Menelusuri Model Pergantian Karyawan dan Kekambuhan Pasien Sirosis Hati: Pendekatan Prediktif dengan Metode Regresi Cox dan Random Survival Forest

Revaldy Hazza Daniswara[1], Torang Pandapotan Simangunsong[2], Rafael Wicaksono Hadi[3]

[1,2,3]Departemen Matematika, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstrak

Analisis data survival adalah metode statistik penting untuk memahami durasi hingga terjadinya peristiwa tertentu, seperti pergantian karyawan dan kekambuhan pasien sirosis hati. Penelitian ini menggunakan metode Regresi Cox dan Random Survival Forest (RSF) untuk menganalisis faktor-faktor yang mempengaruhi kedua kasus tersebut. Regresi Cox memodelkan risiko kejadian tanpa asumsi distribusi waktu survival, sementara RSF menangani hubungan kompleks antar variabel dengan akurasi tinggi. Pada kasus pergantian karyawan, uji residual Schoenfeld mengidentifikasi bahwa seluruh variabel kategorik sudah memenuhi asumsi proporsionalitas hazard. Menggunakan teknik Backward Selection, Model terbaik mencakup variabel age, industry, greywage, way dan extraversion. Untuk sirosis hati, RSF menunjukkan skor 94.961%. Variabel signifikan adalah Days, Status, Cohl, dan Age. Evaluasi model menunjukkan C-Index 0.95, Integrated Brier Score 0.070, dan Time Dependent AUC 0.967 setelah cross validation. Hasil penelitian ini diharapkan dapat meningkatkan strategi manajemen sumber daya manusia dan perawatan medis untuk penderita sirosis hati.

Keywords: Analisis data survival, Regresi Cox, Random Survival Forest, Pergantian karyawan, Sirosis hati

Discovering Employee Turnover and Recurrence of Liver Cirrhosis Patients: A Predictive Approach with Cox Regression and Random Survival Forest Methods

Abstract

Survival data analysis is an important statistical method for understanding the duration until specific events occur, such as employee turnover and recurrence of liver cirrhosis patients. This study utilizes Cox Regression and Random Survival Forest (RSF) methods to analyze the factors influencing these two cases. Cox Regression models event risk without assuming a specific survival time distribution, while RSF handles complex variable relationships with high accuracy. In the case of employee turnover, the Schoenfeld residual test identified that all the categorical variables has already meet the proportional hazards assumption. Using the backward selection technique, the best model includes the variables age, industry, greywage, way and extraversion. For liver cirrhosis, RSF showed a score of 94.961%. Significant variables are Days, Status, Cohl, and Age. The model evaluation showed a C-Index of 0.95, an Integrated Brier Score of 0.070, and a Time-Dependent AUC of 0.967 after cross-validation. The results of this study are expected to improve human resource management strategies and medical care for liver cirrhosis patients.

Keywords: *Survival analysis, Cox Regression, Random Survival Forest, Employee turnover, Liver cirrhosis*

1 Pendahuluan

Analisis data survival merupakan salah satu metode statistik yang penting dalam memahami kejadian-kejadian yang memiliki waktu berlangsung hingga terjadi sebuah peristiwa tertentu. Metode ini digunakan secara luas dalam berbagai bidang, termasuk kesehatan, industri, dan manajemen sumber daya manusia. Dalam penelitian ini, kami akan fokus pada dua kasus utama: pergantian karyawan di suatu perusahaan dan kekambuhan pasien sirosis hati.

Pergantian karyawan merupakan fenomena yang sering dijumpai dalam manajemen sumber daya manusia. Tingginya tingkat pergantian karyawan dapat menunjukkan adanya masalah internal dalam organisasi, seperti ketidakpuasan kerja, lingkungan kerja yang buruk, atau kebijakan perusahaan yang kurang mendukung. Memahami faktor-faktor yang mempengaruhi durasi kerja karyawan di perusahaan sangat penting untuk mengembangkan strategi yang dapat mengurangi turnover dan meningkatkan retensi karyawan.

Di sisi lain, sirosis hati adalah kondisi medis serius yang ditandai dengan kerusakan jangka panjang pada hati. Penyakit ini sering kali menyebabkan kematian, dan analisis survival dapat membantu dalam memahami faktor-faktor yang mempengaruhi kekambuhan penyakit sirosis pada penderita. Informasi ini sangat berguna bagi dokter dan peneliti untuk mengembangkan intervensi yang lebih efektif dan meningkatkan kualitas hidup pasien.

Dalam penelitian ini, kami akan menggunakan dua pendekatan utama dalam analisis survival: Regresi Cox dan Random Survival Forest. Regresi Cox, atau model proporsional hazard, tidak memerlukan asumsi distribusi spesifik untuk waktu survival dan dapat digunakan untuk memodelkan pengaruh beberapa kovariat terhadap risiko kejadian.

Sementara itu, Random Survival Forest adalah sebuah me-

tode dengan pendekatan machine-learning yang tidak memerlukan asumsi distribusi tertentu untuk data survival dan dapat menangani hubungan kompleks antar variabel. Metode ini menggunakan ansambel dari pohon keputusan untuk memodelkan risiko kejadian, sehingga memberikan fleksibilitas dan akurasi yang lebih tinggi dalam berbagai kondisi data.

Dengan menggunakan kedua pendekatan ini, kami berharap dapat memberikan pemahaman yang komprehensif tentang faktor-faktor yang mempengaruhi pergantian karyawan dan kematian penderita sirosis hati. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan strategi manajemen sumber daya manusia yang lebih baik serta perawatan medis yang lebih efektif bagi penderita sirosis hati.

2 Metode Analisis

Penelitian ini menggunakan desain observasional retrospektif untuk menganalisis data survival. Tujuan utama adalah untuk mengevaluasi faktor-faktor yang mempengaruhi waktu hingga terjadinya pergantian karyawan dan kambuhnya penyakit sirosis hati dengan menggunakan metode model Regresi Cox dan Random Forest Survival dalam memprediksi waktu kejadian tersebut. Untuk mengetahui faktor-faktor yang berpengaruh, akan digunakan pendekatan yang berbeda.

Pada kasus pergantian karyawan, akan dilakukan backward selection untuk mengetahui faktor mana saja yang paling berpengaruh dilihat dari signifikansinya terhadap model regresi. Sedangkan pada kasus sirosis hati, akan digunakan *feature importance* yang dapat memperlihatkan serangkaian faktor paling berpengaruh berdasarkan rata-rata atau standar deviasi.

Penggunaan regresi cox didasarkan pada tidak dibutuhkan asumsi distribusi yang harus terpenuhi pada data waktu. Oleh karena itu, penggunaan Cox Proportional Hazard (PH) dirasa lebih cocok untuk data ini.

2.1 Regresi Cox Proportional Hazard (PH)

Model Regresi Cox adalah model regresi hazard proporsional dengan fungsi *baseline hazard* nya dimodelkan secara non-parametrik dan fungsi variabel independennya dimodelkan secara parametrik. Sehingga model ini dikenal juga sebagai *Cox proportional hazards model* atau *Cox Semi parametric hazards model.* Metode ini digunakan untuk memodelkan hubungan antara covariates (variabel prediktor) dan tingkat hazard atau risiko suatu peristiwa terjadi pada suatu waktu.

Danardono [2] mengemukakan bahwa regresi Cox dapat dimodelkan pada persamaan sebagai berikut.

$$h_0(t|x) = h_0(t) \cdot \psi(x, \beta) \tag{1}$$

dengan x = $(x_1, ..., x_p)$ adalah vektor kovariat(variabel independen) dan $\beta^{'} = (\beta_1, ..., \beta_p)$ adalah parameter dari model regresi. Dalam regresi ini, hazard untuk tiap-tiap individu sama dengan baseline hazard $h_0(t)$ apabila dipengaruh variabel independen tidak diperhatikan, atau nilai x = $(x_1, ..., x_p)$ semuanya = 0. Hazard dari masing-masing individu termodifikasi secara multiplikatif oleh karakteristik masing-masing individu, yang diekspresikan dengan $\psi(x, \beta)$.

Kemudian akan digunakan juga Estimasi parameter pada model Regresi Cox yang didasarkan pada Partial Likelihood:

$$L(\beta) = \prod_{k \in D} \frac{exp(x_k \cdot \beta)}{\sum_{j \in R_k} exp(x_j \cdot \beta)} \tag{2}$$

dengan x adalah vektor kovariat (variabel penjelas); β adalah parameter regresi yang akan diestimati D adalah himpunan indeks j dari semua waktu kejadian (semua t_j yang mendapatkan kejadian); R_k adalah himpunan risiko (*risk set*).

Untuk melakukan regresi cox, maka kita harus memastikan agar semua variabel yang dipunya memenuhi uji asumsi proporsionalitas hazard. Pada umumnya, akan dilakukan sebuah pengujian bernama uji residual Schoenfeld. Model regresi cox mengharuskan semua individu memiliki faktor hazard yang sama, hanya berbeda di kovariat nya saja. Oleh karena itu, bentuk dari fungsi hazard sama untuk setiap individu, hanya berbeda pada perkalian skalar per individu.

$$h_i(t) = a_i h(t) \tag{3}$$

Dengan hazard ratio yang harus konstan untuk setiap t dapat dituliskan sebagai berikut.

$$\frac{h_i(t)}{h_j(t)} = \frac{a_i h(t)}{a_j h(t)} = \frac{a_i}{a_j} \tag{4}$$

Jika terdapat variabel yang tidak memenuhi uji asumsi proporsionalitas hazard, maka dapat dilakukan proses stratifikasi. Proses ini akan akan menentukan *baseline hazard* yang berbeda untuk masing-masing strata, tetap parameter β tetap sama untuk setiap strata. Menurut Danardono [1], Model regresi cox terstratifikasi dapat dijabarkan pada persamaan berikut.

$$h_j(t|x) = h_{o_j}(t) \cdot exp(x\beta) \tag{5}$$

dengan j = 1,2,...,s adalah banyaknya strata. Estimasi nilai β atau parameter regresi dapat dilakukan melalui log-partial likelihood sebagai berikut.

$$\ell_{(\beta)} = \ell_{1(\beta)} + \ell_{2(\beta)} + \cdots + \ell_{S(\beta)} \tag{6}$$

dengan $\ell_{j(\beta)}$ adalah log-partial likelihood hanya pada subset dalam strata ke-j. Pendekatan model regresi cox jika terdapat ties (kejadian bersama) dapat berpengaruh pada pembentukan himpunan risiko (*risk set*) pada estimasi parameter regresi. Hal ini dapat ditangani dengan menggunakan metode Breslow yang dirumuskan pada persamaan berikut.

$$L(\beta) = \prod_{k \in D} \frac{\exp(S_k \beta)}{\left[\sum_{j \in R_k} \exp(x_j \beta)\right]^{d_k}} \tag{7}$$

Untuk mengetahui faktor mana saja yang tidak memenuhi uji asumsi proporsionalitas, dapat dilihat melalui grafik hazard kumulatif atapun dengan menggunakan uji residual Schoenfeld. Misal $s_{t,j}$ adalah *scaled* residual Schoenfeld dari variabel j pada waktu t, $\hat{\beta}_j$ merupakan estimasi dari maximum-likelihood dari variabel ke-j, dan $\beta_j(t)$ adalah sebuah alternatif koefisien *varying time.* Dietz et al [3]. menunjukkan bahwa.

$$E[s_{t,j}] + \hat{\beta}_j = \beta_j(t) \tag{8}$$

Asumsi proporsional hazard berimplikasi pada $\hat{\beta}_j = \beta_j(t)$, karena $E[s_{t,j}] = 0$

2.2 Random Survival Forest

Random Forest adalah metode machine learning berbasis ensemble yang dapat digunakan untuk memodelkan data survival. Metode ini lebih dikenal sebagai Random Survival Forests (RSF). Metode ini sangat cocok digunakkan pada data tersensor kanan (*right-censoring data*) [7]. Hal ini karena pada dasarnya RSF berangkat dari pendekatan nonparametrik seperti Nelson-Aalen maupun Kaplan-Meier

Dasar pembelajaran dari metode ini adalah adanya randomisasi. Pada metode machine-learning, random forest yang terbentuk terbagi menjadi dua, diantaranya adalah secara acak mengambil bootstrap dari sampel data untuk dijadikan sebagai *tree*. Selain itu, ada juga pada setiap node dari *tree* secara random dipilih subset dari variabel (kovariat).

