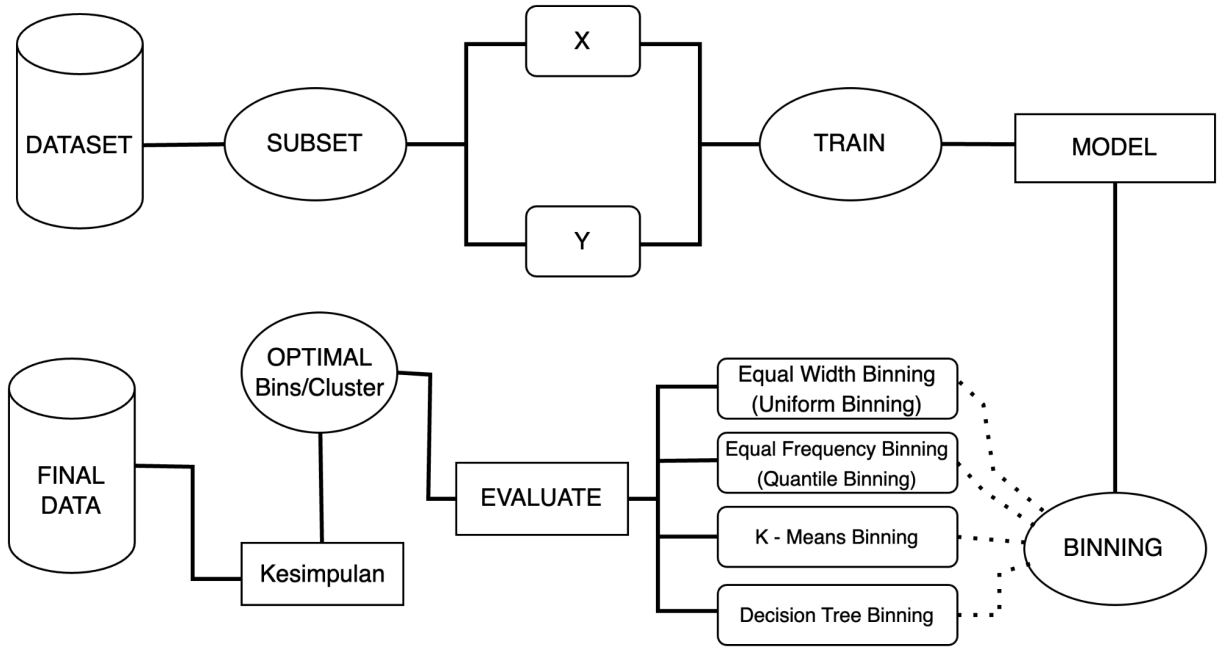
A.) Alur Pengkategorian Numerical Value (Belum Diskrit):



Penjelasan :

1. Diberikan dataset , heart disease : [Heart F.xlsx]
2. Subset Dataset , mengambil yang belum dikategorikan:
3. Memisahkan menjadi X dan Y:
    - X: Variabel Prediktor
    - Y: Variabel Respon
4. Membuat subset Train/Test untuk pemodelan.
5. Menentukan beberapa model machine learning untuk evaluasi.
6. Melakukan Binning menggunakan beberapa metode:
    - Equal Width Binning (Pembagian Lebar yang Sama)
    - Equal Frequency Binning (Pembagian Frekuensi yang Sama)
    - K-Means Binning
    - Decision Tree Binning
7. Menilai setiap Binning/Cluster/Leaf pada setiap model machine learning yang ditentukan.
8. Memilih model terbaik sebagai referensi untuk mengevaluasi Cluster/Bin/Leaf yang optimal.
9. Memilih Bins/Cluster/Leaf optimal untuk semua variabel.
10. Menarik kesimpulan dan menentukan kategori binning/klaster untuk setiap variabel.
11. Mengekspor/Mengunduh Data Final setelah Binning.

Pengkategorian Numerical Value (belum diskrit):

1. Dataset

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358.03	1.5	138	0	0	10	1

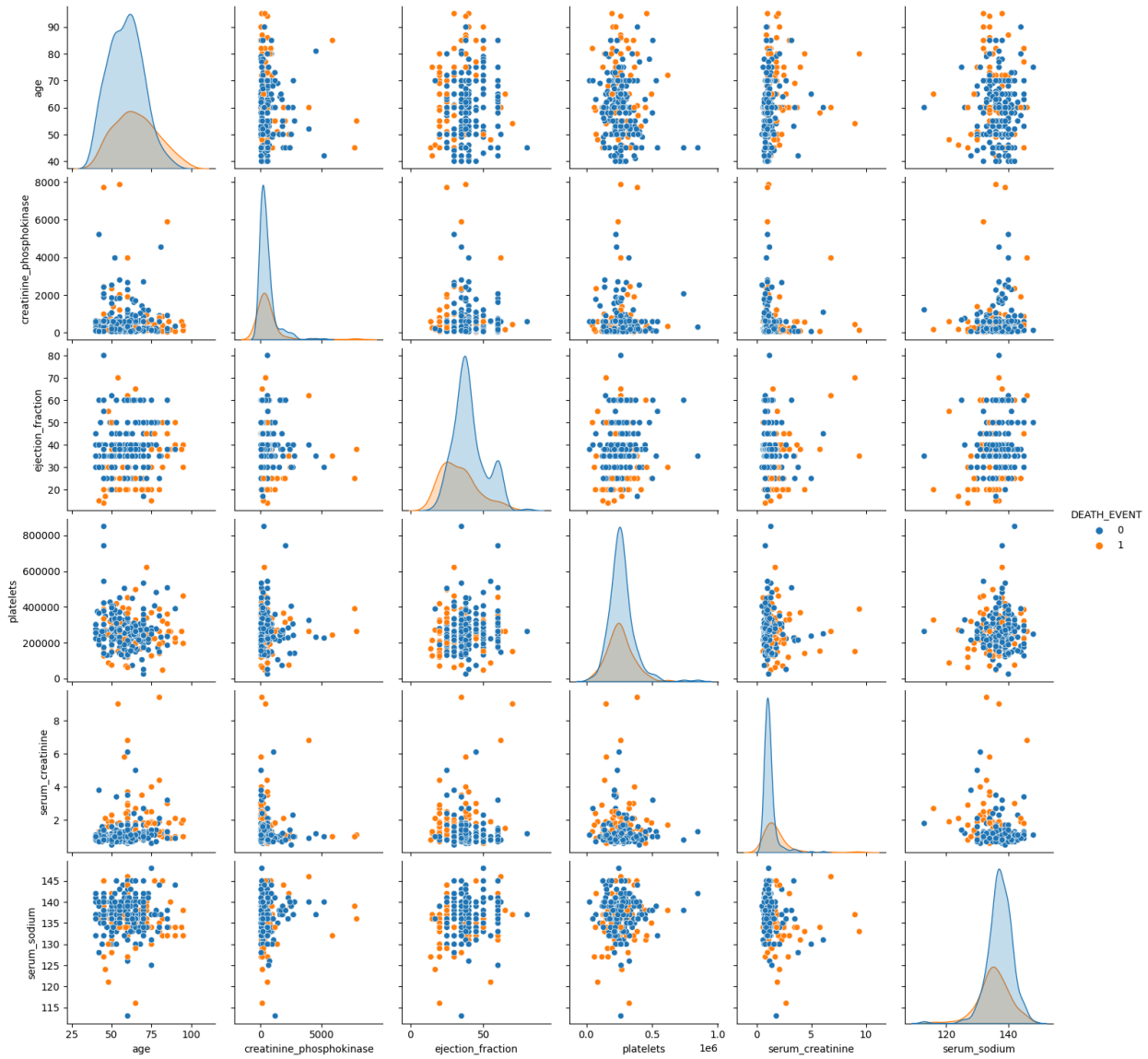
2. Subset Dataset , mengambil yang belum dikategorikan:

age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	DEATH_EVENT
75	582	20	265000	1.9	130	1
55	7861	38	263358.03	1.1	136	1
65	146	20	162000	1.3	129	1
50	111	20	210000	1.9	137	1
65	160	20	327000	2.7	116	1
90	47	40	204000	2.1	132	1
75	246	15	127000	1.2	137	1
60	315	60	454000	1.1	131	1
65	157	65	263358.03	1.5	138	1

Visualisasi Subset:

- **Pair Plot**

```
sns.pairplot(df , markers='o' , diag_kind='kde' , hue='DEATH_EVENT')
```



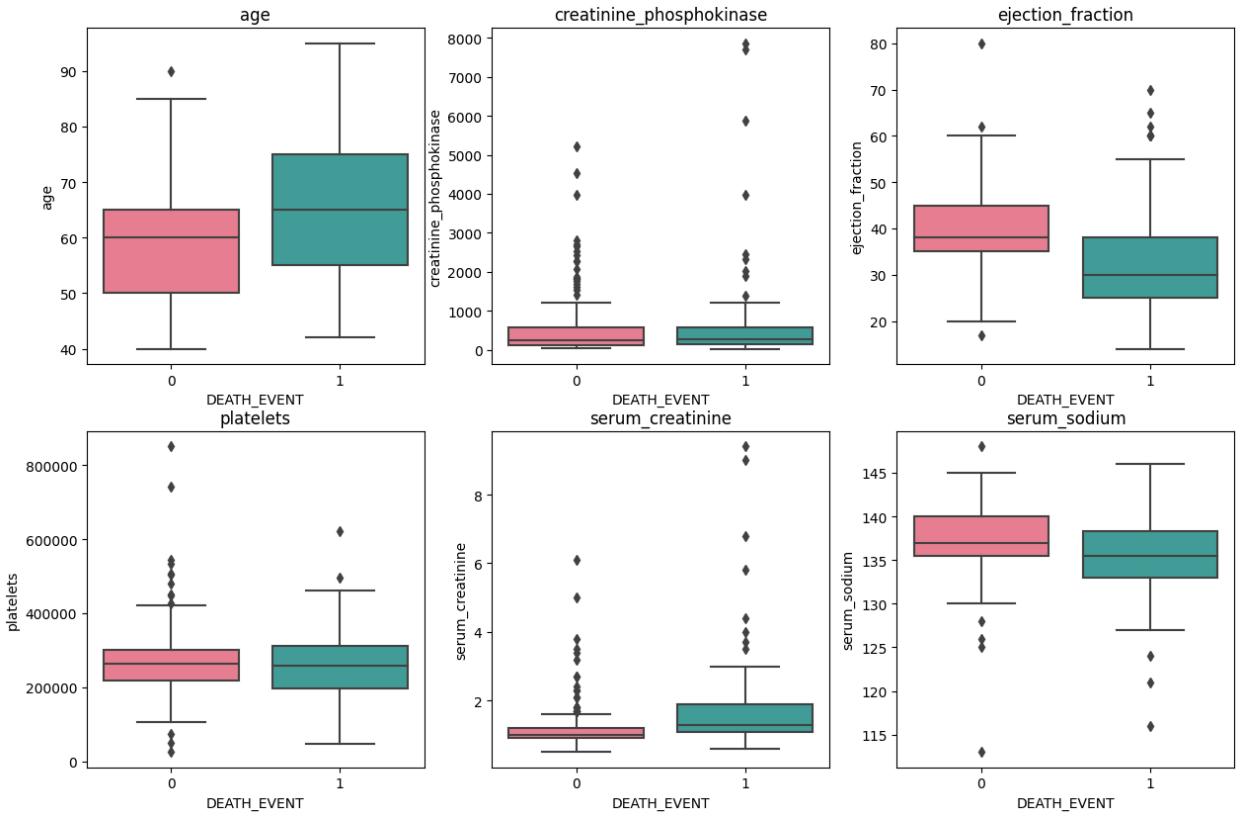
Interpretasi:

Tujuan pembuatan pairplot ini adalah untuk melihat persebaran data death event pada setiap variabel yang belum dikategorikan. Dapat dilihat juga bahwa Death_Event yang bernilai 0 lebih dominan pada seluruh variabel ini menunjukan bahwa lebih banyak pasien yang hidup daripada mati.

- **Box Plot**

```
fig, axes = plt.subplots(3,3, figsize=(15, 15))
axes = axes.flatten()

for i, column in enumerate(df.iloc[:, :6].columns):
```

```
    sns.boxplot(x='DEATH_EVENT', y=column, data=df, palette='husl', ax=axes[i])
    axes[i].set_title(column)

for j in range(i+1, 3*3):
    fig.delaxes(axes[j])
```



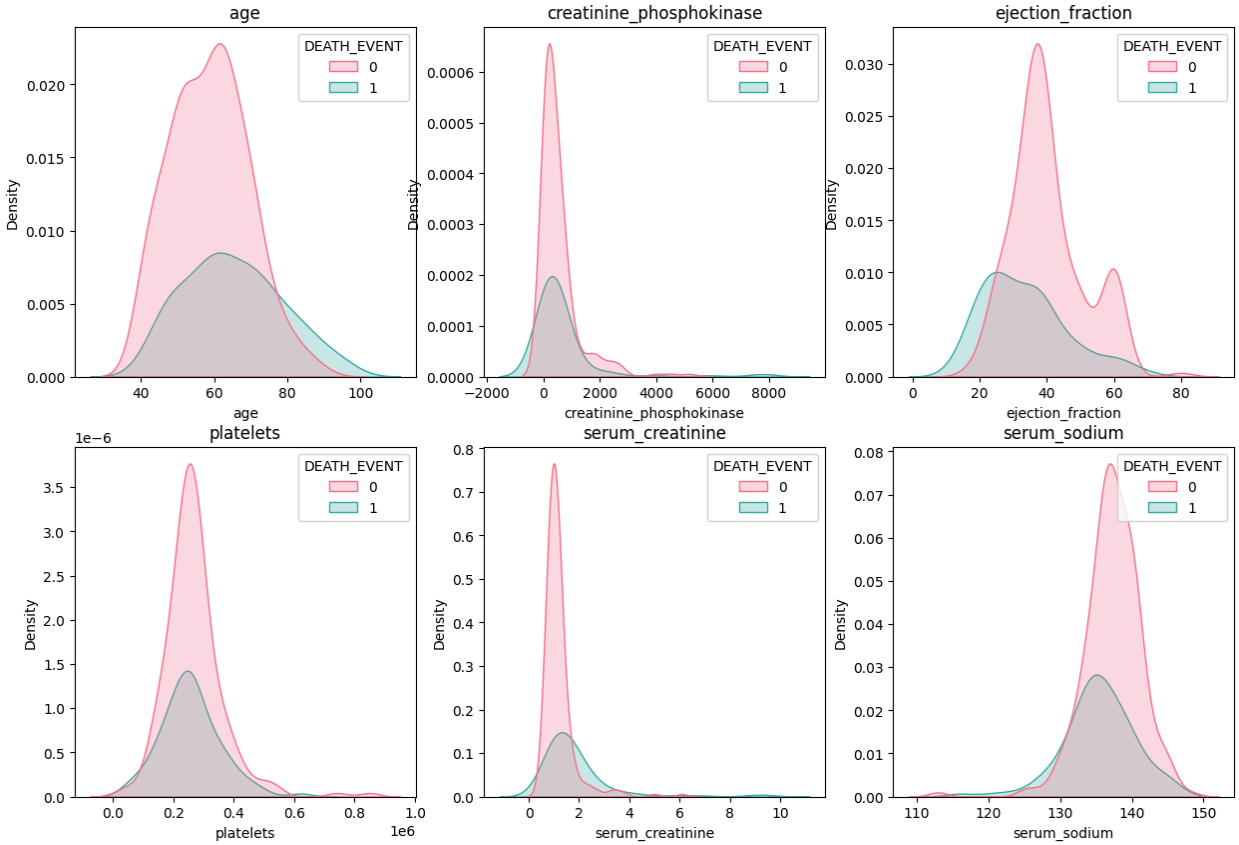
Interpretasi:

Tujuannya dibuat boxplot adalah untuk melihat outlier dari variabel yang sudah disubset. Dari gambar terlihat terdapat beberapa outlier pada setiap variabel. Namun , pada kasus kali ini kami tidak menghapus atau melakukan tindakan terhadap outlier tersebut. Karena kami tidak memiliki pemahaman mendalam terkait data yang nilainya abnormal, sehingga penanganan outliernya tidak dilakukan.

