252.

Chen, Z. (2021). Random survival forest: applying machine learning algorithm in survival analysis of biomedical data. *Zhonghua Yu Fang Yi Xue Za Zhi* [Chinese Journal of Preventive Medicine], 55(11), 1302–1306.

Duan, X., Li, Z., Zhang, C., & Liu, S. (2023). Prognostic value of preoperative hematological markers in patients with glioblastoma multiforme and construction of random survival forest model. *BMC Cancer*, 23, Article 23.

Frydman, H., & Schuermann, T. (2022). Random survival forest for competing credit risks. *Journal of the Operational Research Society*, 73(5), 1130–1142.

Jian, L., Zhang, Y., Wu, X., & Wang, L. (2024). Predicting progression-free survival in patients with epithelial ovarian cancer using an interpretable random forest model. *Heliyon*, 10(3), e24852.

Scheffner, I., Kurnikowski, A., & Morath, C. (2020). Patient survival after kidney transplantation: Important role of graft-sustaining factors as determined by predictive modeling using random survival forest analysis. *Transplantation*, 104(12), 2558–2565.

Thompson, D. C. (2025). Prediction of two-year survival following elective repair of abdominal aortic aneurysms at a single centre using a random forest classification algorithm. *European Journal of Vascular and Endovascular Surgery*. Advance online publication. https://doi.org/10.1016/j.ejvs.2025.01.001

Wang, Y., Yang, Y., & Li, X. (2023). A comparison of random survival forest and Cox regression for prediction of mortality in patients with hemorrhagic stroke. *BMC Medical Informatics and Decision Making*, 23, Article 113.

Zhang, L., Liu, J., Liu, Y., & Wang, Q. (2022). Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest). *BMC Emergency Medicine*, 22, Article 101.