Pada data tersensor kanan, RSF membutuhkan waktu survival dan juga informasi terkait *censoring*. Hal ini juga selaras dengan kebutuhan untuk melakukan *splitting* data.

2.2.1 Algoritma Random Survival Forest

Menurut Ishwaran et al.[7], secara sistematis RSF bekerja melalui algoritma sebagai berikut.

- Gambar B sampel bootstrap dari data. Setiap sampel bootstrap mengecualikan 37% dari data yang disebut sebagai out-of-bag data (OOB data).
- Buat sebuah survival *tree* untuk setiap sampel bootstrap. Pada setiap node dari pohon, secara acak ambil p kandidat variabel. Node dipecah menggunakan kandidat variabel yang memaksimalkan perbedaan survival antara *daughter nodes*.
- Kembangkan *tree* hingga ukuran penuh dengan batasan bahwa sebuah terminal node harus memiliki tidak kurang dari $d_0 > 0$ kematian unik.
- Hitung Cummulative Hazard Function (CHF) dari setiap *tree*. Rata-ratakan untuk mendapat ensemble CHF.
- Dengan menggunakan OOB data, hitung prediksi error untuk ensemble CHF.

2.2.2 Kumulatif Hazard Ensembel

1. **Pohon Survival Biner**
   Survival *tree* adalah pohon yang dikategorikan biner yang tumbuh dengan membagi rekursif node-node dari pohon. Pohon ini mulai tumbuh dari *root node*, yang merupakan bagian atas pohon yang mencakup semua data. Dengan menggunakan kriteria survival yang telah ditetapkan sebelumnya, *root node* dibagi menjadi dua *daughter node* yaitu kiri dan kanan.

   Splitting yang baik untuk node adalah yang memaksimalkan survival antara *daughter nodes*. Hal ini dapat ditemukan melalui mencari semua kemungkinan variabel x, nilai c dan pemilihan x^* serta c^*. Setelah memaksimalkan survival, *tree* akan menjauhkan segala ketidaksamaan. Pada akhirnya, semakin banyak nodes dan ketidaksamaan yang dipisahkan membuat setiap *tree* menjadi homogen dan dikumpulkan dengan kesamaannya.

2. **Prediksi Terminal Node**
   Pohon survival akan mencapai titik saturasi ketika tidak ada *daughter node* yang terbentuk karena tidak memenuhi kriteria terminal node harus memiliki tidak kurang dari $d_0 > 0$ kematian unik. *node* yang paling ekstem di pohon tersaturasi disebut *terminal nodes*. Estimasi CHF dapat menggunakan pendekatan non-parametrik yaitu Nelson-Aalen yang dapat dirumuskan sebagai berikut.

$$\hat{H}_h(t) = \sum_{t_{l,h} \le t} \frac{d_{l,h}}{Y_{l,h}} \tag{9}$$

   dengan $d_{l,h}$ adalah jumlah kematian individual pada waktu $t_{l,h}$ dan $Y_{l,h}$ adalah individual at risk pada waktu $t_{l,h}$. Oleh karena itu, CHF untuk i adalah estimasi Nelson-Aalen untuk terminal node $\mathbf{x}_i$.

$$H(t|\mathbf{x}_i) = \hat{H}_h(t) \tag{10}$$

3. **Bootstrap dan OOB Ensemble CHF**
   CHF diturunkan dari *tree* tunggal. Untuk menghitung sebuah ensemble CHF, kita perlu merata-ratakan terhadap B pohon survival. Ingat kembali jika setiap *tree* di hutan dikembangkan dengan sampel bootsrap independen. Kita bisa mendefinisikan $I_{i,b}$ sebagai berikut.

$$I_{i,b} = \begin{cases} 1 & \text{jika } i \text{ adalah kasus OOB untuk } b \\ 0 & \text{lainnya} \end{cases}$$

   Selanjutnya, dapat dirumuskan untuk ensemble CHF untuk i sebagai berikut.

$$H_e^{**}(t \mid \mathbf{x}_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t \mid \mathbf{x}_i)}{\sum_{b=1}^{B} I_{i,b}} \tag{11}$$

   atau

$$H_e^*(t \mid \mathbf{x}_i) = \frac{1}{B} \sum_{b=1}^{B} H_b^*(t \mid \mathbf{x}_i) \tag{12}$$

   dengan $H_b^{**}(t \mid \mathbf{x}_i)$ adalah CHF untuk *tree* yang dikembangkan dari bth sampel bootstrap.

2.2.3 Pemilihan Metode Permodelan Survival

Pemilihan model regresi Cox pada dataset pergantian karyawan didasarkan pada kemampuan model ini untuk menangani data tersensor, di mana karyawan yang belum terganti masih memiliki kemungkinan untuk berganti di masa depan yang tidak terdeteksi. Menurut Samar [8], Regresi Cox memungkinkan estimasi risiko relatif (hazard ratio) dan penyesuaian terhadap konfounder, memberikan estimasi yang lebih akurat dibandingkan metode Kaplan-Meier. Selain itu, model ini dapat menganalisis variabel kontinu dan kategorikal tanpa memerlukan kategorisasi, sehingga mampu menghitung peningkatan atau penurunan risiko pergantian yang terkait dengan perubahan dalam variabel prediktor. Penggunaan Regresi Cox juga sangat cocok jika semua variabel kategorial memenuhi asumsi proporsionalitas hazard. Ini bisa dibuktikan dengan uji residual Schonfeld.

Sementara itu, pemilihan model random survival forest (RSF) didasarkan pada kemampuannya untuk menangani

korelasi antar *tree* dengan membangun setiap *tree* pada sampel bootstrap berbeda dari data pelatihan asli, dan pada setiap node hanya mengevaluasi subset acak dari fitur dan ambang batas. Prediksi dibuat dengan mengagregasi prediksi dari masing-masing *tree* dalam ensemble. Sebagai metode berbasis machine learning, RSF mampu mengungkap hubungan-hubungan signifikan dari *root node* hingga *leaf node*, dengan menyingkirkan hubungan yang tidak relevan. Meskipun pendekatannya non-parametrik, RSF tetap memungkinkan penggunaan kovariat regresi seperti dalam pendekatan parametrik. Dalam konteks analisis data liver, kovariat regresi (x) dapat digunakan bersama dengan status dan waktu (y), memungkinkan RSF menangani kompleksitas data dan variabel secara lebih fleksibel dibandingkan metode non-parametrik tradisional.

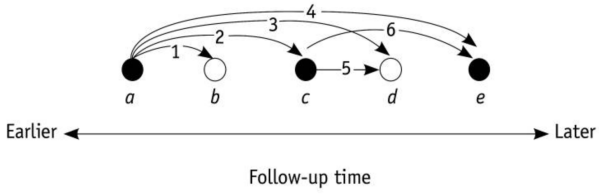
2.3 Metrik Evaluasi

2.3.1 C-Index (Concordance Indec)

C-Index atau concordance index adalah *evaluation metrix* yang digunakan dalam *surival anaylsis* yang melibatakan data *time-to-event*. C-Index atau *concordance index* berikatan dengan area di bawah kurva ROC [5]. C-Indeks memperkirakan probabilitas bahwa dalam pasangan kasus yang dipilih secara acak, kasus yang gagal pertama kali memiliki hasil prediksi terburuk. Dengan kata lain, C-Index mengukur seberapa baik model prediksi dapat mengurutkan kejadian dengan benar berdasarkan hasil yang diprediksi. Interpretasi dari C-Index dapat diartikan sebagai probabilitas kesalahan klasifikasi, yang berarti semakin tinggi nilai C-Index, semakin baik model dalam mengklasifikasikan hasil dengan benar yang merupakan salah satu alasan kami menggunakannya untuk kesalahan prediksi. Fitur menarik lainnya adalah, tidak seperti ukuran lain dari kinerja survival,

C-Index tidak bergantung pada satu waktu yang tetap karena C-Index mengevaluasi kinerja prediksi secara keseluruhan tanpa terbatas pada satu waktu tertentuuntuk evaluasi. C-index juga secara khusus memperhitungkan data yang tersensor. Gambaran nilai C-Index yaitu semakin baik performa model dalam memprediksi urutan kejadian. Nilai maksimum C-Index adalah 1 yang menunjukkan prediksi sempurna, dan nilai minimum adalah 0 yang menunjukkan prediksi acak.

1. **Algoritma C-Index atau Concordance Index**
   Menurut Heagerty dan Zheng [5], C-Index dapat dikalkulasikan melalui langkah-langkah sebagai berikut.
   - *C-Index*, didefinisikan dengan C = Concordance/Permissible.
   - Membentuk semua pasangan kasus yang mungkin terjadi pada data.
   - Hilangkan pasangan-pasangan yang waktu kelangsungan hidupnya lebih pendek yang disensor. Hilangkan pasangan *i* dan *j* jika; $T_i = T$; kecuali setidaknya ada satu yang mati dan biarkan *permissible* menyatakaan jumlah total pasangan yang diizinkan
   - Untuk setiap pasangan yang diizinkan dimana $T_i \neq T$; hitung 1 jika waktu tahan hidup yang lebih pendek memiliki hasil prediksi yang lebih buruk; hitung 0,5 jika hasil prediksi sama. Untuk setiap pasangan yang diizinkan, di mana $T_i = T$; dan keduanya adalah kematian, hitung 1 jika hasil prediksi sama; jika tidak, hitung 0,5. Untuk setiap pasangan yang diperbolehkan dimana $Ti = Tj$, tetapi tidak keduanya adalah kematian, hitung 1 jika kematian memiliki hasil prediksi yang lebih buruk; jika tidak, hitung 0,5. Biarkan Concordance menyatakan jumlah dari semua pasangan yang diizinkan.