- **Kernel Density Estimate Plot:**

```
columns_to_plot = [col for col in df.columns if col != 'DEATH_EVENT']

num_columns = len(columns_to_plot)
num_rows = (num_columns + 2) // 3
```

```
fig, axes = plt.subplots(nrows=num_rows, ncols=3, figsize=(15, 5 * num_rows))
axes = axes.flatten()

for i, column in enumerate(columns_to_plot):
  sns.kdeplot(data=df, x=column, hue='DEATH_EVENT', ax=axes[i], palette='husl', shade=True)
  axes[i].set_title(column)

for j in range(i + 1, len(axes)):
  fig.delaxes(axes[j])
```

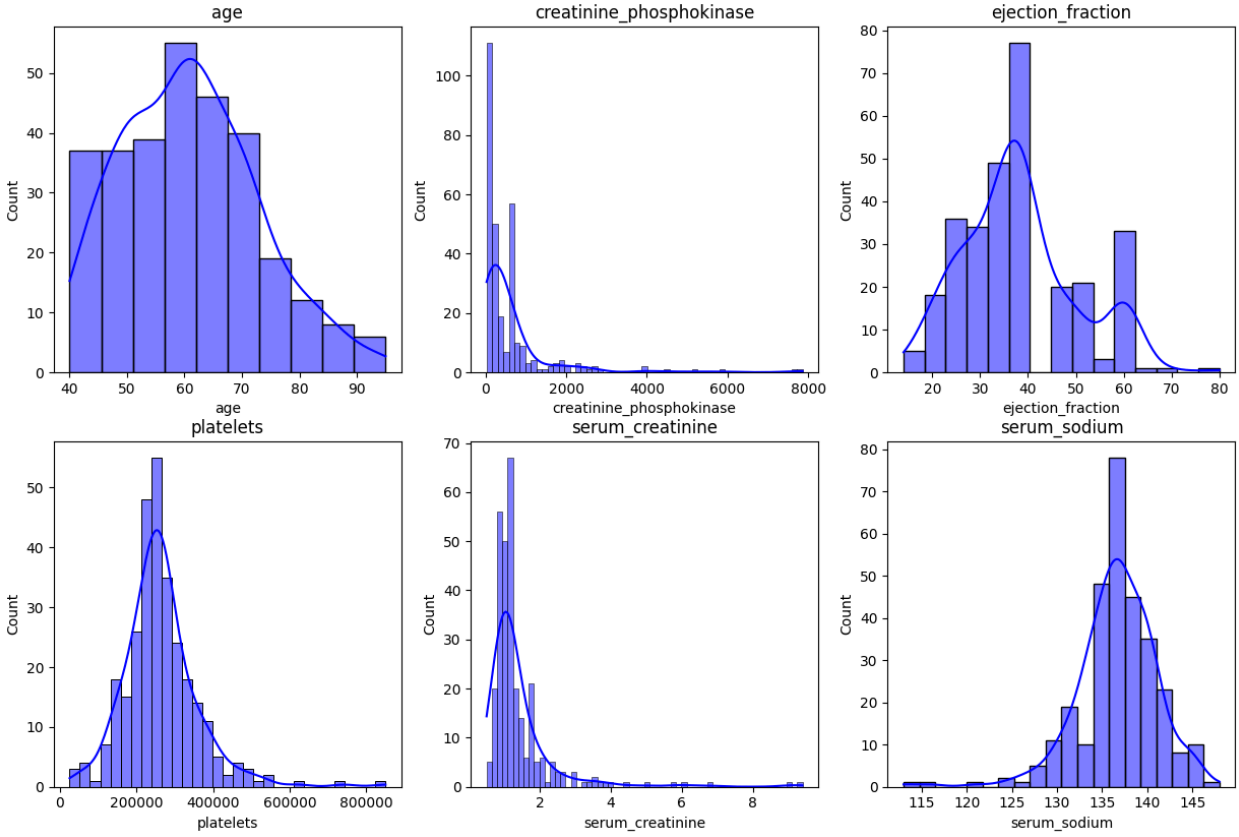


Interpretasi:

Tujuan dari pembuatan plot KDE adalah untuk mengamati distribusi dari semua variabel dengan mempertimbangkan Death Event. Dari plot tersebut, terlihat bahwa terdapat perbedaan dalam distribusi nilai Death Event. Lebih jelasnya, ketika Death Event = 0 dibandingkan dengan Death Event = 1, dapat diamati bahwa terdapat lebih banyak pasien yang bertahan hidup daripada pasien yang meninggal.

- **Histogram Plot:**

```
fig, axes = plt.subplots(3,3, figsize=(15, 15))
axes = axes.flatten()

for i, column in enumerate(df.iloc[:, :6].columns):
  sns.histplot(x=column, data=df, color='blue', ax=axes[i], kde=True)
  axes[i].set_title(column)

for j in range(i+1, 3*3):
  fig.delaxes(axes[j])
```



Interpretasi:

Tujuan dari pembuatan Histogram Plot ini adalah untuk mengamati distribusi keseluruhan dari setiap variabel yang telah dipilih kecuali Death Event.

3. Memisahkan menjadi X dan Y:

```
X = df.drop('DEATH_EVENT', axis=1)
y = df['DEATH_EVENT']
```

X:

age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
75	582	20	265000	1.9	130
55	7861	38	263358.03	1.1	136
65	146	20	162000	1.3	129
50	111	20	210000	1.9	137
65	160	20	327000	2.7	116
90	47	40	204000	2.1	132
75	246	15	127000	1.2	137
60	315	60	454000	1.1	131
65	157	65	263358.03	1.5	138

Y:

DEATH_EVENT
1
1
1
1
1
1
1
1
1

4. Membuat subset Train/Test untuk pemodelan.

```
X_train , X_test , y_train , y_test = train_test_split(X , y , test_size=0.1 , random_state=42)
```

Pada bagian ini subset Train dan Test dibagi menjadi proporsi 0.9 : 0.1 artinya 90% data dilakukan untuk train dan 10% dilakukan untuk testing. Pemilihan proporsi ini didasarkan oleh jumlah observasi dari data yang diperoleh. Karena data hanya ada 299 observasi ini menunjukan bahwa observasi sangat sedikit, sehingga diperlukan data train yang lebih banyak untuk model machine learning yang akan dibuat agar model dapat lebih menangkap pemahaman / mempelajari data lebih baik.

5. Menentukan beberapa model machine learning untuk evaluasi.

Pada bagian ini model machine learning yang diajukan adalah:

1. Random Forest Classifier
2. Gaussian Naive Bayes
3. XGBoost Classifier
4. AdaBoost Classifier
5. Extra Trees Classifier

Alasan pemilihan model ini adalah karena kemampuannya dalam mengklasifikasi beberapa variabel prediktor untuk akhirnya menentukan variabel respons dengan akurat. Classifier ini akan mengelompokkan semua variabel prediktor untuk akhirnya memperoleh variabel respons yang tepat.

Selain itu, model tersebut dapat memahami pola dari berbagai variabel prediktor yang digunakan untuk memprediksi nilai variabel respons. Sehingga, akan dipilih hanya satu model terbaik yang mampu melakukan prediksi variabel prediktor dengan tepat pada akhirnya.

Metrik Penilaian : F1-Score dan ROC AUC

- F1 Score: rata-rata harmonis dari precision dan recall. F1 Score adalah ukuran keseimbangan antara precision dan recall, dengan nilai yang tinggi menunjukkan keseimbangan yang baik antara kedua metrik tersebut.

- ROC AUC: mengukur kemampuan model machine learning untuk membedakan antara data positif dan negatif. Skor ROC AUC berkisar antara 0 dan 1, dengan 1 menunjukkan model machine learning yang sempurna dan 0,5 menunjukkan model machine learning yang buruk.

- Precision: mengukur rasio prediksi positif yang benar dari semua prediksi positif. Precision memberikan informasi tentang seberapa sering model machine learning membuat prediksi positif yang benar. Nilai precision berkisar antara 0 dan 1.

- Recall: mengukur rasio dari data positif yang benar yang ditemukan dari seluruh data positif. Recall memberikan informasi tentang seberapa baik model machine learning menemukan semua data positif. Nilai recall berkisar antara 0 dan 1

Alasan memilih metrik F1 Score dan ROC AUC adalah dikarenakan F1 Score dan ROC AUC score mampu mendiagnosa model dengan baik.

Sehingga diperoleh:

Model	F1 Score	ROC AUC
Random Forest	0.4	0.699074
Naive Bayes	0.352941	0.740741
XGBoost	0.315789	0.694444
AdaBoost	0.47619	0.597222
Extra Trees	0.315789	0.726852

Dapat diamati bahwa semua model memiliki nilai ROC AUC di atas 0.5 dan menunjukkan kinerja yang baik dengan F1 Score yang memuaskan. Oleh karena itu, dapat disimpulkan bahwa semua model tersebut memenuhi standar yang telah ditetapkan. Hal ini juga menandakan bahwa model mampu menjaga keseimbangan antara Precision dan Recall dengan baik. Selain itu, model juga dapat memberikan penilaian yang baik terhadap kinerja model machine learning yang kami gunakan.

6. Melakukan Binning menggunakan beberapa metode:

- Equal Width Binning (Uniform Binning):

Model	Number of Bins	F1 Score	ROC AUC
Random Forest	2	0.444444	0.6875
Random Forest	3	0.5	0.833333
Random Forest	4	0	0.722222
Random Forest	5	0.6	0.708333
Naive Bayes	2	0.285714	0.724537
Naive Bayes	3	0.266667	0.768519

Naive Bayes	4	0.444444	0.74537
Naive Bayes	5	0.333333	0.666667
XGBoost	2	0.526316	0.6875
XGBoost	3	0.5	0.796296
XGBoost	4	0.142857	0.678241
XGBoost	5	0.571429	0.712963
AdaBoost	2	0.153846	0.706019
AdaBoost	3	0.5	0.851852
AdaBoost	4	0.375	0.736111
AdaBoost	5	0.444444	0.733796
Extra Trees	2	0.444444	0.668981
Extra Trees	3	0.5	0.814815
Extra Trees	4	0	0.715278
Extra Trees	5	0.555556	0.675926

Penilaian Berdasarkan ROC AUC terbaik:

Model	Number of Bins	F1 Score	ROC AUC
AdaBoost	3	0.5	0.851852
Extra Trees	3	0.5	0.814815
Naive Bayes	3	0.266667	0.768519
Random Forest	3	0.5	0.833333
XGBoost	3	0.5	0.796296

Penilaian Berdasarkan F1 Score terbaik:

Model	Number of Bins	F1 Score	ROC AUC
AdaBoost	3	0.5	0.851852
Extra Trees	5	0.555556	0.675926
Naive Bayes	4	0.444444	0.74537
Random Forest	5	0.6	0.708333
XGBoost	5	0.571429	0.712963

- Equal Frequency Binning (Quantile Binning):

Model	Number of Bins	F1 Score	ROC AUC
Random Forest	2	0.25	0.625
Random Forest	3	0.375	0.763889
Random Forest	4	0.5	0.719907
Random Forest	5	0.571429	0.766204
Naive Bayes	2	0.47619	0.615741
Naive Bayes	3	0.6	0.759259
Naive Bayes	4	0.421053	0.75463
Naive Bayes	5	0.5	0.777778
XGBoost	2	0.333333	0.62963
XGBoost	3	0.470588	0.699074
XGBoost	4	0.608696	0.736111
XGBoost	5	0.615385	0.722222
AdaBoost	2	0.266667	0.601852
AdaBoost	3	0.555556	0.824074
AdaBoost	4	0.421053	0.759259
AdaBoost	5	0.47619	0.736111
Extra Trees	2	0.25	0.627315
Extra Trees	3	0.285714	0.664352
Extra Trees	4	0.333333	0.685185
Extra Trees	5	0.571429	0.747685

Penilaian Berdasarkan ROC AUC terbaik:

Model	Number of Bins	F1 Score	ROC AUC
AdaBoost	3	0.555556	0.824074
Extra Trees	3	0.285714	0.664352
Naive Bayes	3	0.6	0.759259
Random Forest	3	0.375	0.763889
XGBoost	3	0.470588	0.699074

Penilaian Berdasarkan F1 Score terbaik:

Model	Number of Bins	F1 Score	ROC AUC

AdaBoost	3	0.555556	0.824074
Extra Trees	5	0.571429	0.747685
Naive Bayes	4	0.421053	0.75463
Random Forest	5	0.571429	0.766204
XGBoost	5	0.615385	0.722222

- K-Means Binning:

Model	Number of Clusters	F1 Score	ROC AUC
Random Forest	2	0.352941	0.787037
Random Forest	3	0.4	0.652778
Random Forest	4	0.5	0.643519
Random Forest	5	0.434783	0.712963
Naive Bayes	2	0.375	0.597222
Naive Bayes	3	0.315789	0.587963
Naive Bayes	4	0.4	0.671296
Naive Bayes	5	0.133333	0.5
XGBoost	2	0.352941	0.787037
XGBoost	3	0.421053	0.680556
XGBoost	4	0.5	0.569444
XGBoost	5	0.454545	0.703704
AdaBoost	2	0.352941	0.717593
AdaBoost	3	0.444444	0.736111
AdaBoost	4	0.5	0.75
AdaBoost	5	0.521739	0.715278
Extra Trees	2	0.375	0.787037
Extra Trees	3	0.315789	0.611111
Extra Trees	4	0.5	0.678241
Extra Trees	5	0.56	0.69213

Penilaian Berdasarkan ROC AUC terbaik:

Model	Number of Clusters	F1 Score	ROC AUC
AdaBoost	3	0.444444	0.736111

Extra Trees	3	0.315789	0.611111
Naive Bayes	3	0.315789	0.587963
Random Forest	3	0.4	0.652778
XGBoost	3	0.421053	0.680556

Penilaian Berdasarkan F1 Score terbaik:

Model	Number of Clusters	F1 Score	ROC AUC
AdaBoost	3	0.444444	0.736111
Extra Trees	5	0.56	0.69213
Naive Bayes	4	0.4	0.671296
Random Forest	5	0.434783	0.712963
XGBoost	5	0.454545	0.703704

- Decision Tree Binning:

Model	Max Leaf Nodes	F1 Score	ROC AUC
Random Forest	2	0.333333	0.509259
Random Forest	3	0.4	0.652778
Random Forest	4	0.4	0.703704
Random Forest	5	0.4	0.678241
Naive Bayes	2	0.133333	0.643519
Naive Bayes	3	0.235294	0.652778
Naive Bayes	4	0.222222	0.712963
Naive Bayes	5	0.133333	0.752315
XGBoost	2	0.235294	0.462963
XGBoost	3	0.3	0.597222
XGBoost	4	0.47619	0.666667
XGBoost	5	0.521739	0.648148
AdaBoost	2	0.315789	0.652778
AdaBoost	3	0.315789	0.699074
AdaBoost	4	0.421053	0.685185
AdaBoost	5	0.315789	0.608796
Extra Trees	2	0.333333	0.513889

Extra Trees	3	0.421053	0.643519
Extra Trees	4	0.315789	0.694444
Extra Trees	5	0.333333	0.671296

Penilaian Berdasarkan ROC AUC terbaik:

Model	Max Leaf Nodes	F1 Score	ROC AUC
AdaBoost	3	0.315789	0.699074
Extra Trees	4	0.315789	0.694444
Naive Bayes	5	0.133333	0.752315
Random Forest	4	0.4	0.703704
XGBoost	4	0.47619	0.666667

Penilaian Berdasarkan F1 Score terbaik:

Model	Max Leaf Nodes	F1 Score	ROC AUC
AdaBoost	4	0.421053	0.685185
Extra Trees	3	0.421053	0.643519
Naive Bayes	3	0.235294	0.652778
Random Forest	3	0.4	0.652778
XGBoost	5	0.521739	0.648148

7. Menilai setiap Binning/Cluster/Leaf pada setiap model machine learning yang ditentukan.

Dari langkah keenam, terlihat bahwa F1 Score dari berbagai metode binning tidak mengalami perubahan yang signifikan dan belum mencapai tingkat kinerja yang optimal. Oleh karena itu, fokus kami beralih pada metrik evaluasi ROC AUC untuk menganalisis kinerja model.

- Equal Width Binning(Uniform Binning):

Model	Number of Bins	F1 Score	ROC AUC
AdaBoost	3	0.5	0.851852
Extra Trees	3	0.5	0.814815
Naive Bayes	3	0.266667	0.768519
Random Forest	3	0.5	0.833333
XGBoost	3	0.5	0.796296

Interpretasi:

Setelah proses Equal Width Binning dilakukan, kami berhasil menentukan jumlah Bin terbaik dari setiap model. Namun, Bin ini masih berlaku untuk semua variabel. Sebagai contoh, jika jumlah bin adalah 3, maka akan dibuat 3 pengelompokan untuk setiap variabel prediktor secara keseluruhan. Perlu dicatat juga bahwa setiap model menunjukkan nilai ROC AUC yang lebih dari 0.75, tanpa adanya perbedaan yang signifikan antara satu model dengan yang lainnya.

- Equal Frequency Binning(Quantile Binning):

Model	Number of Bins	F1 Score	ROC AUC
AdaBoost	3	0.555556	0.824074
Extra Trees	3	0.285714	0.664352
Naive Bayes	3	0.6	0.759259
Random Forest	3	0.375	0.763889
XGBoost	3	0.470588	0.699074

Interpretasi:

Setelah melalui proses Equal Frequency Binning, kami berhasil menentukan jumlah Bin terbaik dari setiap model. Namun, Bin ini masih berlaku untuk semua variabel. Sebagai contoh, jika jumlah bin adalah 3, maka akan dibuat 3 pengelompokan untuk setiap variabel prediktor secara keseluruhan. Penting untuk dicatat bahwa terdapat model yang menunjukkan nilai ROC AUC di bawah 0.7, sementara ada juga yang mencapai nilai di atas 0.7. Hal ini menandakan adanya perbedaan yang signifikan antara model-model tersebut, yang mengindikasikan bahwa metode Equal Frequency Binning mungkin kurang tepat dalam mengkategorikan data.

- K-Means Binning:

Model	Number of Clusters	F1 Score	ROC AUC
AdaBoost	3	0.444444	0.736111
Extra Trees	3	0.315789	0.611111
Naive Bayes	3	0.315789	0.587963
Random Forest	3	0.4	0.652778
XGBoost	3	0.421053	0.680556

Interpretasi:

Setelah melalui proses K-Means Binning, kami berhasil menentukan jumlah Cluster terbaik dari setiap model. Namun, Cluster ini masih berlaku untuk semua variabel. Sebagai contoh, jika jumlah Cluster adalah 3, maka akan dibuat 3 pengelompokan untuk setiap variabel prediktor secara keseluruhan. Penting untuk dicatat bahwa hanya satu model yang menunjukkan nilai ROC AUC di atas 0.7, sedangkan 4 model lainnya mencapai nilai di bawah 0.7. Perbedaan signifikan ini menandakan bahwa terdapat variasi yang cukup besar antara model-model tersebut, dan nilai ROC AUC yang rendah mengindikasikan bahwa metode K-Means Binning mungkin tidak optimal dalam mengkategorikan data.

- Decision Tree Binning:

Model	Max Leaf Nodes	F1 Score	ROC AUC
AdaBoost	3	0.315789	0.699074
Extra Trees	4	0.315789	0.694444
Naive Bayes	5	0.133333	0.752315
Random Forest	4	0.4	0.703704
XGBoost	4	0.47619	0.666667

Interpretasi:

Setelah proses Decision Tree Binning dilakukan, kami berhasil menentukan jumlah daun terbaik dari setiap model. Namun, jumlah daun ini masih berlaku untuk semua variabel. Sebagai contoh, jika jumlah daun adalah 3, maka akan dibuat 3 pengelompokan untuk setiap variabel prediktor secara keseluruhan. Perlu dicatat juga bahwa setiap model menunjukkan nilai ROC AUC yang signifikan di rentang 0.66-0.70. Namun, karena nilai ROC AUC dari kelima model tersebut relatif rendah, metode ini berkemungkinan kurang optimal dalam mengkategorikan data.
Kesimpulan :

- Equal Width Binning:
- Setiap model menunjukkan nilai ROC AUC lebih dari 0.75.
- Tidak ada perbedaan signifikan antara model-model.
- Metode ini menunjukkan kinerja yang cukup baik secara konsisten.

- Equal Frequency Binning:

- Terdapat variasi yang signifikan antara model-model, dengan beberapa model memiliki nilai ROC AUC di bawah 0.7 dan beberapa di atas 0.7.
- Metode ini menunjukkan bahwa tidak semua model berkinerja baik, sehingga kurang konsisten.

- K-Means Binning:
- Hanya satu model yang memiliki nilai ROC AUC di atas 0.7, sementara empat lainnya di bawah 0.7.
- Perbedaan signifikan ini menandakan variasi besar dan nilai ROC AUC yang rendah, mengindikasikan metode ini mungkin tidak optimal.

- Decision Tree Binning:
- Nilai ROC AUC untuk setiap model berada di rentang 0.66-0.70.
- Nilai ini relatif rendah, menunjukkan bahwa metode ini mungkin kurang optimal.

Berdasarkan hasil di atas, **Equal Width Binning (Uniform Binning)** adalah metode terbaik dari keempat analisis tersebut. Metode ini menunjukkan nilai ROC AUC yang konsisten di atas 0.75 untuk semua model, tanpa perbedaan yang signifikan, yang mengindikasikan kinerja yang baik dan stabil dalam mengkategorikan data.

8. Memilih model terbaik sebagai referensi untuk mengevaluasi Cluster/Bin/Leaf yang optimal.

Equal Width Binning(Uniform Binning):

Model	Number of Bins	F1 Score	ROC AUC
AdaBoost	3	0.5	0.851852
Extra Trees	3	0.5	0.814815
Naive Bayes	3	0.266667	0.768519
Random Forest	3	0.5	0.833333
XGBoost	3	0.5	0.796296

Interpretasi:

1. AdaBoost memiliki ROC AUC tertinggi (0.851852) dan F1 Score yang cukup baik (0.5).

2. Random Forest juga menunjukkan ROC AUC yang tinggi (0.833333) dengan F1 Score yang sama (0.5).

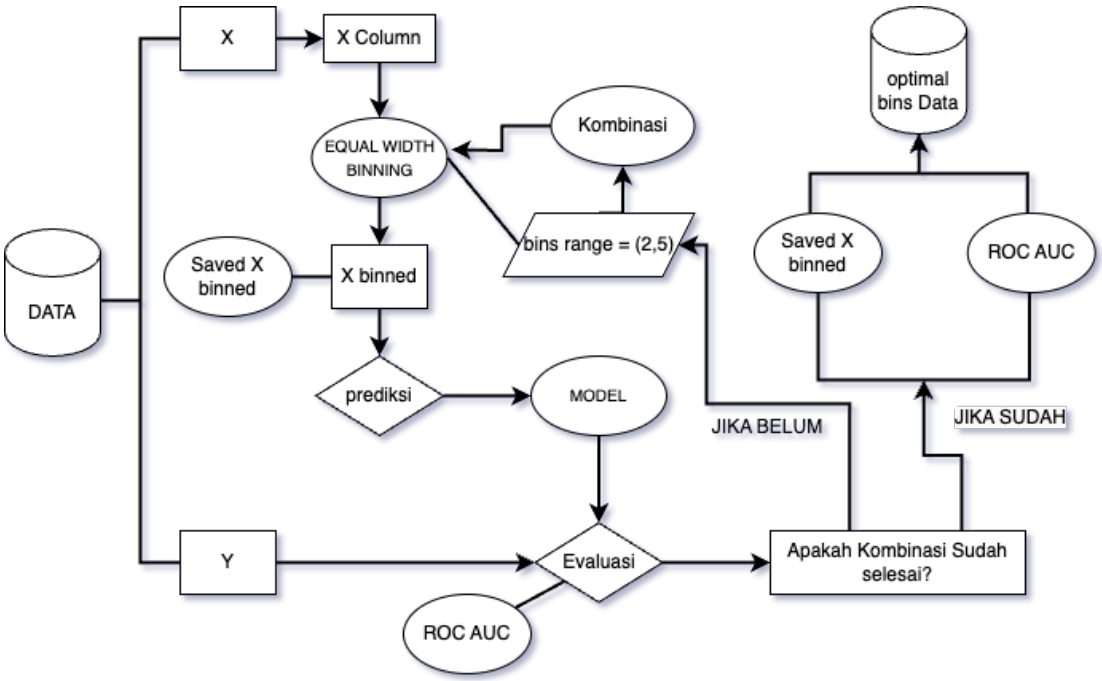
3. Extra Trees memiliki ROC AUC (0.814815) dan F1 Score (0.5), tetapi sedikit lebih rendah dibandingkan AdaBoost dan Random Forest pada nilai ROC AUC .

4. XGBoost menunjukkan ROC AUC (0.796296) dan F1 Score (0.5), yang sedikit lebih rendah dibandingkan dengan AdaBoost dan Random Forest pada nilai ROC AUC.

5. Naive Bayes memiliki nilai yang paling rendah dalam hal ROC AUC (0.768519) dan F1 Score (0.266667).

Kesimpulan:
Karena fokus metrik dalam penelitian ini adalah ROC AUC, maka AdaBoost akan menjadi pilihan terbaik sebagai referensi untuk mengevaluasi Cluster/Bin/Leaf yang optimal. Ini dikarenakan AdaBoost memiliki nilai ROC AUC tertinggi di antara semua model yang diuji.

9. Memilih Bins/Cluster/Leaf optimal untuk semua variabel.



Syntax Python:

```
model = AdaBoostClassifier(random_state=42)
def equal_width_binning(X_train, X_test, n_bins):
  discretizer = KBinsDiscretizer(n_bins=n_bins, encode='ordinal', strategy='uniform' , random_state=42)
  binned_X_train = discretizer.fit_transform(X_train)
  binned_X_test = discretizer.transform(X_test)
  return binned_X_train, binned_X_test

n_bins_range = range(2, 6)
columns = X.columns
all_combinations = list(product(n_bins_range, repeat=len(columns)))

results = []

for combination in all_combinations:
  X_train_binned = X_train.copy()
  X_test_binned = X_test.copy()
  for i, n_bins in enumerate(combination):
    col = columns[i]
          X_train_binned[[col]],X_test_binned[[col]]= equal_width_binning(X_train[[col]],  X_test[[col]], n_bins)
  model.fit(X_train_binned, y_train)
  y_pred = model.predict_proba(X_test_binned)[:, 1]
  roc_auc = roc_auc_score(y_test, y_pred)
```

```
    result = list(combination) + [roc_auc]
    results.append(result)

results_df = pd.DataFrame(results, columns=[f'n_bins_{col}' for col in columns] + ['roc_auc'])

hasil = results_df
```

Sehingga diperoleh hasil Optimal Bins Data :

	n_bins_age	n_bins_creatinine_phosphokinase	n_bins_ejection_fraction	n_bins_platelets	n_bins_serum_creatinine	n_bins_serum_sodium	roc_auc
0	2	2	2	2	2	2	0.706019
1	2	2	2	2	2	3	0.787037
2	2	2	2	2	2	4	0.726852
3	2	2	2	2	2	5	0.803241
4	2	2	2	2	3	2	0.706019
...	...	...	...	...	...	...	...
4091	5	5	5	5	4	5	0.775463
4092	5	5	5	5	5	2	0.680556
4093	5	5	5	5	5	3	0.701389
4094	5	5	5	5	5	4	0.678241
4095	5	5	5	5	5	5	0.733796

10. Menarik kesimpulan dan menentukan kategori binning/klaster untuk setiap variabel.

Selanjutnya akan dicari score roc auc yang maximal setelah kombinasi binning tiap variabel.

```
hasil['roc_auc'].max()
```

Diperoleh:

0.9074074074074074

Maka akan dicari observasi yang memiliki nilai ROC AUC score tersebut.

```
hasil[hasil['roc_auc'] == 0.9074074074074074]
```

Diperoleh hasil:

	n_bins_age	n_bins_creatinine_phosphokinase	n_bins_ejection_fraction	n_bins_platelets	n_bins_serum_creatinine	n_bins_serum_sodium	roc_auc
327	2	3	3	2	3	5	0.907407

Kesimpulan:

1. Optimal Binning: Nilai ROC AUC tertinggi diperoleh saat menggunakan kombinasi pada Uniform Binning dengan jumlah bin tertentu untuk setiap variabel. Secara spesifik, jumlah bin optimal untuk masing-masing variabel adalah sebagai berikut:

- Usia (Age): 2 bin
- Kreatinin Fosfokinase (Creatinine Phosphokinase): 3 bin
- Fraksi Ejeksi (Ejection Fraction): 3 bin
- Trombosit (Platelets): 2 bin
- Kreatinin Serum (Serum Creatinine): 3 bin
- Sodium Serum (Serum Sodium): 5 bin

2. ROC AUC Score: ROC AUC score maksimal yang ditemukan setelah kombinasi binning menggunakan Uniform Binning adalah 0.9074074074074074.

Oleh karena itu, dengan menggunakan Uniform Binning dengan jumlah bin yang optimal untuk setiap variabel dapat meningkatkan kinerja model dalam memprediksi, ini ditunjukkan dengan peningkatan ROC AUC score menjadi 0.9074074074074074.

11. Mengekspor/Mengunduh Data Final setelah Binning.

	age	anaemia	creatinine_pho	diabetes	ejection_fra	high_blood	platelets	serum_creat	serum_sodi	sex	smoking	time	DEATH_EV

			sphokinase		ction	_pressure		inine	um				ENT
0	1	0	0	0	0	1	0	0	2	1	0	4	1
1	0	0	2	0	1	0	0	0	3	1	0	6	1
2	0	0	0	0	0	0	0	0	2	1	1	7	1
3	0	1	0	0	0	0	0	0	3	1	0	7	1
4	0	1	0	1	0	0	0	0	0	0	0	8	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
294	0	0	0	1	1	1	0	0	4	1	1	270	0
295	0	0	0	0	1	0	0	0	3	0	0	271	0
296	0	0	0	1	2	0	1	0	3	0	0	278	0
297	0	0	0	0	1	0	0	0	3	1	1	280	0
298	0	0	0	0	1	0	0	0	3	1	1	285	0

Penjelasan Binning

-Age (2 bin):

Setelah melakukan binning pada variabel usia (age), didapatkan hasil sebagai berikut:

Terdapat dua bin yang dihasilkan: bin 0 dan bin 1.

- Bin 0 memiliki nilai 0.0 dan terdapat 214 observasi.
- Bin 1 memiliki nilai 1.0 dan terdapat 85 observasi.

Dengan demikian, proporsi pembagian setelah binning adalah sebagai berikut:

- Proporsi observasi pada bin 0 (usia < 68)
- Proporsi observasi pada bin 1 (usia >= 68)

-Creatinine (3bin):

Setelah melakukan binning pada variabel kreatinin fosfokinase (creatinine phosphokinase), didapatkan hasil sebagai berikut:

Terdapat tiga bin yang dihasilkan: bin 0, bin 1, dan bin 2:

- Bin 0 memiliki nilai 0.0 dan terdapat 289 observasi.
- Bin 1 memiliki nilai 1.0 dan terdapat 7 observasi.
- Bin 2 memiliki nilai 2.0 dan terdapat 3 observasi.

Berikut adalah rentang nilai untuk masing-masing bin:

- Bin 0: Rentang nilai kreatinin fosfokinase adalah dari 0 hingga 2522.0.

- Bin 1: Rentang nilai kreatinin fosfokinase adalah dari 2600 hingga 5209.0.
- Bin 2: Rentang nilai kreatinin fosfokinase adalah dari 5800 hingga 7861.0.

-Ejection Fraction (3bin):

Setelah melakukan binning pada variabel fraksi ejeksi (ejection fraction), didapatkan hasil sebagai berikut:

Terdapat tiga bin yang dihasilkan: bin 0, bin 1, dan bin 2:

- Bin 0 memiliki nilai 0.0 dan terdapat 142 observasi.
- Bin 1 memiliki nilai 1.0 dan terdapat 121 observasi.
- Bin 2 memiliki nilai 2.0 dan terdapat 36 observasi.

Berikut adalah rentang nilai untuk masing-masing bin:

- Bin 0: Rentang nilai fraksi ejeksi adalah dari 0 hingga 35
- Bin 1: Rentang nilai fraksi ejeksi adalah dari 36 hingga 55
- Bin 2: Rentang nilai fraksi ejeksi adalah dari 56 hingga 80

-Platelets (2bin):

Setelah melakukan binning pada variabel trombosit (platelets), didapatkan hasil sebagai berikut:

Terdapat dua bin yang dihasilkan: bin 0 dan bin 1:

- Bin 0 memiliki nilai 0.0 dan terdapat 285 observasi.
- Bin 1 memiliki nilai 1.0 dan terdapat 14 observasi.

Berikut adalah rentang nilai untuk masing-masing bin:

- Bin 0: Rentang nilai trombosit adalah dari 0 hingga 427,000.0.
- Bin 1: Rentang nilai trombosit adalah dari 448,000 hingga 850,000.0.

-Serum Creatinine (3bin):

Setelah melakukan binning pada variabel kreatinin serum (serum creatinine), didapatkan hasil sebagai berikut:

Terdapat tiga bin yang dihasilkan: bin 0, bin 1, dan bin 2:

- Bin 0 memiliki nilai 0.0 dan terdapat 287 observasi.
- Bin 1 memiliki nilai 1.0 dan terdapat 9 observasi.
- Bin 2 memiliki nilai 2.0 dan terdapat 3 observasi.

Berikut adalah rentang nilai untuk masing-masing bin:

- Bin 0: Rentang nilai kreatinin serum adalah dari 0.0 hingga 3.4.
- Bin 1: Rentang nilai kreatinin serum adalah dari 3.5 hingga 6.1.

- Bin 2: Rentang nilai kreatinin serum adalah dari 6.5 hingga 9.4.

-Serum Sodium (5bin):

Setelah melakukan binning pada variabel sodium serum (serum sodium), didapatkan hasil sebagai berikut:

Terdapat lima bin yang dihasilkan: bin 0, bin 1, bin 2, bin 3, dan bin 4.