Gambar 1. Gambaran Mekanisme C-Index

2.3.2 Integrated Brier Score

Integrated Brier Score adalah sebuah metrik evaluasi yang merupakan perkembangan dari Brier Score. Metrik ini muncul ketika Brier Score tidak mampu untuk mengakomodasi semua titik-titik waktu pada survival. Integrated Brier Score menampilkan penilaian sebuah model survival yang menyeluruh dari waktu t_1 hingga t_{max}. IBS bernilai 0 hingga 1, semakin kecil nilai IBS, semakin baik modelnya. Integrated Brier Score secara sederhana dapat ditulis sebagai berikut.

$$IBS = \int_{t_1}^{t_{\max}} \frac{t}{t_{max}} \cdot BS(t)\ dt \tag{13}$$

Dengan t_1 dan t_{max} adalah titik waktu minimum dan maksimum secara berturut-turut. BS(t) adalah Brier Score pada waktu t yang dapat dihitung pada persamaan di bawah ini.

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N} w_{it}(\hat{\gamma}_{it} - \gamma_{it})^2$$

w_{it} dapat dimaknai sebagai bobot yang bisa diestimasi menggunakan pendekatan non-parametrik yaitu Kaplan-Meier dari distribusi *censoring* G pada data. w_{it} dapat dijawabarkan dalam *piecewise function* berikut.

$$w_{it} = \begin{cases} \delta_i / G(\gamma_i) & \text{jika } \gamma_i \leq t \\ 1/G(\gamma_i) & \text{jika } \gamma_i \geq t \end{cases}$$

Di sisi lain, jika terdapat status *censoring* akan digunakan Integrated Brier Score untuk *censored observation* [4]. Untuk perhitungannya juga berbeda, namun secara sederhana dapat dijabarkan sebagai berikut.

$$IBS^c = \frac{1}{t_{max} - t_1} \cdot \int_{t_1}^{t_{\max}} BS^c(t)\ dt \tag{14}$$

Ekspresi $BS^c(t)$ merupakan Time Dependent Brier Score untuk *censored data* yang dapat dituliskan pada persamaan berikut.

$$BS^c(t) = \frac{1}{N}\sum_{i=1}^{N}\left[I(y_i \leq t, \delta_i = 1)\left(\frac{(0-\hat{S}(t|x^i))^2}{\hat{G}(y_i)}\right) + I(y_i > t)\left(\frac{(1-\hat{S}(t|x^i))^2}{\hat{G}(y_i)}\right)\right] \tag{15}$$

1. **Algoritma Independent Brier Score**
   Menurut Husnaqilati [6], untuk menghitung Independent Brier Score dapat melalui langkah-langkah berikut.

- Hitung brier score BS(t) untuk setiap titik waktu t yang berbeda menggunakan probabilitas survival yang sudah diprediksi dan *true event atau censoring status*.
- Aplikasikan *inverse probability of censoring weights* (IPCW) untuk menangani data tersensor ketika menghitung BS(t).
- Integralkan Brier scores berbobot $\frac{t}{t_{max}}$BS(t) untuk setiap titik waktu menggunakan integral numerik seperti aturan trapezoidal.

2. **Keunggulan Integrated Brier Score**
Sebagai sebuah metrik evaluasi, pemilihan metrik berdasarkan kelebihan adalah pertimbangan yang sangat penting. Untuk itu, kami sajikan kelebihan dari IBS sebagai berikut.

- Menghadirkan penilaian (*assesment*) yang menyeluruh pada keseluruhan interval waktu, bukan hanya sekadar pada titik spesifik waktu saja.
- Dapat mengakomodasi data tersensor dengan bobot IPC (Inverse Probability of Censoring).
- Memberikan bobot lebih pada titik waktu yang lebih lama, sering kali lebih relevan dalam komputasi.

3 Dataset

3.1 Penjelasan Dataset

Pada bagian ini, akan dibahas mengenai dataset secara keseluruhan. Dataset secara keseluruhan merujuk pada kumpulan data lengkap yang digunakan dalam suatu analisis atau penelitian. Penyajian head dari data mengacu pada tampilan beberapa baris pertama dari dataset untuk memberikan gambaran awal tentang struktur dan isi datanya.

Berikutnya, penjelasan variabel dalam data mencakup deskripsi dari setiap kolom atau atribut yang ada dalam dataset. Hal ini meliputi nama variabel, statistik deskriptif, serta jenis data (misalnya numerik, kategorikal, atau teks)

3.1.1 Data Pergantian Karyawan

Selanjutnya, kami akan menampilkan sedikit mengenai data pertama yang akan digunakan. Sekilas terkait data turnover atau pergantian karyawan beserta statistik deskriptif untuk variabel numerik dapat dilihat pada tabel berikut.

stag	event	gender	age	industry	profession	traffic	coach	head_gender	greywage	way
7.030800821	1	m	35	Banks	HR	rabrecNErab	no	f	white	bus
22.9650924	1	m	33	Banks	HR	empjs	no	m	white	bus
15.93429158	1	f	35	PowerGeneration	HR	rabrecNErab	no	m	white	bus
15.93429158	1	f	35	PowerGeneration	HR	rabrecNErab	no	m	white	bus
8.410677618	1	m	32	Retail	Commercial	youjs	yes	f	white	bus

Tabel 1. Data Turnover

	stag	event	gender	age	industry	profession	traffic	coach	head_gender	greywage	way	extraversion	independ	selfcontrol	anxiety	novator
count	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.000	1129.00
mean	36.628	0.506	0.244	31.067	8.615	6.561	4.179	0.839	0.517	0.888	0.500	5.592	5.478	5.597	5.666	5.880
std	34.097	0.500	0.430	6.996	4.233	2.514	2.253	0.608	0.500	0.316	0.677	1.852	1.703	1.980	1.709	1.904
min	0.394	0.000	0.000	18.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.700	1.000
25%	11.729	0.000	0.000	26.000	5.000	6.000	2.000	0.000	0.000	1.000	0.000	4.600	4.100	4.100	4.800	4.400
50%	24.345	1.000	0.000	30.000	10.000	6.000	4.000	1.000	1.000	1.000	0.000	5.400	5.500	5.700	5.600	6.000
75%	51.318	1.000	0.000	36.000	12.000	6.000	7.000	1.000	1.000	1.000	1.000	7.000	6.900	7.200	7.100	7.500
max	179.450	1.000	1.000	58.000	15.000	14.000	7.000	2.000	1.000	1.000	2.000	10.000	10.000	10.000	10.000	10.000

Tabel 2. Statistik Deskriptif Data Turnover

Variabel tentunya memiliki keterangan dan karakteristik yang berbeda-beda. Setelah itu, guna memperjelas dataset akan disajikan keterangan untuk setiap variabel pada data turnover sebagai berikut.

- stag: waktu atau lama bekerja
- event: pergantian karyawan (1 = Ya, 0 = Tidak)
- gender: jenis kelamin (f = perempuan, m = laki-laki)
- age: usia pekerja (tahun)
- industry: bidang pekerjaan (Ritel, Manufaktur, dan lainnya)
- profession: pekerjaan (HR, IT, dan lainnya)
- traffic: asal pekerja (menghubungi langsung perusahaan atau lainnya)
- coach: apakah ada pelatihan selama masa probation?
- head_gender: jenis kelamin supervisor
- greywage: gaji yang tidak dilaporkan kepada otoritas pajak

3.1.2 Data Sirosis Hati

Berikutnya, kami akan menampilkan sedikit mengenai data kedua yang akan digunakan. Sekilas terkait data sirosis hati atau sirosis hati beserta penjelasan setiap variabelnya sebagai berikut.

Status	Days	Drug	Age	Sex	Ascites	Hepatom	Spiders	Edema	Bill	Cohl	Albumin	Copper	Alk Phos	Sgot	Trig	Platelet	Protime	Stage
0	4500	1	20617	1	0	1	1	0	1.1	302	4.14	54	7394.8	113.52	88	221	10.6	3
0	1832	2	20284	1	0	1	0	0	1	322	4.09	52	824	60.45	213	204	9.7	3
0	3577	2	16688	1	0	0	0	0	0.7	281	3.85	40	1181	88.35	130	244	10.6	3
0	3672	2	14772	1	0	0	0	0	0.7	204	3.66	28	685	72.85	58	198	10.8	3
0	4232	1	18102	1	0	1	0	0.5	0.7	235	3.56	39	1881	93	123	209	11	3

Tabel 3. Data Sirosis Hati

- status: kekambuhan (0 = tersensor, 1 = kambuh)
- days: jumlah hari antara registrasi hingga kematian, transplantasi atau masih masuk ke dalam studi.
- drug: perlakuan obat (1 = D-penicilline, 2 = placebo)
- sex : jenis kelamin (0 = laki-laki, 1 = perempuan)
- ascites = apakah ada ascites? (0 = tidak, 1 = ada)
- hepatom = apakah ada hepatomegaly? (0 = tidak, 1 = ada)
- spiders = apakah ada spiders? (0 = tidak, 1 = ada)
- edema = apakah ada edema? (0 = tidak ada dan tidak ada terapi diuretik untuk edema, 1 = ada, dan terdapat terapi diuretik untuk edema)
- bill = serum bilirubin dalam mg/dl
- cohl = serum kolesterol dalam mg/dl
- albumin = jumlah albumin dalam g/dl
- alk phos = jumlah alkaline fosfat dalam u/liter
- sgot = jumlah SGOT dalam u/ml
- trig = jumlah trigliserida dalam mg/dl
- platelet = jumlah platelet dalam mg/dl
- protime = waktu terbentuknya protrombin dalam detik
- stage = tingkat histologis dari penyakit

Kemudian, untuk statistik deskriptif nya dapat disajikan pada tabel sebagai berikut.