- Bin 0 memiliki nilai 0.0 dan terdapat 2 observasi.
- Bin 1 memiliki nilai 1.0 dan terdapat 4 observasi.
- Bin 2 memiliki nilai 2.0 dan terdapat 45 observasi.
- Bin 3 memiliki nilai 3.0 dan terdapat 206 observasi.
- Bin 4 memiliki nilai 4.0 dan terdapat 42 observasi.

Berikut adalah rentang nilai untuk masing-masing bin:

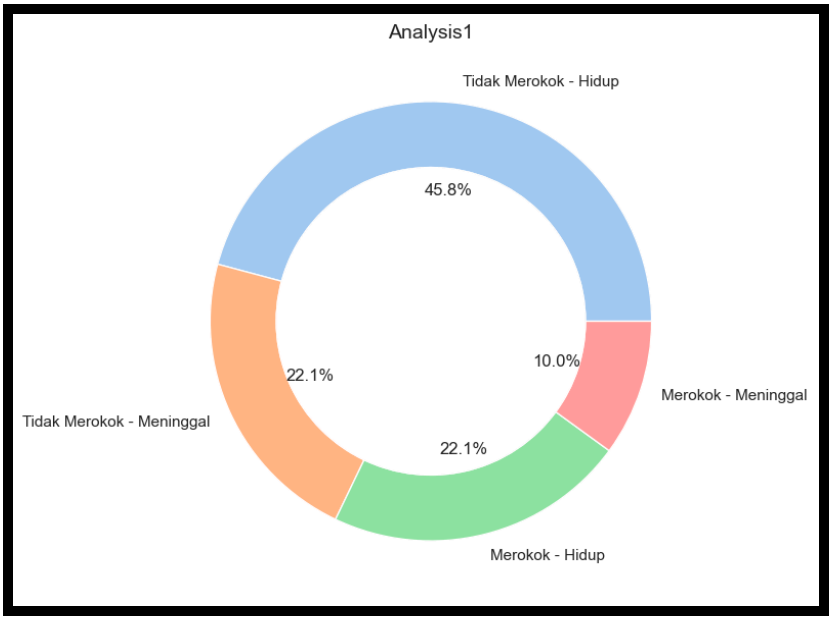
- Bin 0: Rentang nilai sodium serum adalah dari 0 hingga 116
- Bin 1: Rentang nilai sodium serum adalah dari 117 hingga 126
- Bin 2: Rentang nilai sodium serum adalah dari 127 hingga 133
- Bin 3: Rentang nilai sodium serum adalah dari 134 hingga 140
- Bin 4: Rentang nilai sodium serum adalah dari 141 hingga 148

B.) Analisis Survival:

Exploratory Data Analysis (EDA):

- Smoking:

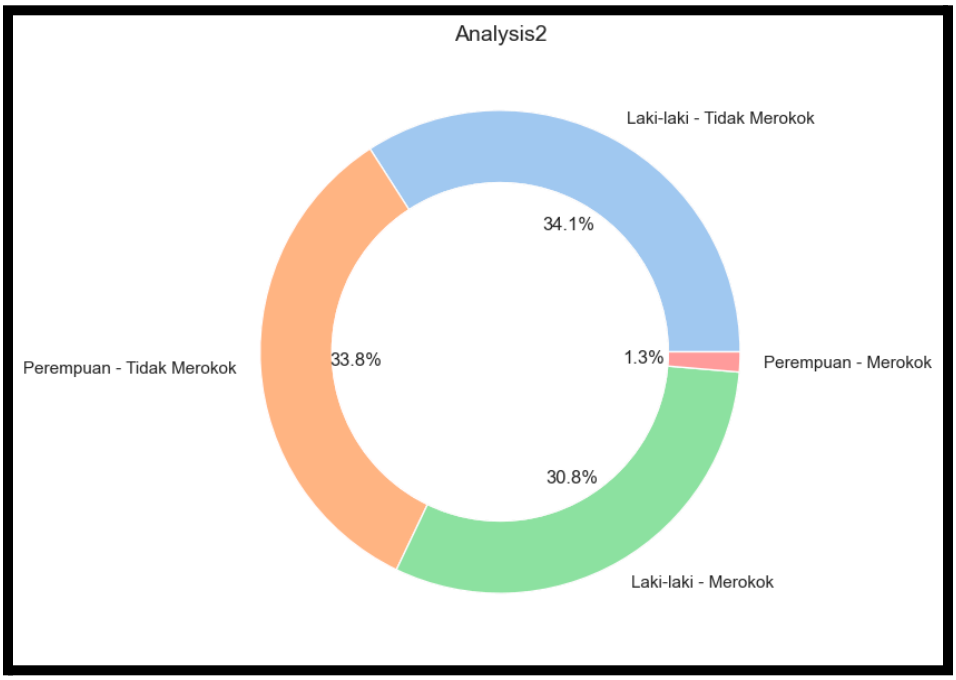
Visualisasi Analisis Hubungan Aktivitas Merokok dan Kematian



Interpretasi:

1. Berdasarkan visualisasi di atas, terlihat bahwa proporsi pasien yang tidak merokok dan masih hidup adalah yang tertinggi, yaitu 45,8%.
2. Proporsi pasien yang tidak merokok dan meninggal sama dengan pasien yang merokok dan masih hidup, masing-masing sebesar 22,1%.
3. Proporsi pasien yang merokok dan meninggal adalah yang paling kecil, hanya 10%.

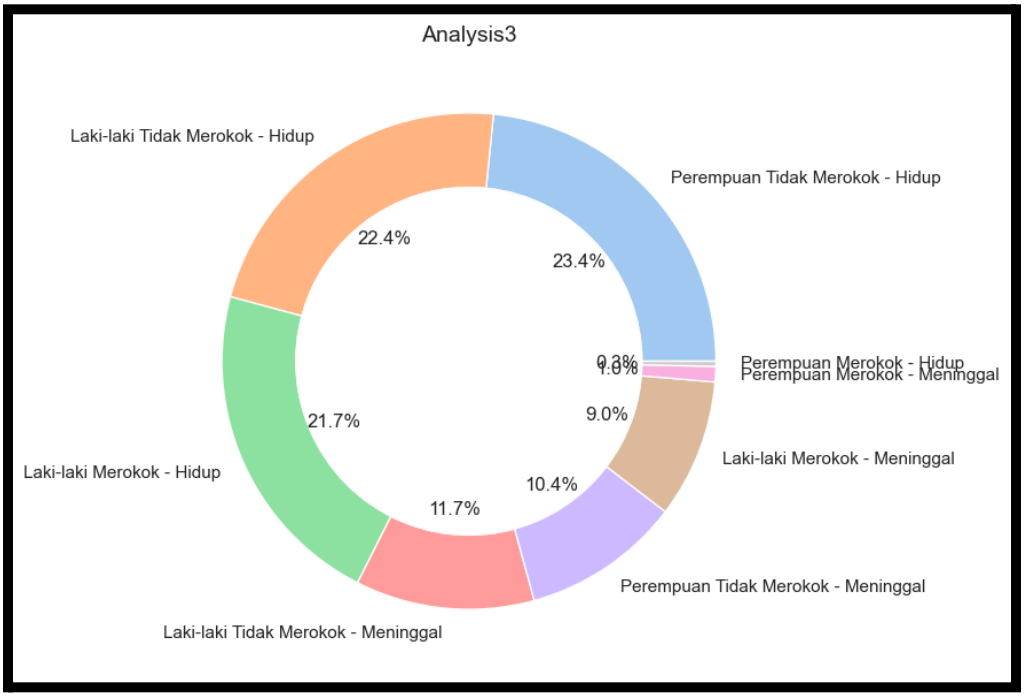
Visualisasi Analisis Hubungan Jenis Kelamin dan Aktivitas Merokok



Interpretasi:

1. Berdasarkan visualisasi data tersebut, 34,1% pasien adalah laki-laki yang tidak merokok.
2. Berdasarkan visualisasi data tersebut, 33,8% pasien adalah perempuan yang tidak merokok.
3. Laki-laki yang merokok juga memiliki proporsi yang cukup besar, yaitu 30,8%.
4. Sedangkan perempuan yang merokok sangat sedikit, hanya 1,3% dari pasien.

Visualisasi Analisis Hubungan Jenis Kelamin dan Aktivitas Merokok dengan Kematian



Interpretasi:

1. Proporsi jenis kelamin yang tidak merokok dan hidup cukup besar dengan proporsi laki - laki tidak merokok hidup (22.4%) dan perempuan tidak merokok hidup (23.4%)
2. Proporsi jenis kelamin yang merokok dan hidup yaitu Laki - laki merokok hidup (21.7%) dan perempuan merokok hidup (0.3%)
3. Proporsi jenis kelamin yang tidak merokok dan meninggal yaitu laki - laki tidak merokok meninggal (11.7%) dan perempuan tidak merokok meninggal (10.4%)
4. Proporsi jenis kelamin yang merokok dan meninggal yaitu laki - laki merokok meninggal (9%) dan perempuan (1%)

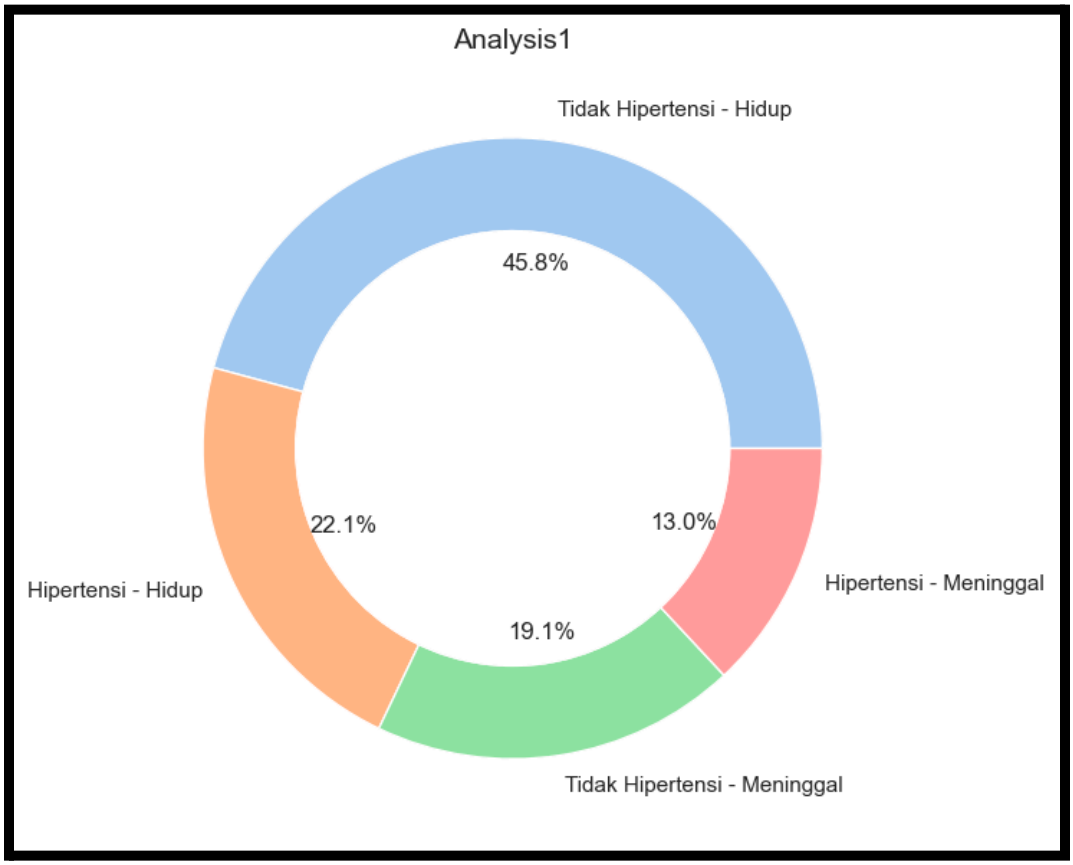
Kesimpulan

Dari data tersebut, dapat disimpulkan bahwa proporsi perempuan yang merokok sangat kecil, baik dalam hal kehidupan maupun kematian. Hal ini terlihat dari proporsi perempuan yang merokok hidup (0.3%) dan yang meninggal akibat merokok (1%). Sementara itu, proporsi laki-laki yang merokok lebih besar, baik dalam hal kehidupan (21.7%) maupun kematian (9%).

Secara umum, perempuan cenderung memiliki proporsi yang lebih kecil dalam hal merokok dan konsekuensi kesehatannya, baik hidup maupun meninggal, dibandingkan dengan laki-laki. Meskipun demikian, penting untuk diingat bahwa setiap individu, tanpa memandang jenis kelaminnya, memiliki risiko kesehatan yang perlu diperhatikan terkait kebiasaan merokok.

- Blood Pressure:

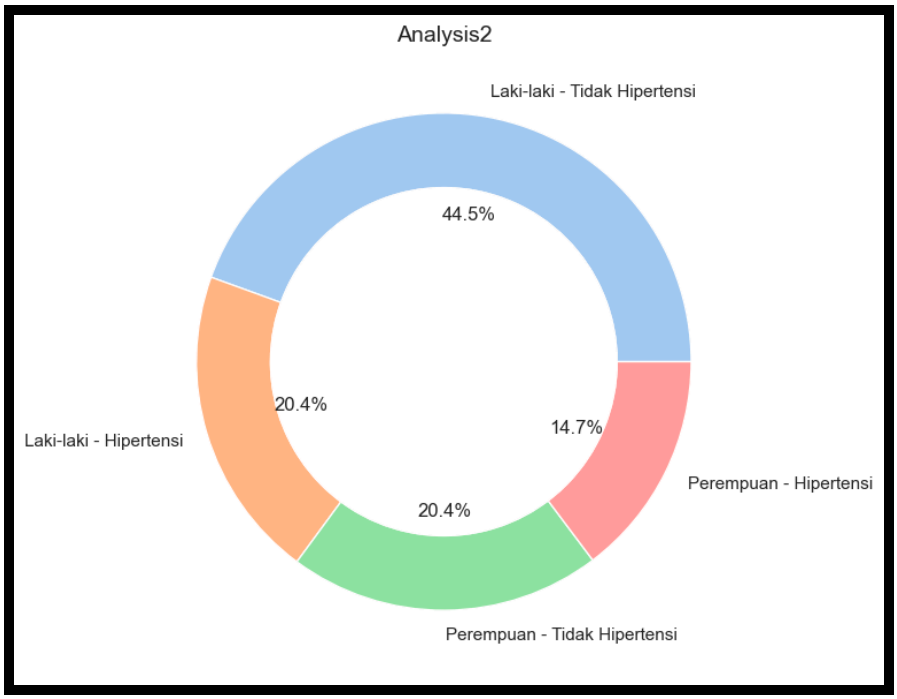
Visualisasi Analisis Hubungan Blood Pressure dan Kematian



Interpretasi:

1. Berdasarkan visualisasi di atas, terlihat bahwa proporsi pasien yang hipertensi dan masih hidup adalah yang tertinggi, yaitu 45,8%.
2. Proporsi pasien yang tidak hipertensi dan meninggal tidak jauh berbeda dengan pasien yang hipertensi dan masih hidup, masing-masing sebesar hipertensi hidup (22,1%) , tidak hipertensi meninggal (19,1%)
3. Proporsi pasien yang hipertensi dan meninggal adalah yang paling kecil, hanya (13%)

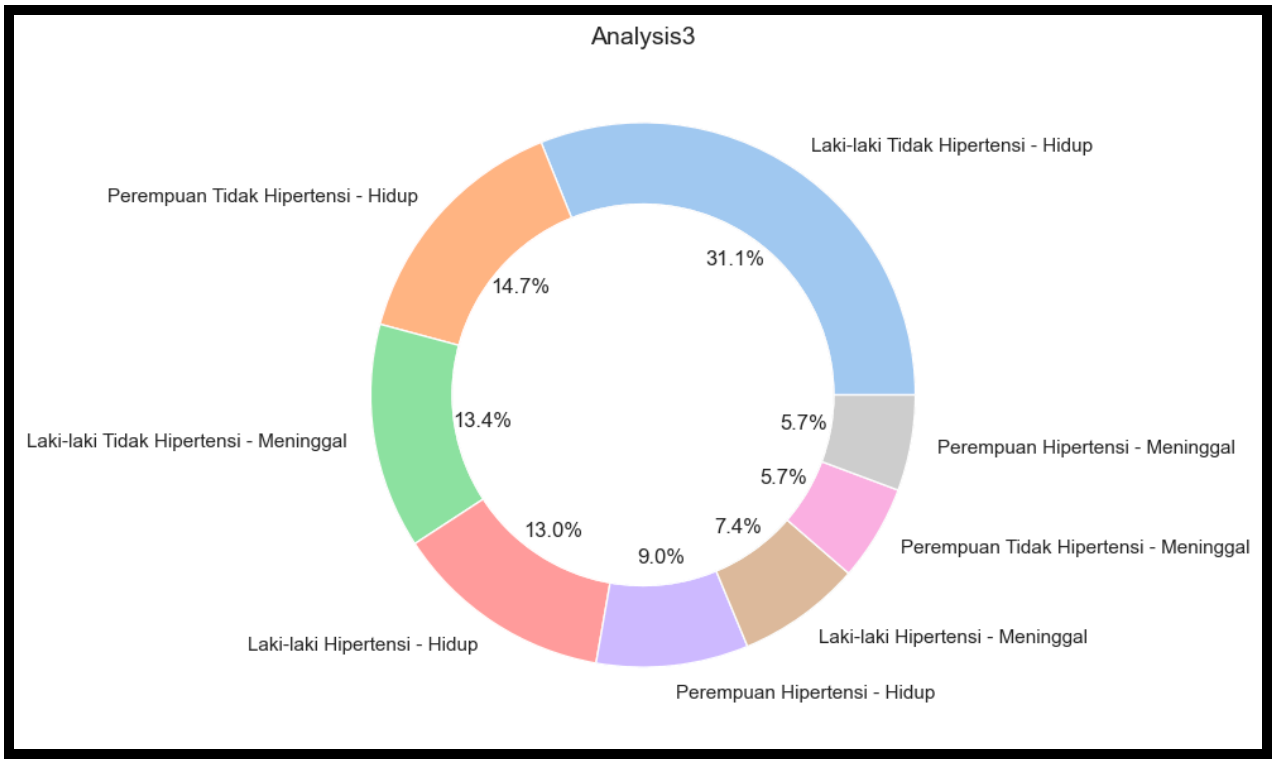
Visualisasi Analisis Hubungan Jenis Kelamin dan Blood Pressure



Interpretasi:

1. Berdasarkan visualisasi data tersebut, 44,5% pasien adalah laki-laki yang tidak mengalami hipertensi.
2. Berdasarkan visualisasi data tersebut, pasien perempuan yang tidak mengalami hipertensi memiliki proporsi yang sama dengan pasien laki-laki yang mengalami hipertensi yaitu 20,4%.
3. Sedangkan perempuan yang mengalami hipertensi sangat sedikit, hanya 14.7 % dari seluruh pasien.

Visualisasi Analisis Hubungan Jenis Kelamin dan Blood Pressure dengan Kematian



Interpretasi:

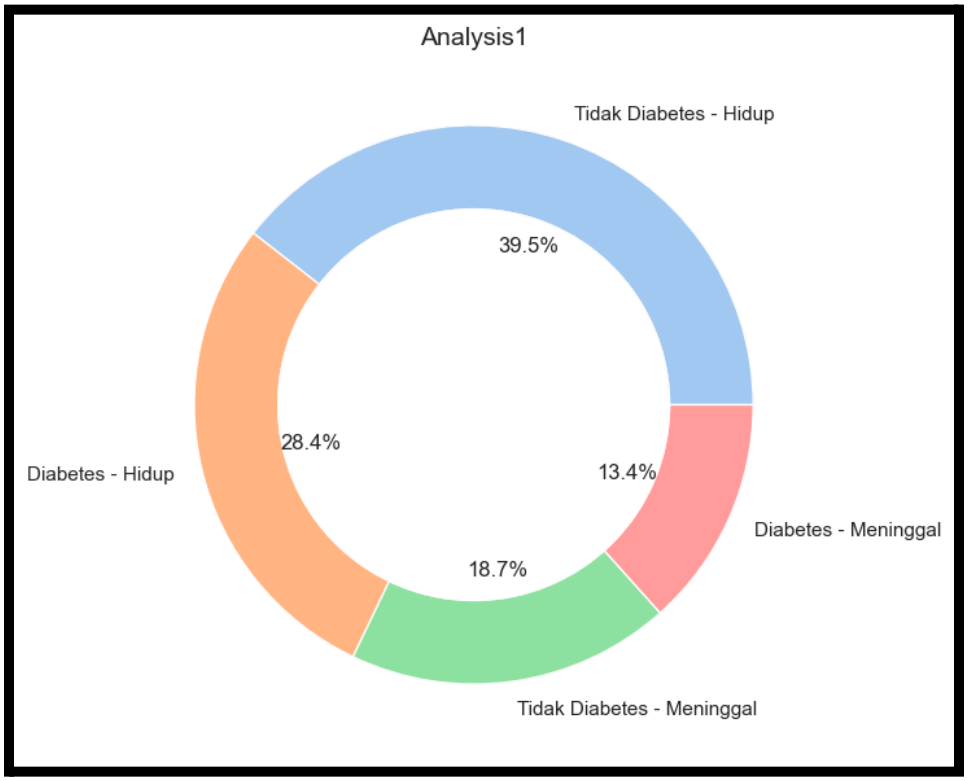
1. Proporsi laki-laki yang tidak menderita hipertensi dan tetap hidup (31.1%) jauh lebih tinggi daripada proporsi perempuan yang tidak menderita hipertensi dan tetap hidup (14.7%).
2. Proporsi laki-laki yang menderita hipertensi dan masih hidup (13%) lebih tinggi daripada proporsi perempuan yang menderita hipertensi dan masih hidup (9%).
3. Proporsi laki-laki yang tidak menderita hipertensi dan meninggal (13.4%) lebih tinggi daripada proporsi perempuan yang tidak menderita hipertensi dan meninggal (5.7%).
4. Proporsi laki-laki yang menderita hipertensi dan meninggal (7.4%) lebih tinggi daripada proporsi perempuan yang menderita hipertensi dan meninggal (5.7%).

Kesimpulan

Berdasarkan data tersebut, terlihat bahwa proporsi laki-laki yang mengalami hipertensi, baik yang masih hidup maupun yang sudah meninggal, cenderung lebih tinggi daripada perempuan yang mengalami kondisi yang sama. Meskipun demikian, proporsi laki-laki yang tidak menderita hipertensi namun meninggal juga relatif tinggi jika dibandingkan dengan perempuan yang tidak menderita hipertensi dan meninggal.

- Diabetes:

Visualisasi Analisis Hubungan Diabetes dan Kematian

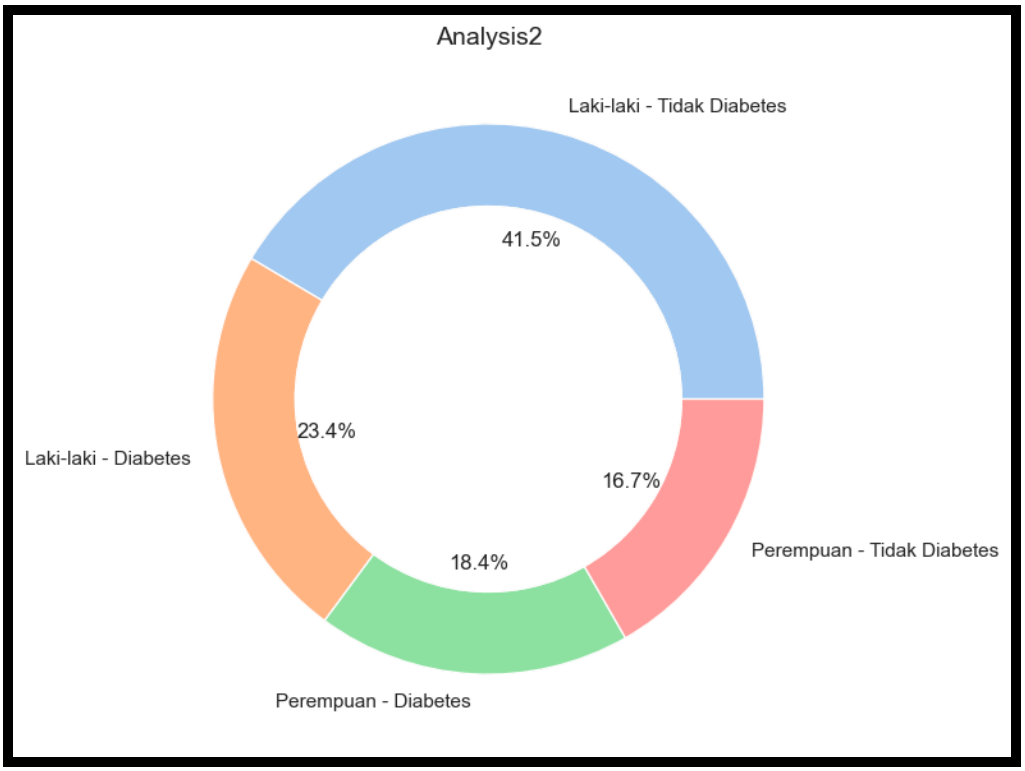


Interpretasi:

1. Berdasarkan visualisasi di atas, terlihat bahwa proporsi pasien yang tidak diabetes dan masih hidup adalah yang tertinggi, yaitu 39,5%.
2. Proporsi pasien yang diabetes dan masih hidup berada di peringkat kedua tertinggi yaitu 28,4 %
3. Proporsi pasien yang tidak diabetes dan meninggal adalah 18,7%
4. Proporsi pasien yang merokok dan meninggal adalah yang paling kecil, hanya 13,4%.

Namun jika dilihat persebaran proporsinya lumayan merata (tidak terlalu dominan)

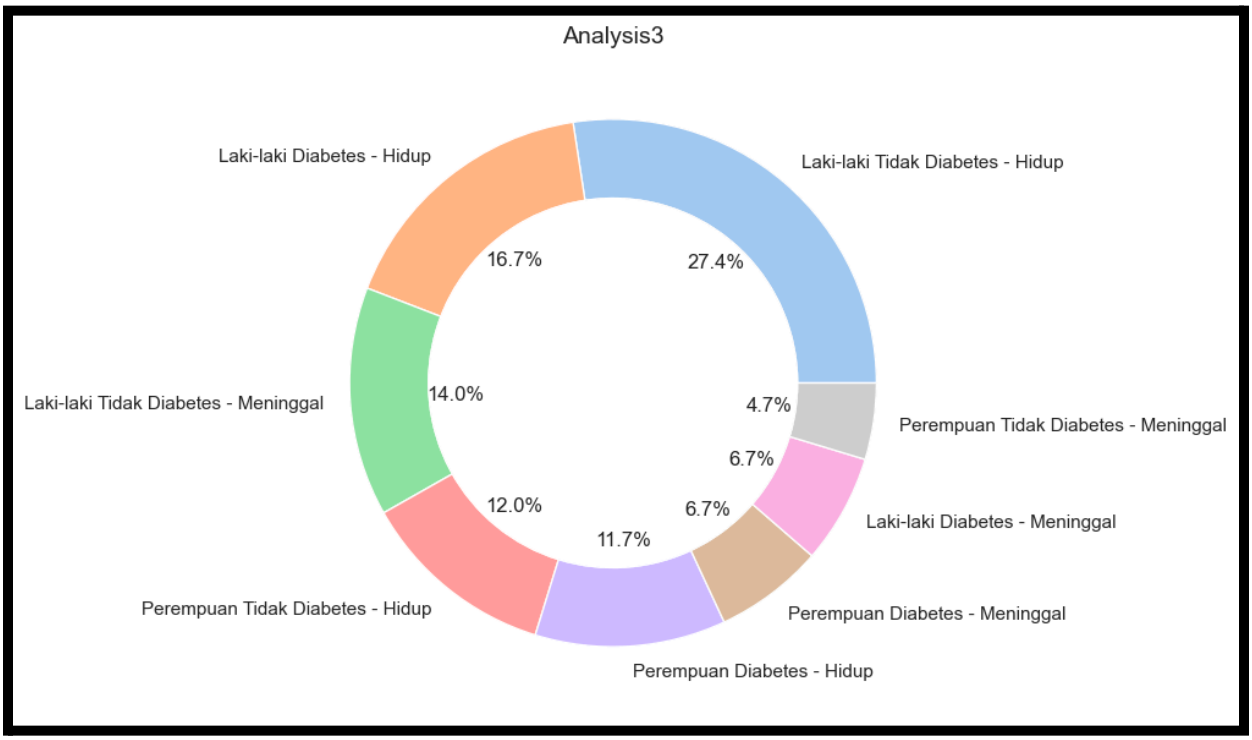
Visualisasi Analisis Hubungan Jenis Kelamin dan Diabetes



Interpretasi:

1. Sebanyak 41,5% dari total pasien merupakan laki-laki yang tidak menderita diabetes.
2. Proporsi laki-laki yang mengalami diabetes adalah sebesar 23,4%, berdasarkan visualisasi data tersebut.
3. Perempuan yang mengalami diabetes mencapai 18,4% dari total pasien, menurut data visualisasi yang disajikan tersebut.
4. Meskipun demikian, jumlah perempuan yang mengalami diabetes relatif sedikit, hanya sekitar 16,7% dari keseluruhan pasien, seperti yang terlihat dalam visualisasi data tersebut.

Visualisasi Analisis Hubungan Jenis Kelamin dan Diabetes dengan Kematian



Interpretasi:

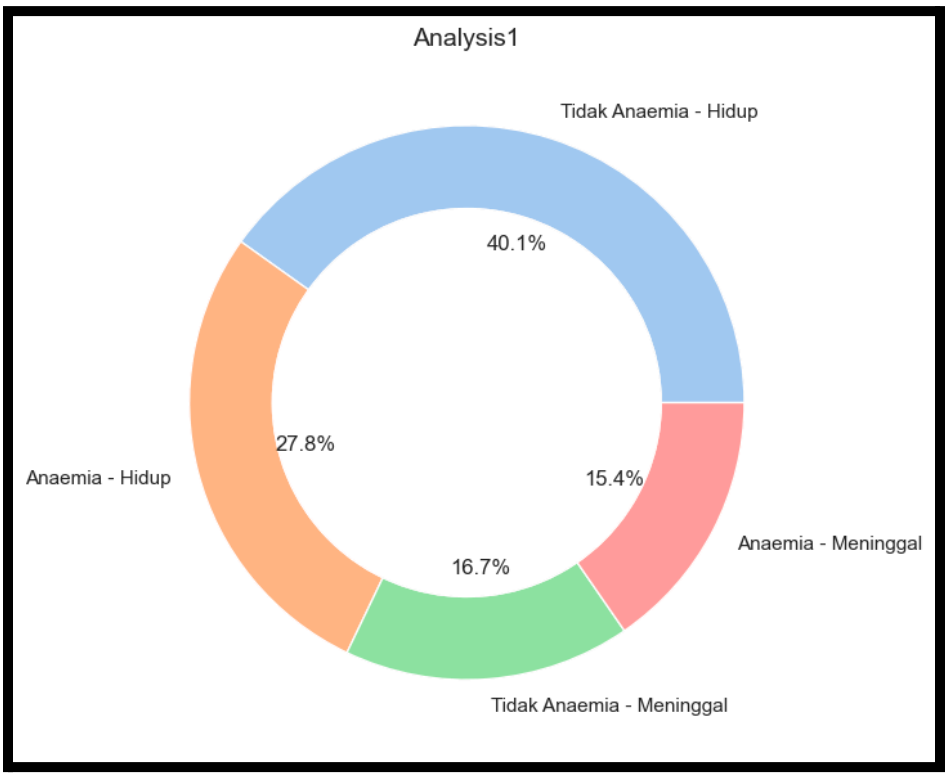
1. Proporsi laki-laki yang tidak menderita diabetes dan tetap hidup (27.4%) jauh lebih tinggi daripada proporsi perempuan yang tidak menderita diabetes dan tetap hidup (12%).
2. Proporsi laki-laki yang menderita diabetes dan masih hidup (16.7%) lebih tinggi daripada proporsi perempuan yang menderita diabetes dan masih hidup (11.7%).
3. Proporsi laki-laki yang tidak menderita diabetes dan meninggal (14%) lebih tinggi daripada proporsi perempuan yang tidak menderita diabetes dan meninggal (4.7%).
4. Proporsi laki-laki yang menderita diabetes dan meninggal (6.7%) sama dengan proporsi perempuan yang menderita diabetes dan meninggal (6.7%).

Kesimpulan

Berdasarkan data tersebut, dapat dilihat bahwa proporsi laki-laki yang mengalami diabetes, baik yang masih hidup maupun yang sudah meninggal, cenderung lebih tinggi daripada proporsi perempuan yang mengalami kondisi yang sama. Meskipun demikian, proporsi laki-laki yang tidak menderita diabetes namun meninggal juga relatif tinggi jika dibandingkan dengan perempuan yang tidak menderita diabetes dan meninggal.