	Status	Days	Drug	Age	Sex	Ascites	Hepatom	Spiders	Edema	Bill	Cohl	Albumin	Copper	Alk Phos	Sgot	Trig	Platelet	Protime	Stage
count	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.000	418.00
mean	0.385	1917.782	1.115	18533.352	0.895	0.057	0.383	0.215	0.160	3.221	251.655	3.497	72.419	1479.877	91.477	84.129	250.261	10.680	2.980
std	0.487	1104.673	0.781	3815.845	0.307	0.233	0.487	0.412	0.253	4.408	257.528	0.425	85.223	2040.235	72.439	79.257	105.402	1.261	0.946
min	0.000	41.000	0.000	9598.000	0.000	0.000	0.000	0.000	0.000	0.300	0.000	1.960	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	0.000	1092.750	0.000	15644.500	1.000	0.000	0.000	0.000	0.000	0.800	0.000	3.243	0.000	0.000	0.000	0.000	181.000	10.000	2.000	2.000
50%	0.000	1730.000	1.000	18628.000	1.000	0.000	0.000	0.000	0.000	1.400	252.500	3.530	50.500	1009.000	90.450	85.000	248.000	10.600	3.000
75%	1.000	2613.500	2.000	21272.500	1.000	0.000	1.000	0.000	0.000	3.400	347.750	3.770	100.750	1707.750	135.750	127.750	315.500	11.100	4.000
max	1.000	4795.000	2.000	28650.000	1.000	1.000	1.000	1.000	1.000	28.000	1775.000	4.640	588.000	13862.400	457.250	598.000	721.000	18.000	4.000

Tabel 4. Statistik Deskriptif Data Sirosis Hati

4 Hasil dan Pembahasan

4.1 Data Pergantian Karyawan

Dengan memanfaatkan library lifelines untuk keperluan survival di Python, telah dilakukan sebuah pemodelan regresi dengan menggunakan Regresi Cox. Sebelum itu, akan dilakukan uji asumsi proporsionalitas hazard atau yang biasa disebut uji residual Schoenfeld. Uji hipotesis akan disajikan dalam beberapa pernyataan sebagai berikut.

- **Uji Hipotesis**
  H_0 : asumsi proporsionalitas hazard terpenuhi
  H_a : asumsi proporsionalitas hazard tidak terpenuhi
- **Tingkat Signifikansi**
  $\alpha = 0.05$
- **Statistik Uji**

	test_statistic	p	-log2(p)
age			
km	0.47	0.49	1.02
rank	0.45	0.50	0.99
anxiety			
km	1.39	0.24	2.07
rank	1.05	0.30	1.72
coach			
km	1.13	0.29	1.80
rank	1.80	0.18	2.47
extraversion			
km	4.02	0.04	4.48
rank	3.11	0.08	3.68
gender			
km	0.23	0.63	0.66
rank	0.14	0.71	0.49
greywage			
km	0.63	0.43	1.22
rank	0.73	0.39	1.34
head_gender			
km	0.00	0.95	0.07
rank	0.04	0.84	0.26
independ			
km	0.66	0.42	1.26
rank	0.15	0.70	0.51
industry			
km	0.41	0.52	0.94
rank	0.77	0.38	1.39
novator			
km	2.65	0.10	3.28
rank	3.39	0.07	3.93
profession			
km	0.11	0.74	0.44
rank	0.14	0.70	0.51
selfcontrol			
km	5.75	0.02	5.92
rank	3.85	0.05	4.33
traffic			
km	0.89	0.35	1.54
rank	0.46	0.50	1.01
way			
km	0.78	0.38	1.40
rank	0.43	0.51	0.97

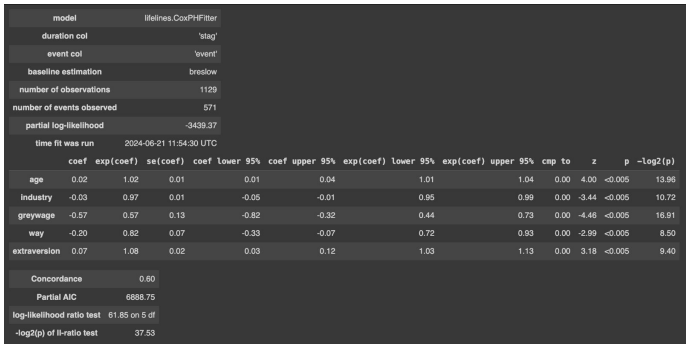
Tabel 5. Hasil Uji Residual Schoenfeld pada Data Pergantian Karyawan

- **Daerah Kritik**
  H_0 ditolak ketika p-value < 0.05
- **Kesimpulan**
  Setelah melakukan uji residual Schoenfeld, terdapat dua variabel yang tidak memenuhi uji asumsi proporsionalitas hazard. Variabel tersebut diantaranya adalah extraversion dan selfcontrol yang memiliki p-value sebesar 0.04 dan 0.02 secara berturut-turut. P-value yang didapatkan < $\alpha = 0.05$, sehingga H_0 ditolak variabel tersebut tidak memenuhi uji asumsi proporsionalitas hazard. Untuk variabel lainnya sudah lolos dalam uji asumsi proporsionalitas hazard. Kemudian kedua variabel tersebut tidak akan dilakukan stratifikasi. Stratifikasi dilakukan untuk membuat tingkatan pada data yang bersifat kategorik, dikarenakan variabel extraversion dan selfcontrol merupakan data numerik, maka tidak dilakukan stratifikasi. Untuk mendapatkan model terbaik, akan dilakukan eliminasi variabel tidak signifikan menggunakan *backward selection*. Oleh karena itu, setelah melakukan beberapa *fitting* menggunakan regresi cox dan turunannya, didapat hasil metrik evaluasi sebagai berikut.

Model	C-index	AIC	Partial AIC
Cox Model1	0.616	-	6896.6374
Cox Model2	0.616	-	6894.6679
Cox Model3	0.616	-	6892.8474
Cox Model4	0.616	-	6891.0071
Cox Model5	0.616	-	6889.2893
Cox Model6	0.615	-	6888.0008
Cox Model7	0.612	-	6887.7551
Cox Model8	0.602	-	6888.7499
Cox Model9	0.605	-	6887.5567
Cox Model10	0.602	-	6888.7499
Cox Spline (Penalized)	-	5969.0413	-
Cox Piecewise	-	5991.4181	-

Tabel 6. Metrik Evaluasi Kumpulan Model Regresi Cox

Model terbaik yang akan dipilih adalah model 10, dengan ringkasan regresi sebagai berikut.



model	lifelines.CoxPHFitter
duration col	'stag'
event col	'event'
baseline estimation	breslow
number of observations	1129
number of events observed	571
partial log-likelihood	-3439.37
time fit was run	2024-06-21 11:54:30 UTC

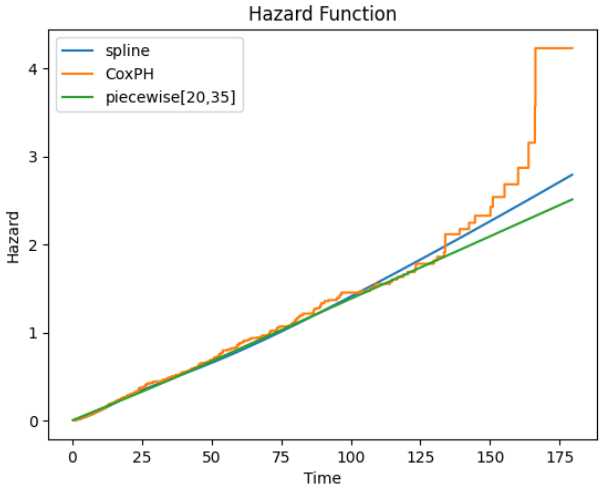
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
age	0.02	1.02	0.01	0.01	0.04	1.01	1.04	0.00	4.00	<0.005	13.96
industry	-0.03	0.97	0.01	-0.05	-0.01	0.95	0.99	0.00	-3.44	<0.005	10.72
greywage	-0.57	0.57	0.13	-0.82	-0.32	0.44	0.73	0.00	-4.46	<0.005	16.91
way	-0.20	0.82	0.07	-0.33	-0.07	0.72	0.93	0.00	-2.99	<0.005	8.50
extraversion	0.07	1.08	0.02	0.03	0.12	1.03	1.13	0.00	3.18	<0.005	9.40

Concordance	0.60
Partial AIC	6888.75
log-likelihood ratio test	61.85 on 5 df
-log2(p) of ll-ratio test	37.53

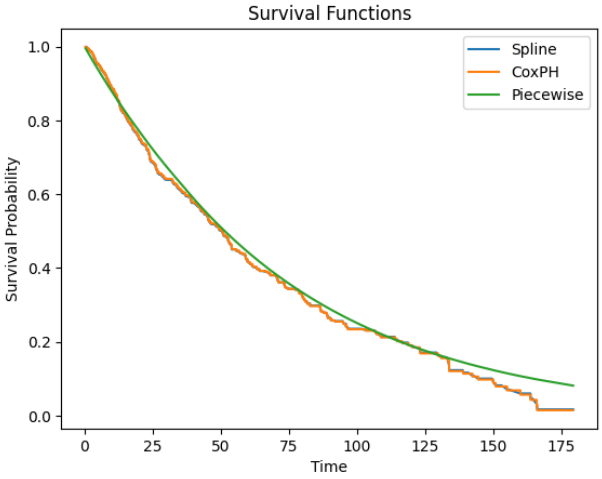
Gambar 2. Hasil Regresi Cox Model 10

Model terbaik (model10) ini merupakan hasil backward selection untuk menghapus variabel yang tidak signifikan. Oleh karena itu, didapatkan variabel signifikan yaitu age, industry, greywage, way dan extraversion. Kemudian, dapat juga ditampilkan prediksi fungsi hazard dan survival dari regresi cox untuk model terbaik (model 10) serta perbandingan dengan model turunan cox yang lain, yaitu model cox penalized dan piecewise.

$$h_0(t|x) = h_o(t) \cdot \exp(0.02 \cdot \text{age} - 0.03 \cdot \text{industry} - 0.57 \cdot \text{greywage} - 0.20 \cdot \text{way} + 0.07 \cdot \text{extraversion})$$



Gambar 3. Perbandingan Hazard Function untuk Beberapa Model



Gambar 4. Perbandingan Survival Function untuk Beberapa Model

4.2 Data Sirosis Hati

Untuk kasus data sirosis hati, akan diterapkan Random Survival Forest. Metode ini menggunakan pendekatan non-parametrik berupa Nelson-Aalen. Oleh karena itu, akan dibuat model regresi yang terdiri dari variabel independen (x) nya berupa kovariat-kovariat yang ada pada data sirosis hati, dan variabel dependen (y) berupa sebuah array yang berisi *censoring information* dan waktu observasi. Sebagai gambaran, berikut adalah array untuk variabel dependen (y) yang akan digunakan pada model RSF ini.

(False, 4500.)	(False, 1832.)	(False, 3577.)	(False, 3672.)
(False, 4232.)	(False, 3445.)	(False, 4127.)	(False, 4509.)
(False, 2468.)	(False, 2475.)	(False, 2241.)	(False, 2837.)
(False, 1435.)	(False, 1732.)	(False, 2737.)	(False, 1301.)
(False, 1542.)	(False, 1084.)	(False, 1447.)	(False, 1067.)
(False, 1901.)	(False, 1877.)	(False, 2533.)	(False, 3021.)
(False, 1617.)	(False, 2267.)	(False, 1347.)	(False, 1725.)
(False, 1022.)	(True, 2400.)	(True, 1012.)	(True, 1925.)

Tabel 7. Array untuk Variabel Independen (y)

Selanjutnya, akan dilakukan fitting menggunakan RSF pada X dan y, dengan X sudah dilakukan *encoding* untuk variabel kategorik. Proses ini menggunakan spesifikasi model RSF sebagai berikut.