- Anaemia:

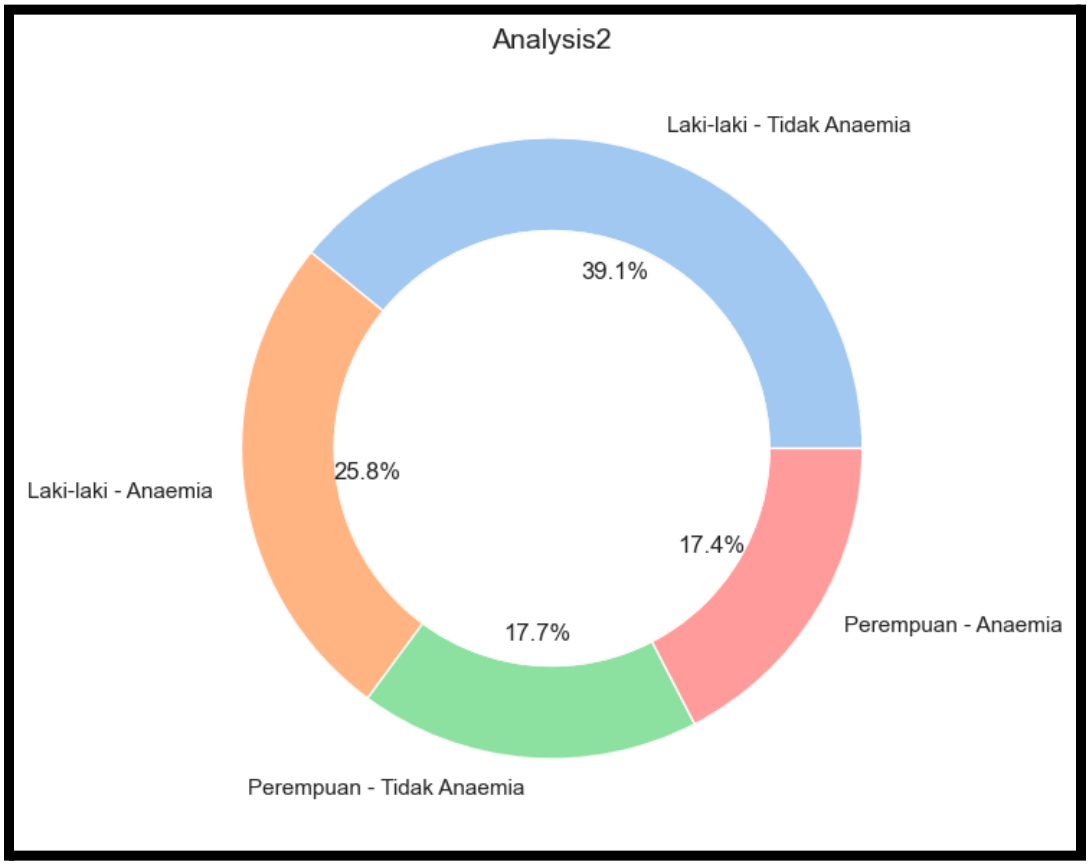
Visualisasi Analisis Hubungan Anaemia dan Kematian



Interpretasi:

1. Berdasarkan visualisasi di atas, terlihat bahwa proporsi pasien yang tidak anaemia dan masih hidup adalah yang tertinggi, yaitu 40,1%.
2. Proporsi pasien yang anaemia dan masih hidup berada di peringkat kedua tertinggi yaitu 27,8 %
3. Proporsi pasien yang tidak anaemia dan meninggal adalah 16,7%
4. Proporsi pasien yang merokok dan meninggal adalah yang paling kecil, hanya 15,4%.

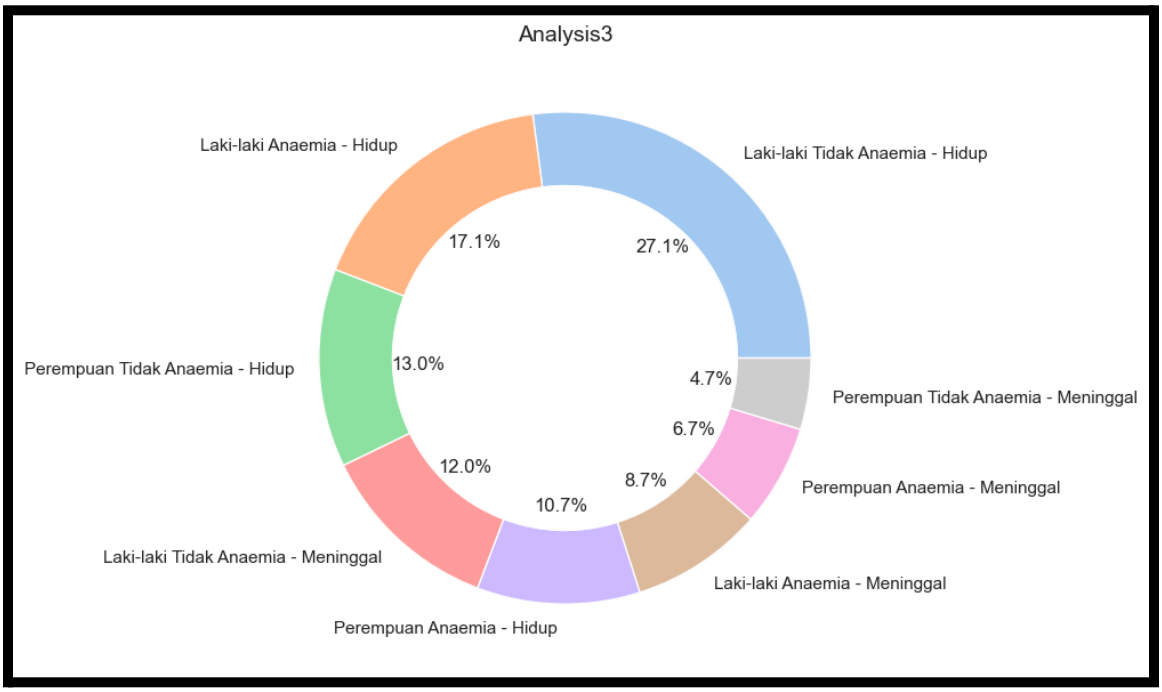
Visualisasi Analisis Hubungan Anaemia dan Jenis Kelamin



Interpretasi:

1. Sebanyak 39,1% dari total pasien merupakan laki-laki yang tidak menderita anaemia.
2. Proporsi laki-laki yang mengalami anaemia adalah sebesar 25,8%, berdasarkan visualisasi data tersebut.
3. Perempuan yang tidak mengalami anaemia mencapai 17,7% dari total pasien, menurut data visualisasi yang disajikan tersebut.
4. Meskipun demikian, jumlah perempuan yang mengalami anaemia relatif sedikit, hanya sekitar 17,4% dari keseluruhan pasien, seperti yang terlihat dalam visualisasi data tersebut.

Visualisasi Analisis Hubungan Jenis Kelamin dan Anaemia dengan Kematian



Interpretasi:

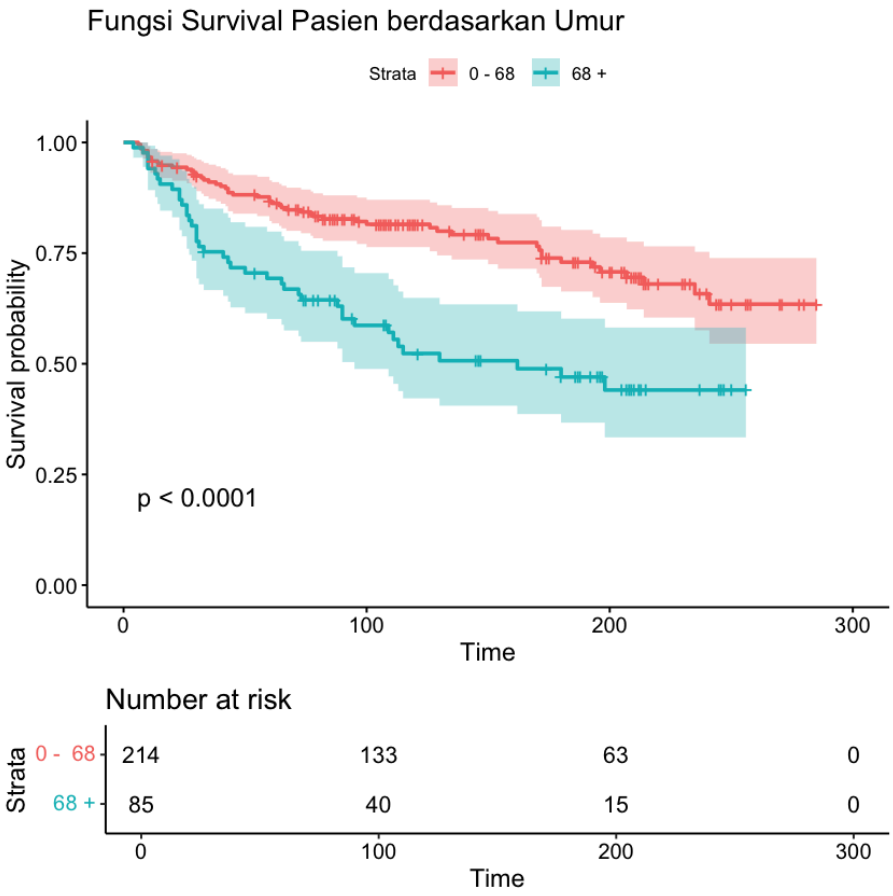
1. Proporsi laki-laki yang tidak menderita anaemia dan tetap hidup (27.1%) jauh lebih tinggi daripada proporsi perempuan yang tidak menderita anaemia dan tetap hidup (13%).
2. Proporsi laki-laki yang menderita anaemia dan masih hidup (17.1%) lebih tinggi daripada proporsi perempuan yang menderita anaemia dan masih hidup (10.7%).
3. Proporsi laki-laki yang tidak menderita anaemia dan meninggal (12%) lebih tinggi daripada proporsi perempuan yang tidak menderita anaemia dan meninggal (4.7%).
4. Proporsi laki-laki yang menderita anaemia dan meninggal (8.7%) lebih tinggi daripada proporsi perempuan yang menderita anaemia dan meninggal (6.7%).

Kesimpulan

Berdasarkan data tersebut, terlihat bahwa proporsi laki-laki yang menderita anaemia, baik yang masih hidup maupun yang sudah meninggal, cenderung lebih tinggi daripada proporsi perempuan yang mengalami kondisi yang sama. Meskipun demikian, proporsi laki-laki yang tidak menderita anaemia namun meninggal juga relatif tinggi jika dibandingkan dengan perempuan yang tidak menderita anaemia dan meninggal.

Kaplan Meier dan Uji - K sampel:

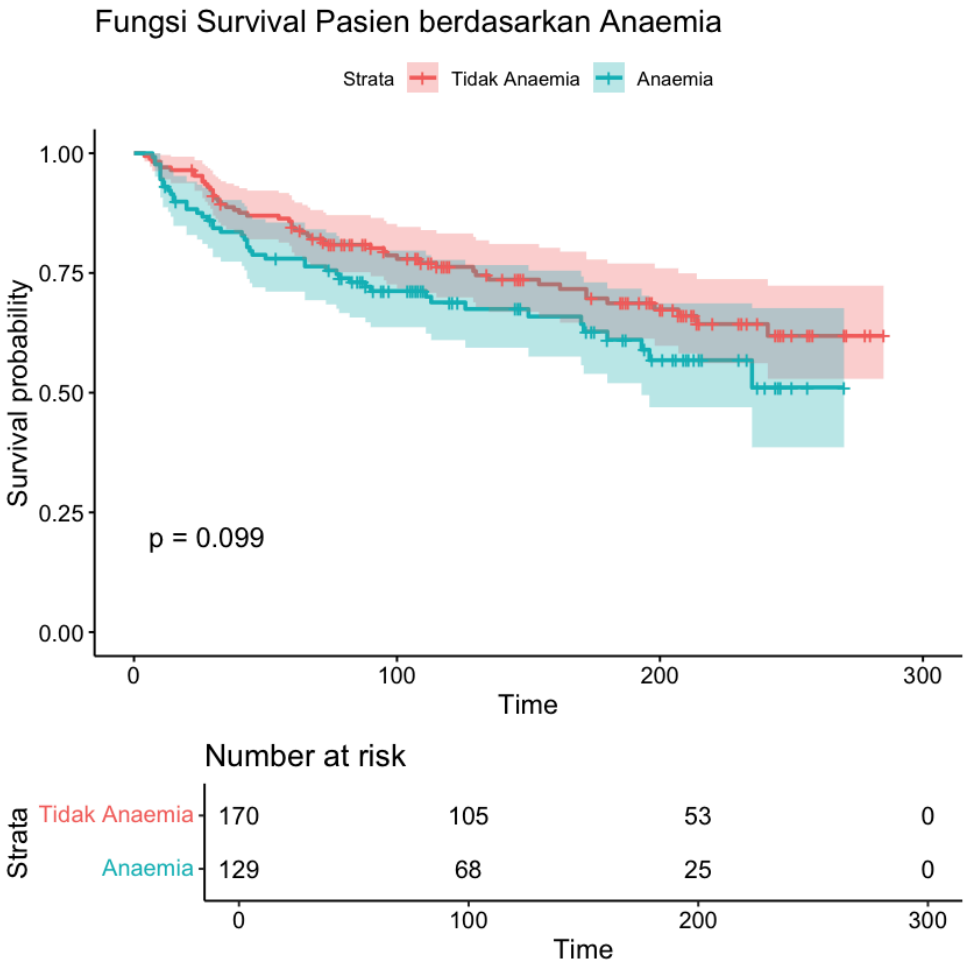
1. Age



Akan dilakukan uji $K = 2$ sampel dengan $H_0: h_1(t) = h_2(t)$ dengan $\alpha = 0.05$. Karena $p - value < \alpha = 0.05$, maka H_0 ditolak pada $a = 0.05$. Artinya ada perbedaan survival experience antara 2 kategori Umur (Age) tersebut.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ age, data = df)

        N Observed Expected (O-E)^2/E (O-E)^2/V
age=0 214       54     72.1      4.54      18.4
age=1  85       42     23.9     13.71      18.4

 Chisq= 18.4  on 1 degrees of freedom, p= 2e-05
```

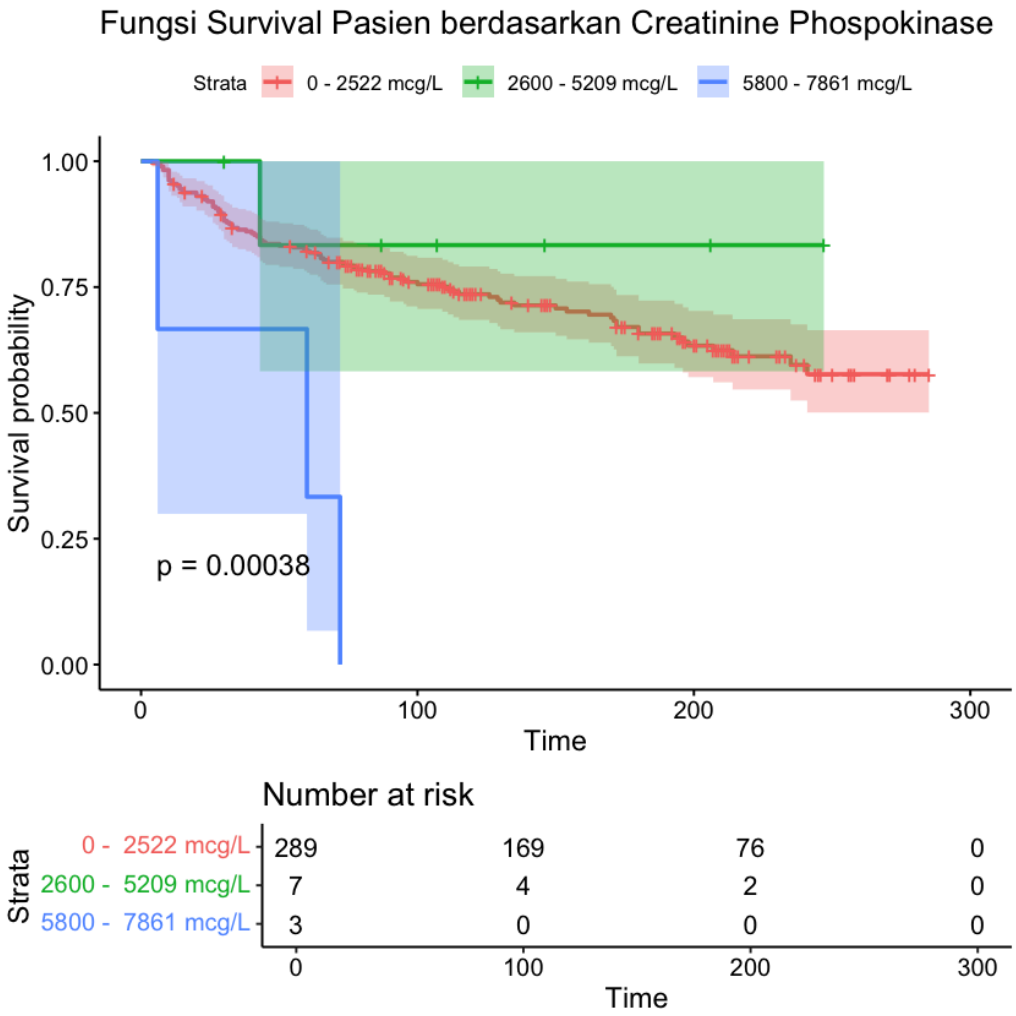
2. Anaemia



Akan dilakukan uji $K = 2$ sampel dengan $H_0: h_1(t) = h_2(t)$ dengan $\alpha = 0.05$. Karena $p - value > \alpha = 0.05$, maka H_0 gagal ditolak pada $a = 0.05$. Artinya tidak ada perbedaan survival experience antara pasien yang mengalami anaemia dan tidak mengalami anaemia.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ anaemia, data = df)

            N Observed Expected (O-E)^2/E (O-E)^2/V
anaemia=0 170       50     57.9      1.07      2.73
anaemia=1 129       46     38.1      1.63      2.73

 Chisq= 2.7  on 1 degrees of freedom, p= 0.1
```

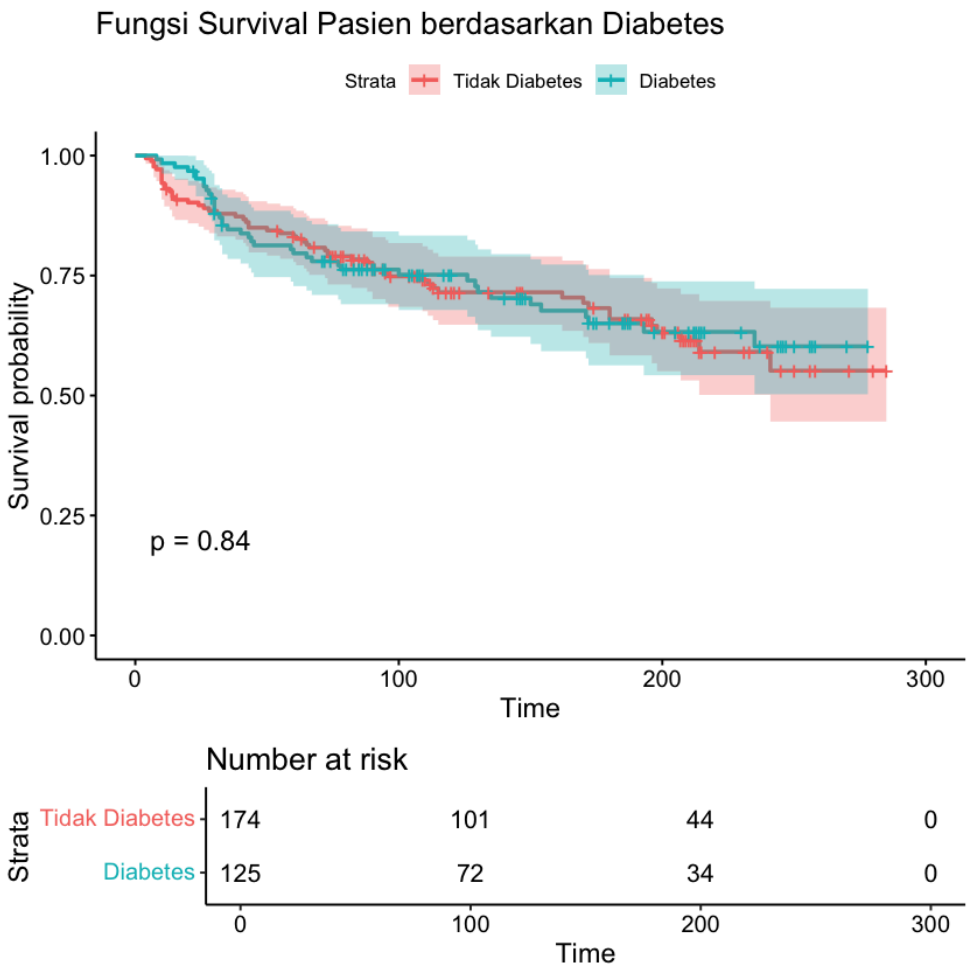
3. Creatinin Phosphokinase



Akan dilakukan uji $K = 3$ sampel dengan $H_0: h_1(t) = h_2(t) = h_3(t)$ dengan $\alpha = 0.05$. Karena $p - value < \alpha = 0.05$, maka H_0 ditolak pada $a = 0.05$. Artinya ada perbedaan survival experience antara 3 kategori Creatinine Phosphokinase tersebut.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ creatinine_phosphokinase,
    data = df)

                               N Observed Expected (O-E)^2/E (O-E)^2/V
creatinine_phosphokinase=0 289       92   93.373    0.0202     0.742
creatinine_phosphokinase=1   7        1    2.188    0.6450     0.663
creatinine_phosphokinase=2   3        3    0.439   14.9499    15.129

 Chisq= 15.7  on 2 degrees of freedom, p= 4e-04
```

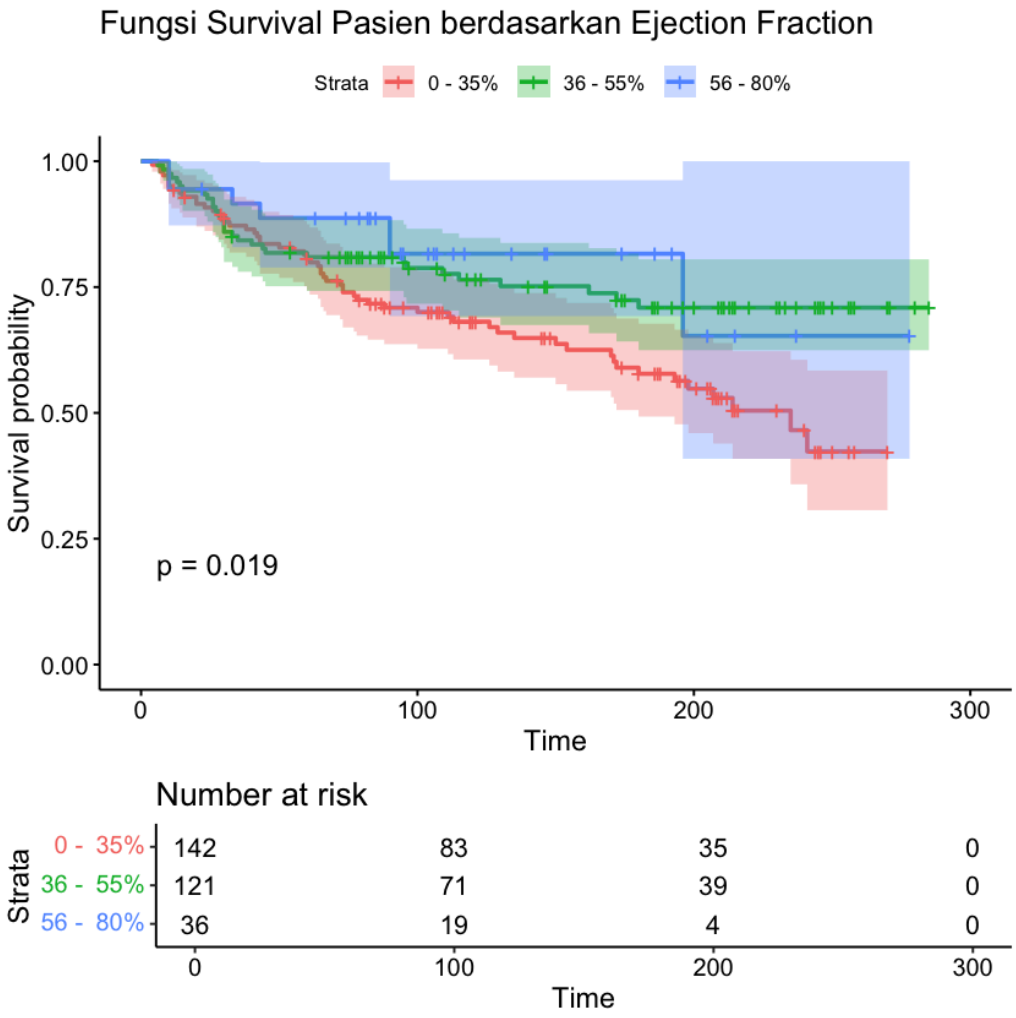
4. Diabetes



Akan dilakukan uji $K = 2$ sampel dengan $H_0: h_1(t) = h_2(t)$ dengan $\alpha = 0.05$. Karena $p - value > \alpha = 0.05$, maka H_0 gagal ditolak pada $a = 0.05$. Artinya tidak ada perbedaan survival experience antara pasien yang mengalami diabetes dan tidak mengalami diabetes.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ diabetes, data = df)

             N Observed Expected (O-E)^2/E (O-E)^2/V
diabetes=0 174       56       55    0.0172    0.0405
diabetes=1 125       40       41    0.0231    0.0405

 Chisq= 0  on 1 degrees of freedom, p= 0.8
```

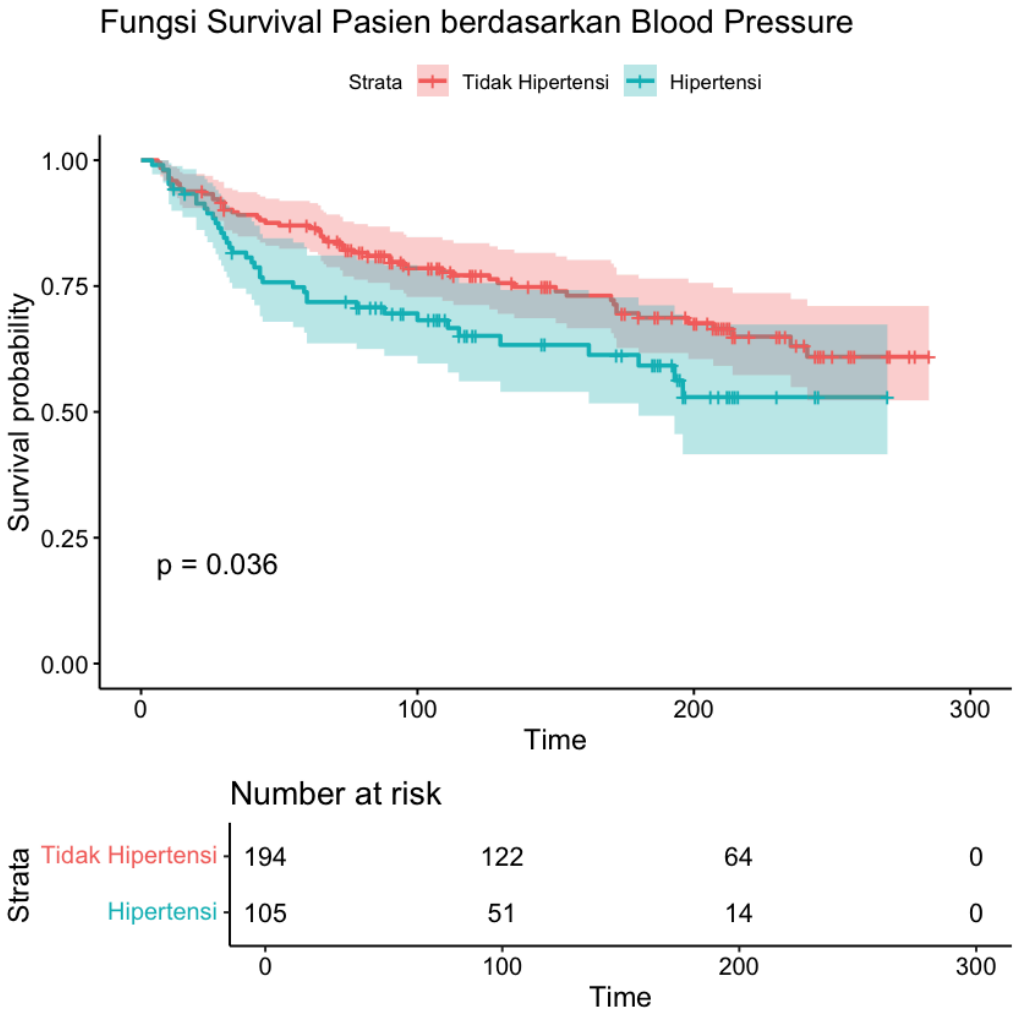
5. Ejection Fraction



Akan dilakukan uji $K = 3$ sampel dengan $H_0: h_1(t) = h_2(t) = h_3(t)$ dengan $\alpha = 0.05$. Karena $p-value < \alpha = 0.05$, maka H_0 ditolak pada $a = 0.05$. Artinya ada perbedaan survival experience antara 3 kategori Ejection Fraction tersebut.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ ejection_fraction,
    data = df)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
ejection_fraction=0 142       58     44.4      4.19      7.83
ejection_fraction=1 121       31     40.8      2.36      4.15
ejection_fraction=2  36        7     10.8      1.34      1.53