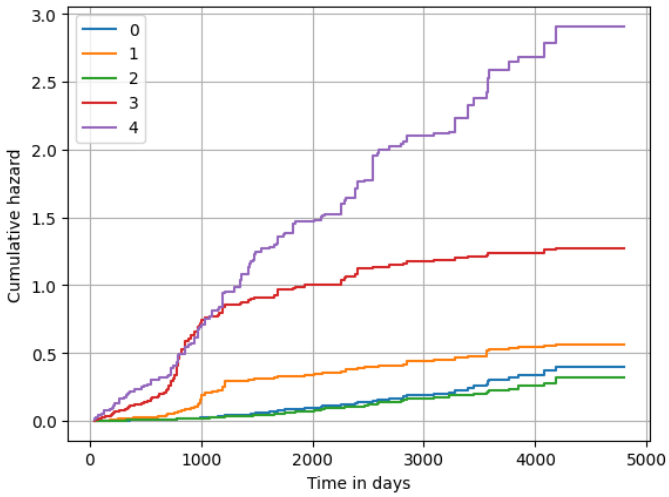
- max_depth = 3
- min_samples_leaf = 15
- min_samples_split = 10
- num_estimators = 1000
- num_jobs = 1
- random_state = 100

Didapatkan skor untuk random forest sebesar 0.94961 atau 94.961%. Selain itu, akan ditampilkan tabel *feature importance* atau sebuah tabel yang berisi variabel-variabel mana saja yang paling memengaruhi dalam penelitian. Penilaian tersebut dapat dilihat pada tabel terurut berikut.

	importances_mean	importances_std
Status	1.2675e-01	0.024348
Days	9.4157e-02	0.011420
Cohl	2.1450e-03	0.001151
Age	1.2701e-03	0.001483
Platelet	4.7981e-04	0.000706
Bill	2.5402e-04	0.004119
Trig	5.6449e-05	0.000211
Copper	-1.4802e-17	0.002120
Alk Phos	-2.5402e-04	0.001497
Albumin	-1.2701e-03	0.003103
Sgot	-1.7217e-03	0.001251
Protime	-3.9796e-03	0.002084

Tabel 8. Importance Mean dan Std Deviation pada RSF untuk Data Liver

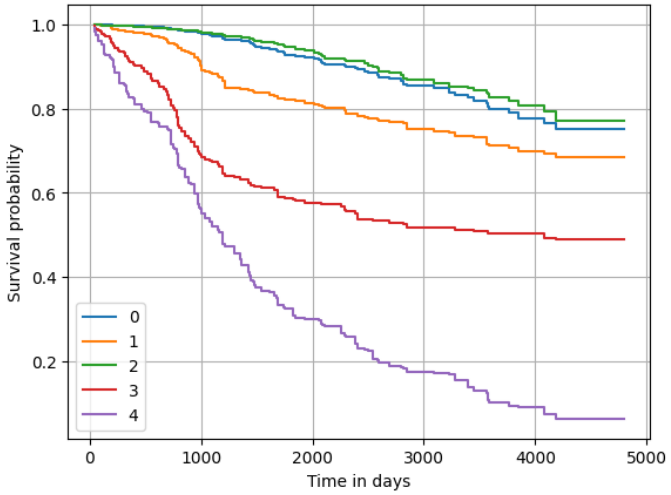
Berdasarkan tabel, kita telah mendapatkan bahwa variabel yang paling memengaruhi penelitian selain Days dan Status adalah Cohl dan Age. Oleh karena itu, mari kita lihat prediksi cummulative hazard dan survival secara berturut-turut untuk *head* dan *tail* dari 3 data terurut berdasar kolesterol terendah dan tertinggi.



Gambar 5. Prediksi Hazard untuk Kolesterol

Dengan melakukan fitting menggunakan RSF, maka didapatkan beberapa metrik evaluasi yang sebagai berikut.

Penggunaan Integrated Brier Score didasari pada keinginan untuk melihat perliaku survival dari pasien pengidap sirosis hati secara menyeluruh mulai dari waktu



Gambar 6. Prediksi Survival untuk Kolesterol

Metrik Evaluasi	Nilai
C-Index	0.95
Integrated Brier Score	0.081
Time Dependent AUC	0.952

Tabel 9. Metrik Evaluasi untuk Random Survival Forest

paling awal hingga waktu akhir. Hal ini dibuktikan dengan adanya deklarasi waktu penelitian mulai dari waktu 94 hingga 4556 dengan jangka pengamatan yaitu tahunan. Pendeklarasian interval waktu menyeluruh juga membantu dalam memunculkan metrik evaluasi lain seperti Time Dependent AUC.

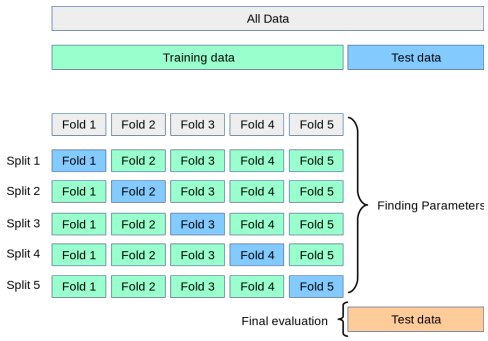
4.3 Cross Validation

Cross validation adalah metode yang digunakan untuk menghindari overfitting pada data uji serta menemukan parameter terbaik untuk model dalam penelitian. Teknik ini melibatkan pembagian dataset menjadi beberapa subset atau "folds", di mana model dilatih pada beberapa fold dan diuji pada fold yang tersisa. Proses ini diulang beberapa kali sehingga setiap fold digunakan sebagai data uji sekali, memberikan estimasi performa model yang lebih akurat.

GridSearch Cross Validation adalah jenis cross validation yang menggabungkan grid search dengan cross validation. Grid search adalah teknik untuk mencari parameter terbaik dari model dengan mencoba berbagai kombinasi parameter yang telah ditentukan sebelumnya. Dengan GridSearch Cross Validation, setiap kombinasi parameter dievaluasi menggunakan cross validation untuk mengukur performa model, sehingga memungkinkan pemilihan kombinasi parameter yang memberikan performa terbaik berdasarkan hasil cross validation.

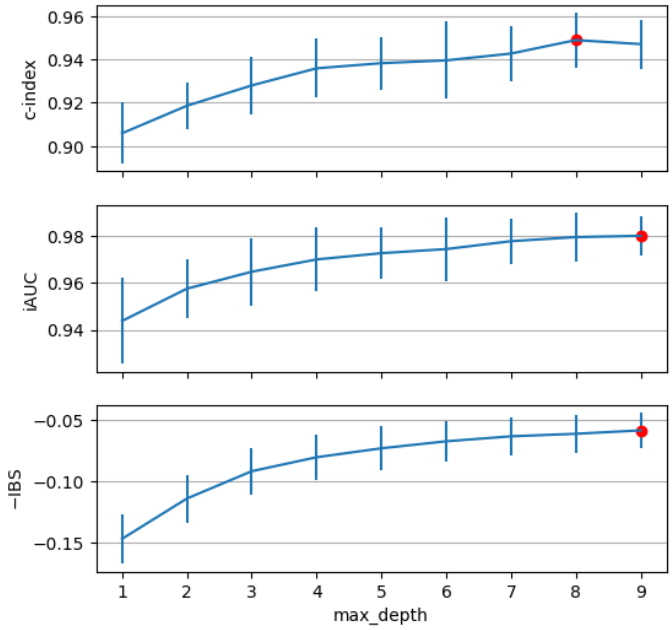
Keuntungan utama dari GridSearch Cross Validation adalah kemampuannya menghindari overfitting dan mengoptimalkan parameter model, memberikan estimasi performa yang lebih akurat dibandingkan evaluasi pada satu subset data saja. Teknik ini membantu menemukan kombinasi parameter terbaik untuk model, memastikan model yang dihasilkan memiliki performa optimal dan dapat diandalkan untuk prediksi pada data baru.

Mekanisme GridSearch Cross Validation dapat dilihat pada skema tersebut. Pada skema tersebut, data train akan dilatih menggunakan k-1 fold. Untuk kasus ini, kita menggunakan 5 fold dengan base max_ depth seperti di awal yaitu 3.



Gambar 7. Prediksi Hazard untuk Kolesterol

Selanjutnya, akan dilakukan cross validation untuk data sirosis hati pada beberapa metrik, diantaranya adalah IBS, C-Index dan Time Dependent AUC. Grafik yang menunjukkan max_depth terbaik untuk setiap metrik adalah sebagai berikut.



Gambar 8. Max_Depth Terbaik untuk Beberapa Metrik Evaluasi

Oleh karena itu, diperoleh max_depth terbaik untuk c-index adalah 8, Time Dependent AUC adalah 9, dan IBS adalah 9.

Setelah itu, dilakukan pemodelan ulang untuk mendapatkan hasil maksimal. Hasil dari pemodelan menggunakan maxdepth terbaik adalah sebagai berikut.

Metrik Evaluasi	Nilai
C-Index	0.95
Integrated Brier Score	0.070
Time Dependent AUC	0.967

Tabel 10. Metrik Evaluasi Setelah CV pada Random Survival Forest

Dengan melakukan cross validation, skor random forest

meningkat menjadi 0.96147 atau 96.147%. Selain itu, feature importance dari setiap variabel juga meningkat. Peningkatannya ditampilkan pada tabel berikut.

	importances_mean	importances_std
Status	0.134378	0.024015
Days	0.102681	0.013610
Cohl	0.003443	0.001196
Bill	0.003133	0.003514
Age	0.002314	0.001795
Alk Phos	0.001778	0.000955
Platelet	0.000931	0.001376
Copper	0.000536	0.001270
Trig	0.000056	0.000486
Sgot	-0.000536	0.001031
Albumin	-0.000762	0.003419
Protime	-0.001919	0.001338

Tabel 11. Imp Mean dan Standard Deviation Setelah Cross Validation

5 Kesimpulan

Pada analisis data pergantian karyawan, uji residual Schoenfeld dilakukan dan ditemukan bahwa dua variabel, yaitu extraversion dan selfcontrol, tidak memenuhi asumsi proporsionalitas hazard. Untuk variabel lainnya, asumsi ini telah terpenuhi. Variabel extraversion dan selfcontrol tidak di-stratifikasi karena keduanya adalah variabel numerik, bukan kategorik. Kemudian, menggunakan teknik backward selection, variabel yang tidak signifikan dieliminasi dari model. Model terbaik (model10) dihasilkan setelah beberapa kali fitting menggunakan regresi Cox dan turunannya dengan hasil akhir menunjukkan bahwa variabel yang signifikan adalah age, industry, profession, greywage, way dan extraversion.

Untuk kasus sirosis hati, digunakan metode Random Survival Forest (RSF) dengan pendekatan non-parametrik Nelson-Aalen. Hasil fitting menunjukkan skor RSF sebesar 0.94961 atau 94.961%. Analisis feature importance menunjukkan bahwa selain Days dan Status, variabel Cohl dan Age memiliki pengaruh paling signifikan. Metrik evaluasi untuk model RSF adalah sebagai berikut: C-Index sebesar 0.95, Integrated Brier Score sebesar 0.081, dan Time Dependent AUC sebesar 0.952. Penggunaan Integrated Brier Score dimaksudkan untuk melihat perilaku survival pasien pengidap sirosis hati dari waktu paling awal hingga akhir, dengan interval waktu penelitian dari 94 hingga 4556 hari.

Untuk menghindari overfitting dan menemukan parameter terbaik, digunakan teknik GridSearch Cross Validation dengan 5 fold. Cross validation meningkatkan skor RSF menjadi 96.147% dengan C-Index sebesar 0.95, Integrated Brier Score sebesar 0.070, dan Time Dependent AUC sebesar 0.967, menunjukkan peningkatan performa model yang signifikan.

Penelitian ini berhasil mengidentifikasi variabel-variabel signifikan yang memengaruhi pergantian karyawan dan survival pasien sirosis hati menggunakan model regresi Cox dan Random Survival Forest. Dengan teknik validasi dan pemodelan yang tepat, diperoleh model yang optimal untuk kedua kasus, menunjukkan kinerja yang baik berdasarkan metrik evaluasi yang digunakan.

Daftar Pustaka

[1] Danardono. (2012). Analisis Data Survival: Model Regresi Cox.

[2] Danardono. (2012). Diktat Analisis Data Survival.

[3] Dietz, K., Gail, M., Krickeberg, K., Samet, J., Tsiatis, A. (2000). Statistics for Biology and Health Series.

[4] Goswami, R., Dey, A. K. (2022). Integrated Brier Score based Survival Cobra – A regression based approach. http://arxiv.org/abs/2210.12006

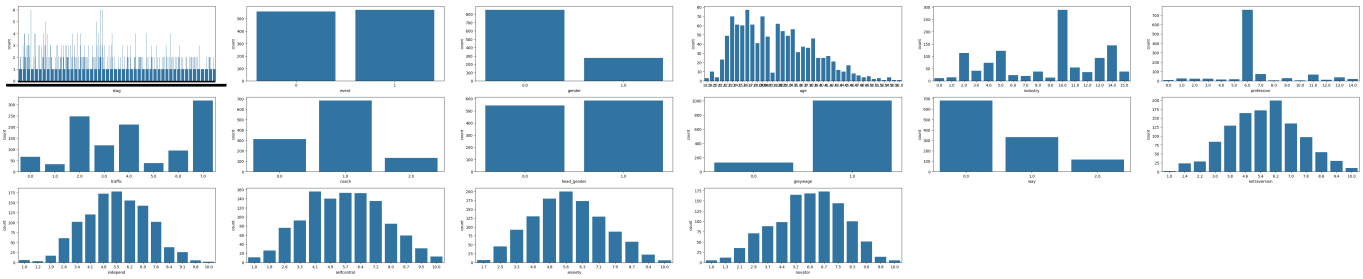
[5] Heagerty, P. J., Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves. In Biometrics (Vol. 61). https://academic.oup.com/biometrics/article/61/1/92/7305694

[6] Husnaqilati, A. (2024). Evaluation Metrics.

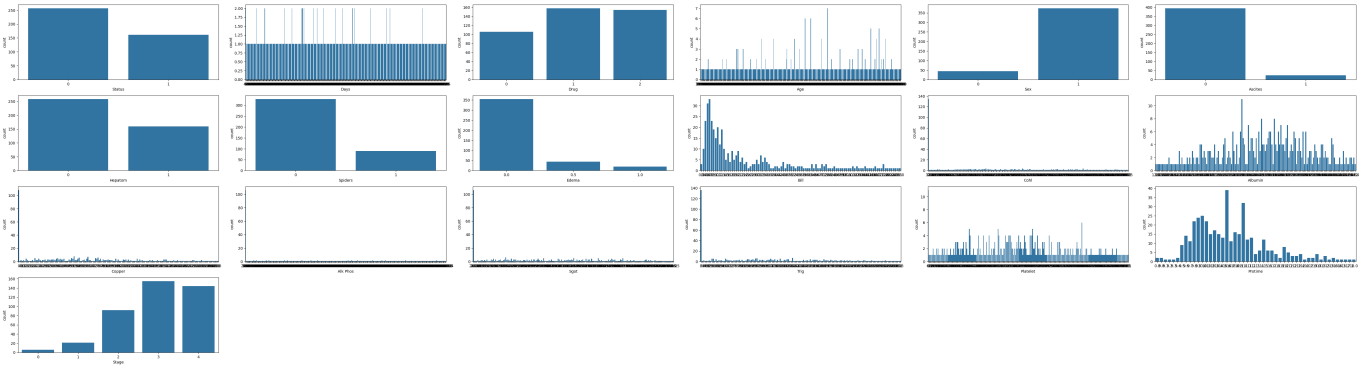
[7] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. (2008). Random survival forests. Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/08-AOAS169

[8] Abd ElHafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G., Roumeliotis, S. (2021). Methods to analyze time-to-event data: the Cox regression analysis. Oxidative medicine and cellular longevity, 2021(1), 1302811.

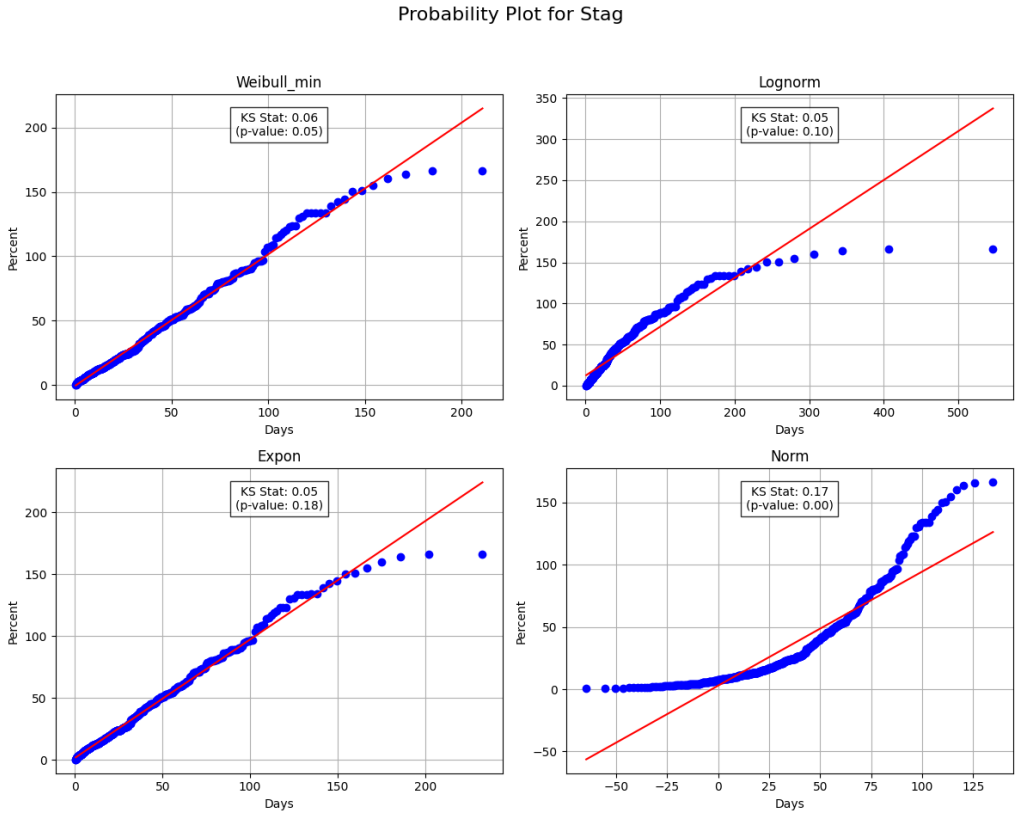
Apendiks A



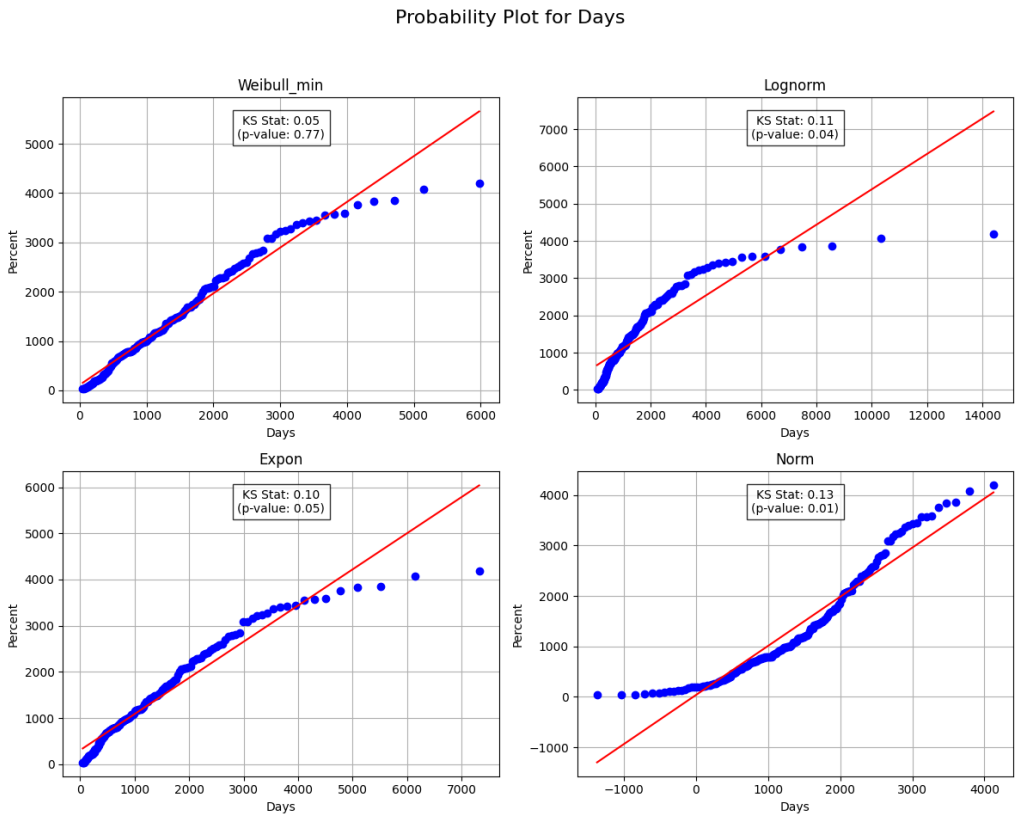
Gambar A1. Countplot Data Turnover



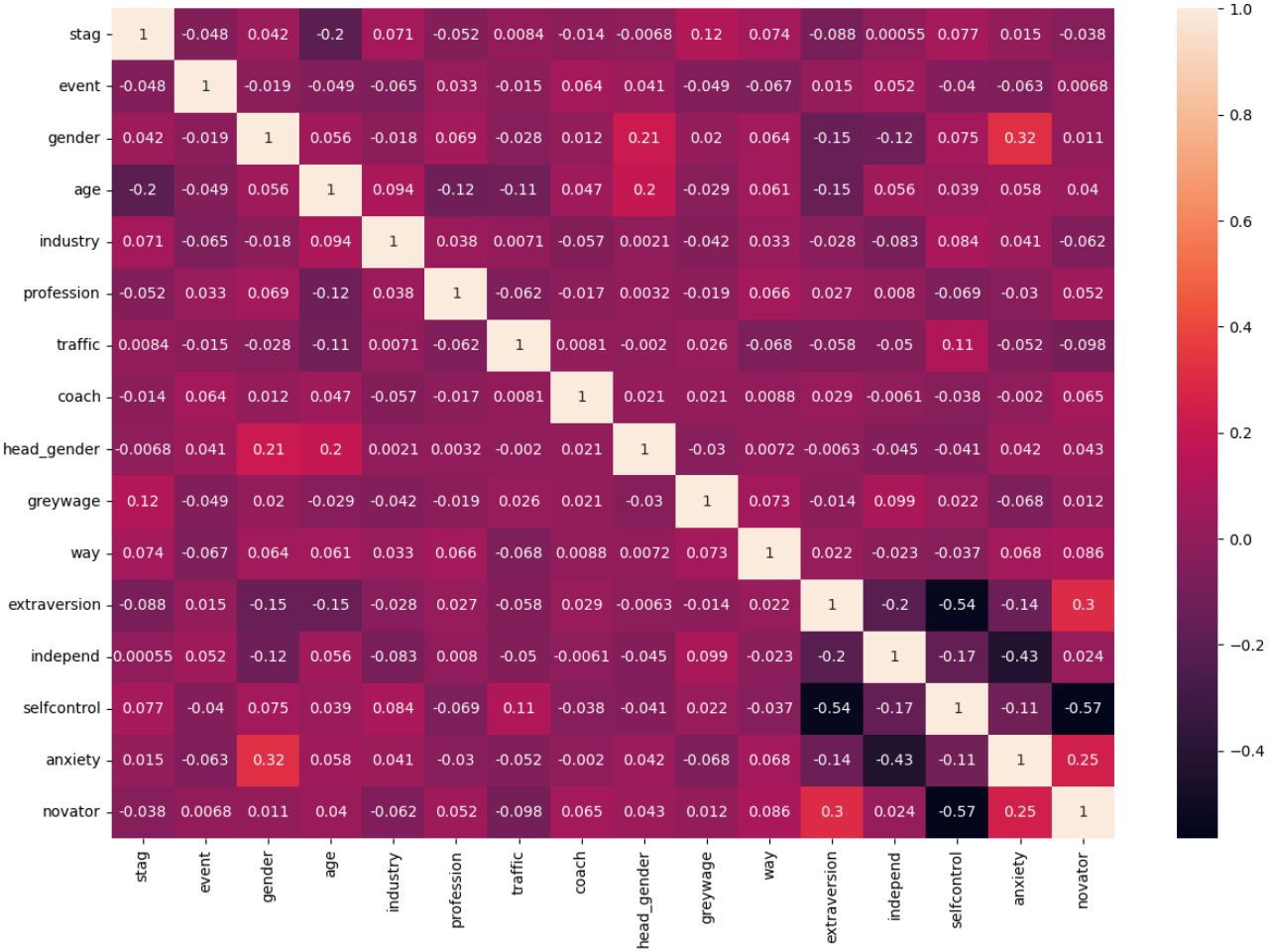
Gambar A2. Countplot Data Sirosis Hati



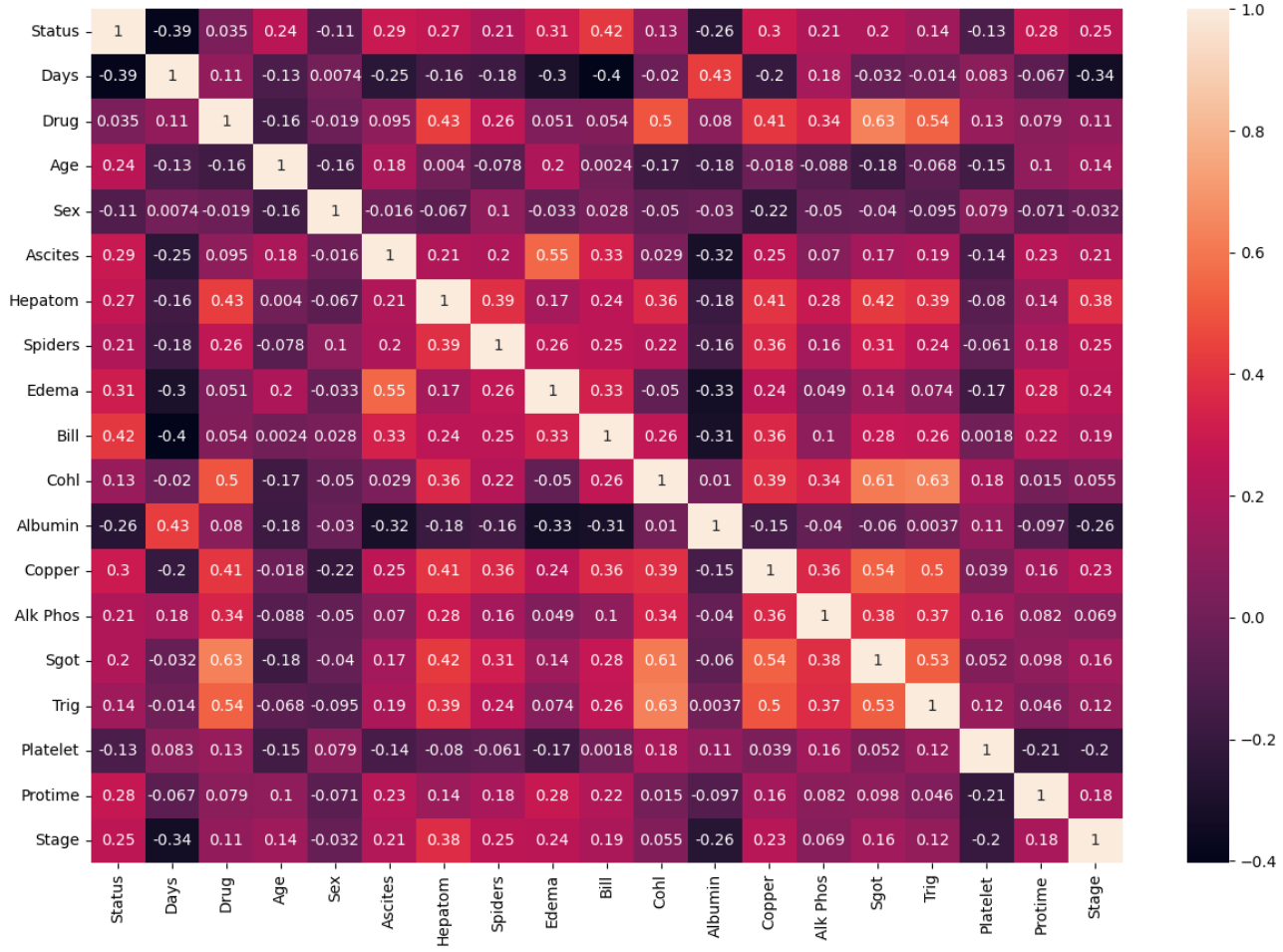
Gambar A3. Probability Plot untuk Stag pada Data Turnover



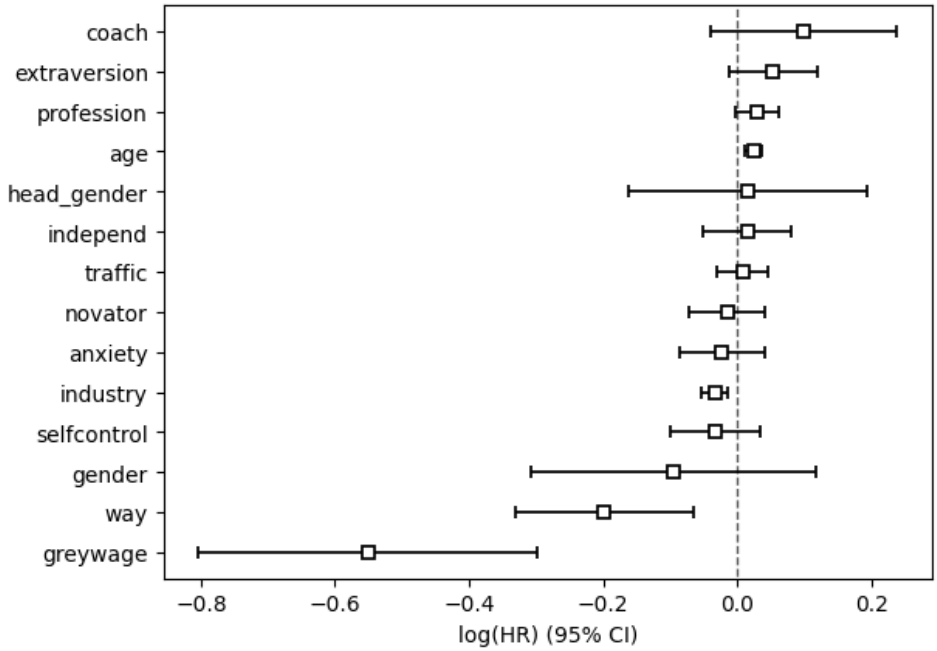
Gambar A4. Probability Plot untuk Days pada Data Sirosis Hati



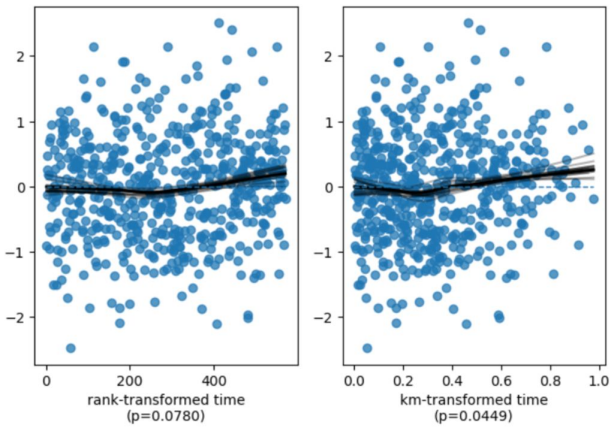
Gambar A5. Heatmap Data Turnover



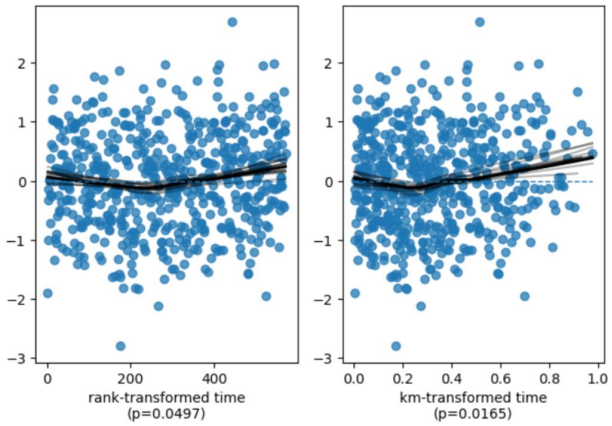
Gambar A6. Heatmap Data Sirosis Hati



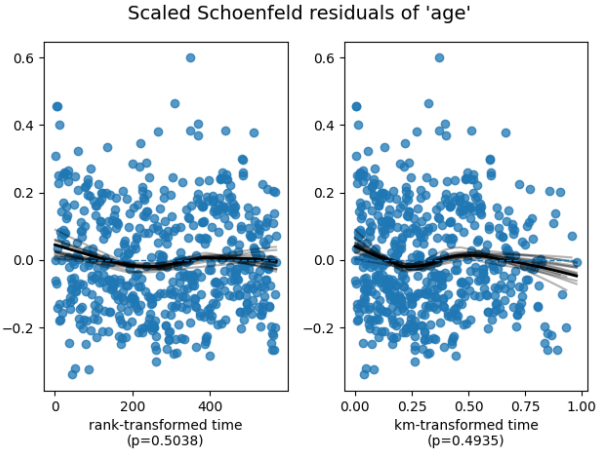
Gambar A7. Cox Proportional Hazard untuk Data Turnover



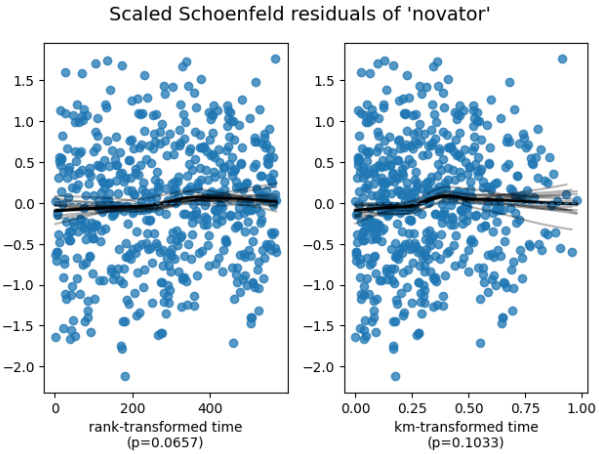
Gambar A8. Scaled Schoenfeld untuk extraversion



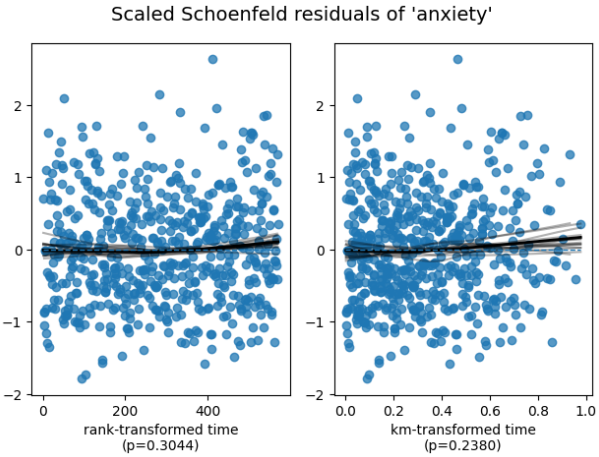
Gambar A9. Scaled Schoenfeld untuk selfcontrol



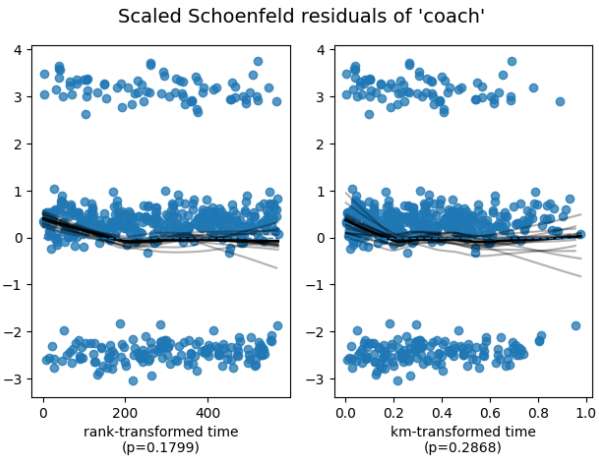
Gambar A10. Scaled Schoenfeld untuk age



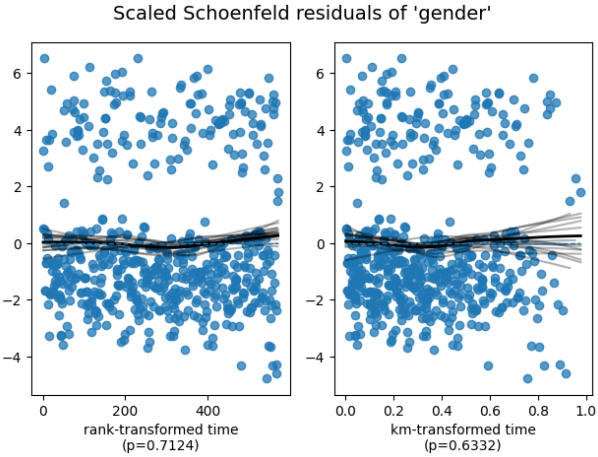
Gambar A11. Scaled Schoenfeld untuk novator



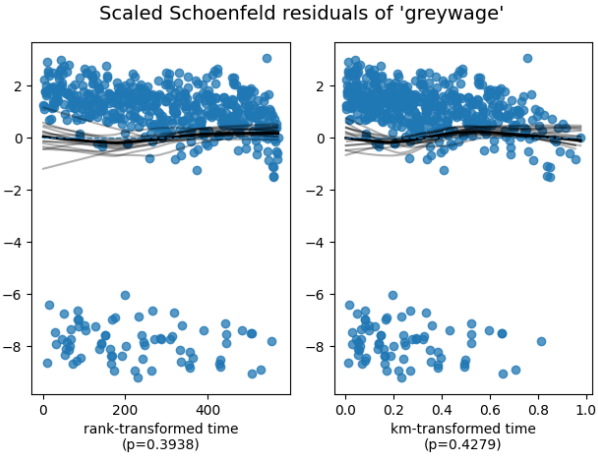
Gambar A12. Scaled Schoenfeld untuk anxiety



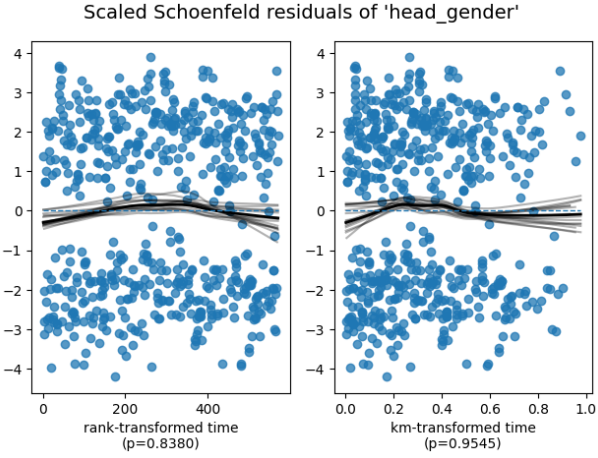
Gambar A13. Scaled Schoenfeld untuk coach



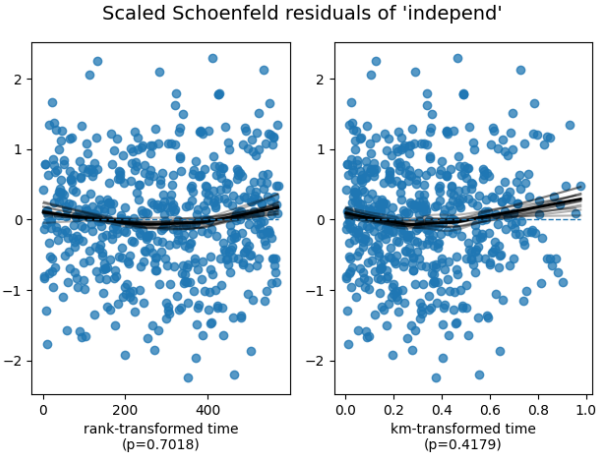
Gambar A14. Scaled Schoenfeld untuk gender



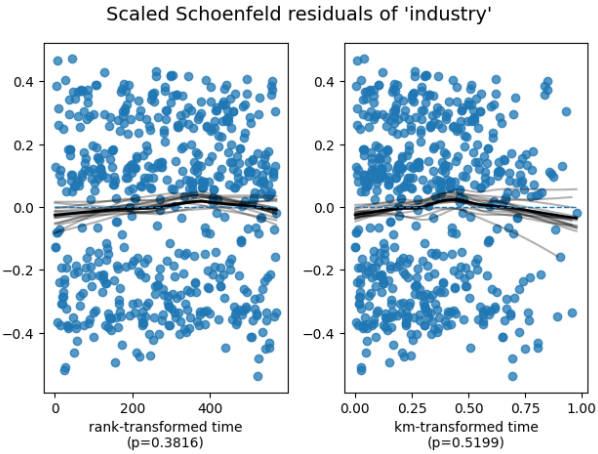
Gambar A15. Scaled Schoenfeld untuk greywage



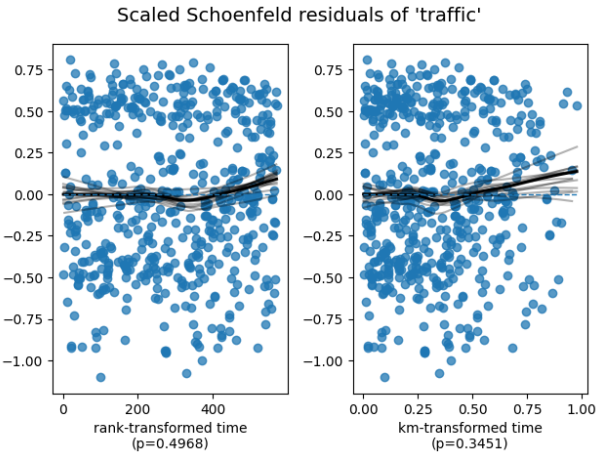
Gambar A16. Scaled Schoenfeld untuk head$_{g}ender$



Gambar A17. Scaled Schoenfeld untuk independ



Gambar A18. Scaled Schoenfeld untuk industry



Gambar A19. Scaled Schoenfeld untuk traffic