 Chisq= 7.9  on 2 degrees of freedom, p= 0.02
```

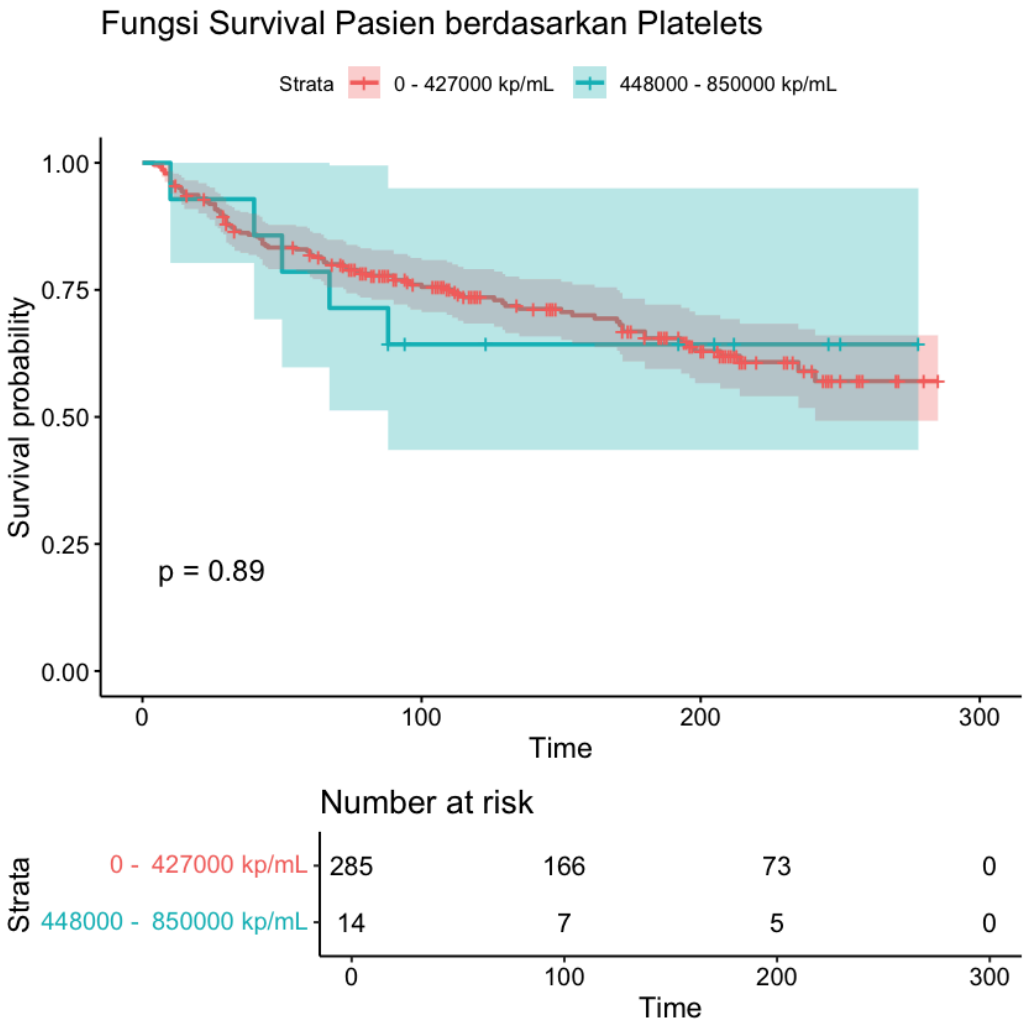
6. Blood Pressure



Akan dilakukan uji $K = 2$ sampel dengan $H_0: h_1(t) = h_2(t)$ dengan $\alpha = 0.05$. Karena $p - value < \alpha = 0.05$, maka H_0 ditolak pada $a = 0.05$. Artinya ada perbedaan survival experience antara pasien yang mengalami hipertensi dan tidak mengalami hipertensi.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ high_blood_pressure,
    data = df)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
high_blood_pressure=0 194       57     66.4      1.34      4.41
high_blood_pressure=1 105       39     29.6      3.00      4.41

 Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```

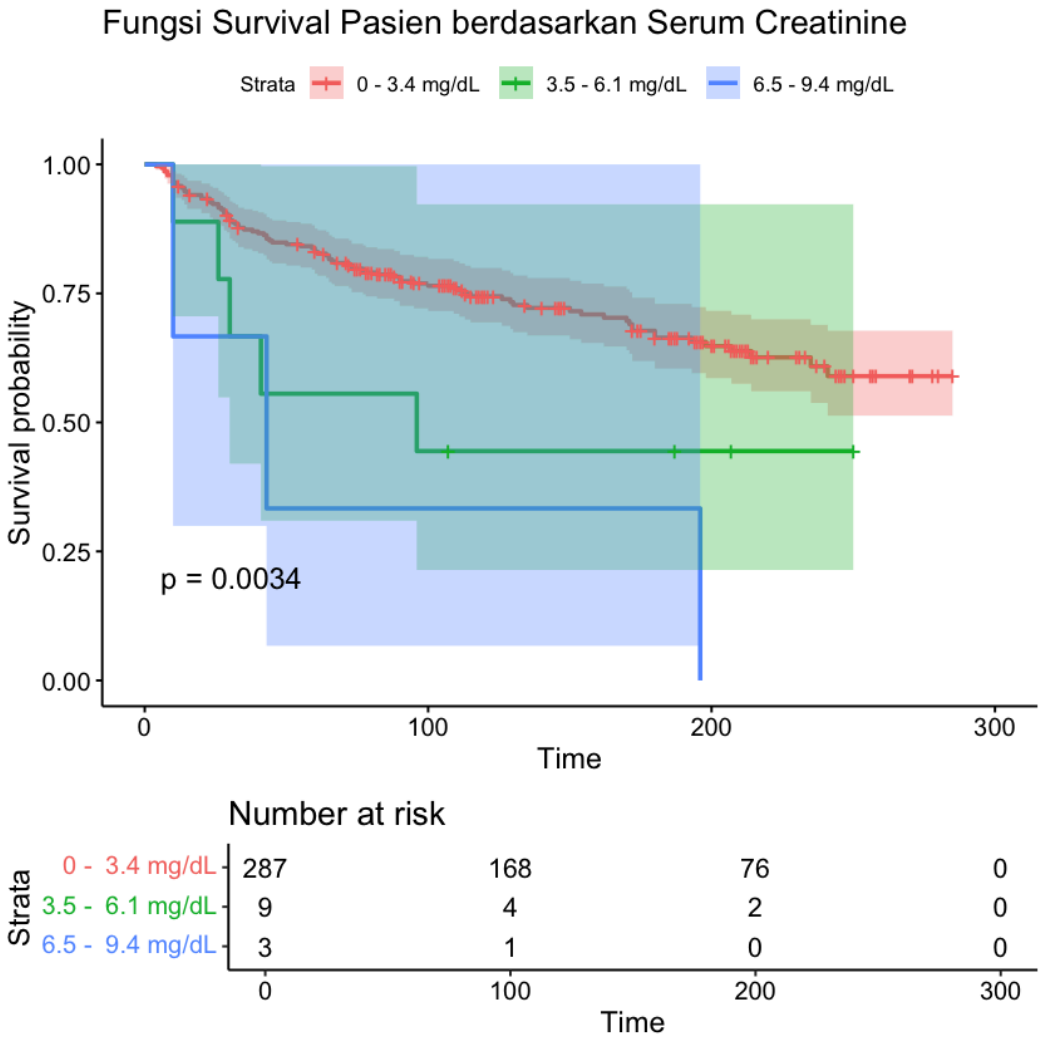
7. Platelets



Akan dilakukan uji $K = 2$ sampel dengan $H_0: h_1(t) = h_2(t)$ dengan $\alpha = 0.05$. Karena $p-value > \alpha = 0.05$, maka H_0 gagal ditolak pada $a = 0.05$. Artinya tidak ada perbedaan survival experience antara 2 kategori Platelets tersebut.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ platelets, data = df)

               N Observed Expected (O-E)^2/E (O-E)^2/V
platelets=0 285       91    91.29   0.00095    0.0195
platelets=1  14        5     4.71   0.01843    0.0195

 Chisq= 0  on 1 degrees of freedom, p= 0.9
```

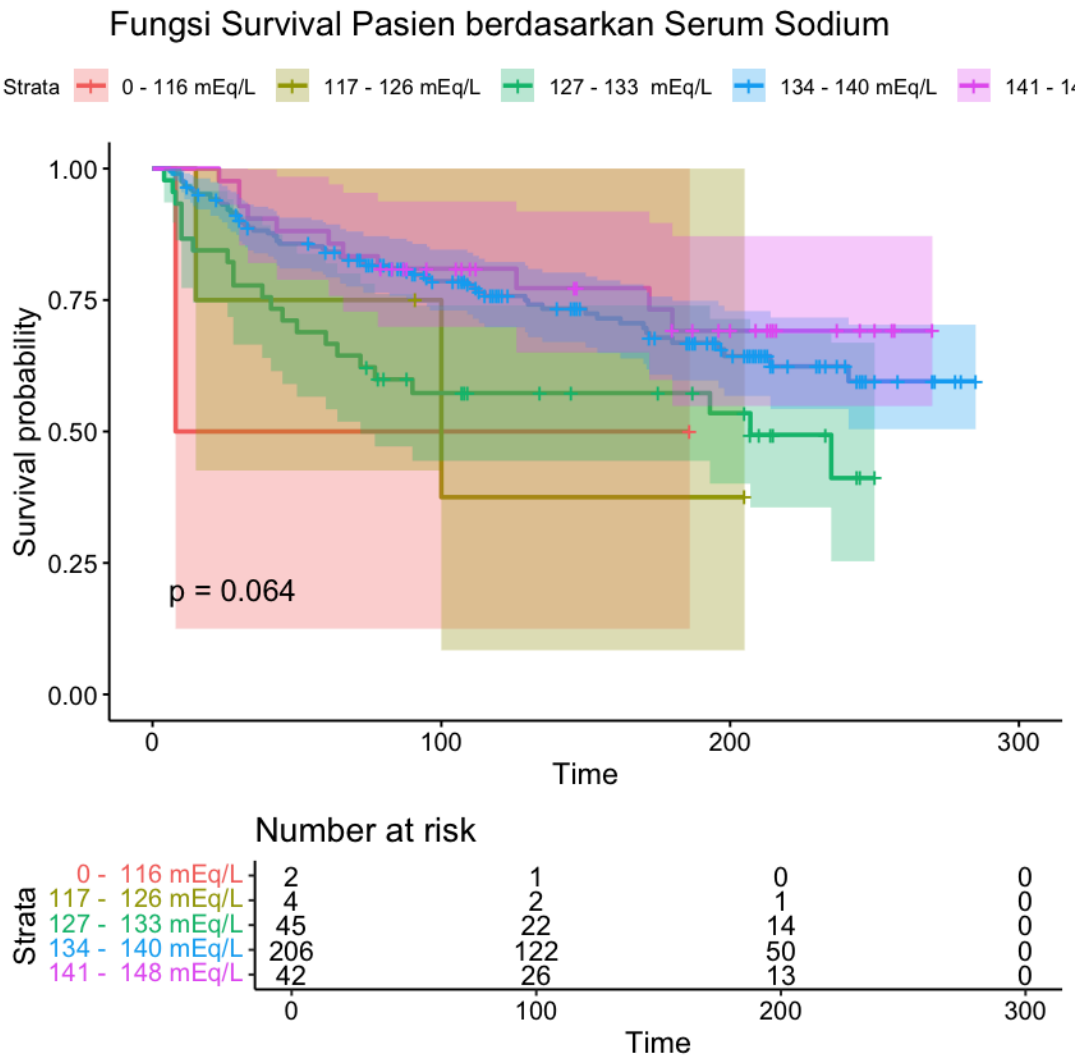
8. Serum Creatinine



Akan dilakukan uji $K = 3$ sampel dengan $H_0: h_1(t) = h_2(t) = h_3(t)$ dengan $\alpha = 0.05$. Karena $p - value < \alpha = 0.05$, maka $H0$ ditolak pada $a = 0.05$. Artinya ada perbedaan survival experience antara 3 kategori Serum Creatinine tersebut.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ serum_creatinine,
    data = df)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
serum_creatinine=0 287       88   92.913      0.26      8.12
serum_creatinine=1   9        5    2.428      2.73      2.81
serum_creatinine=2   3        3    0.659      8.32      8.42

 Chisq= 11.4  on 2 degrees of freedom, p= 0.003
```

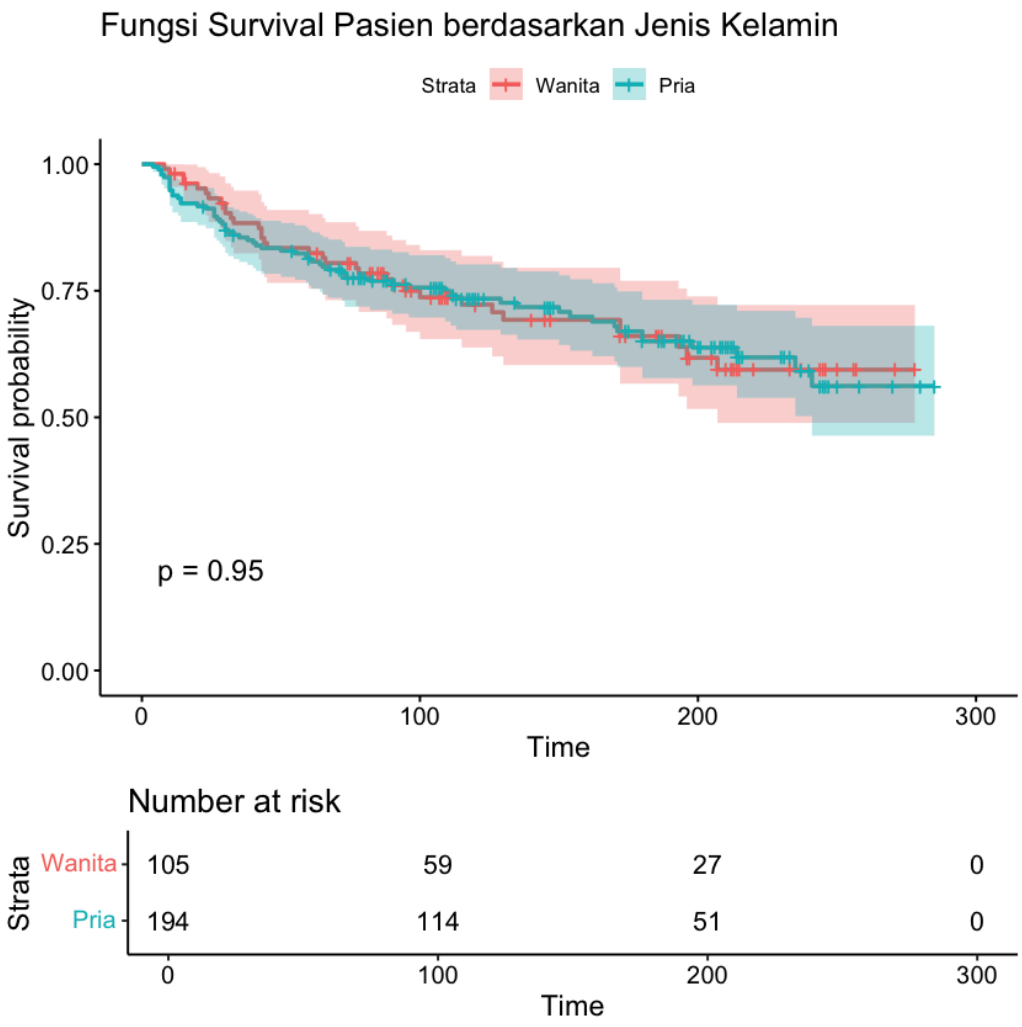
9. Serum Sodium



Akan dilakukan uji $K = 5$ sampel dengan $H_0: h_1(t) = h_2(t) = h_3(t) = h_4(t) = h_5(t)$ dengan $\alpha = 0.05$. Karena $p-value > \alpha = 0.05$, maka H_0 gagal ditolak pada $a = 0.05$. Artinya ada perbedaan survival experience antara 5 kategori Serum Sodium tersebut.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ serum_sodium, data = df)

                N Observed Expected (O-E)^2/E (O-E)^2/V
serum_sodium=0  2        1    0.443     0.702     0.708
serum_sodium=1  4        2    1.081     0.782     0.795
serum_sodium=2 45       22   13.257     5.766     6.723
serum_sodium=3 206      60   66.445     0.625     2.040
serum_sodium=4  42       11   14.775     0.964     1.144

 Chisq= 8.9  on 4 degrees of freedom, p= 0.06
```

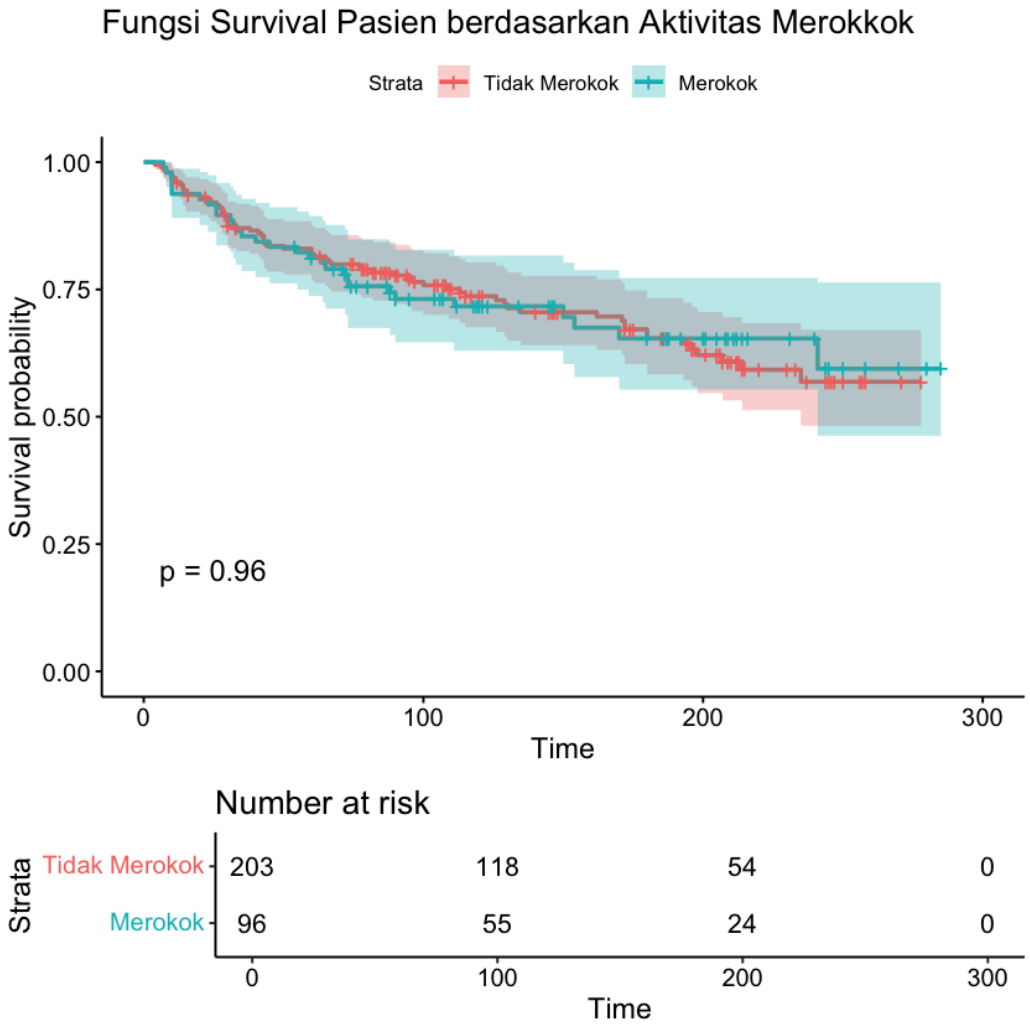
10. Sex



Akan dilakukan uji $K = 2$ sampel dengan $H_0 : h_1(t) = h_2(t)$ dengan $\alpha = 0.05$. Karena $p - value > \alpha = 0.05$, maka H_0 gagal ditolak pada $a = 0.05$. Artinya tidak ada perbedaan survival experience antara pasien berjenis kelamin pria maupun wanita.

```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ sex, data = df)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=0 105       34     34.3   0.00254   0.00397
sex=1 194       62     61.7   0.00141   0.00397

 Chisq= 0  on 1 degrees of freedom, p= 0.9
```

11. Smoking

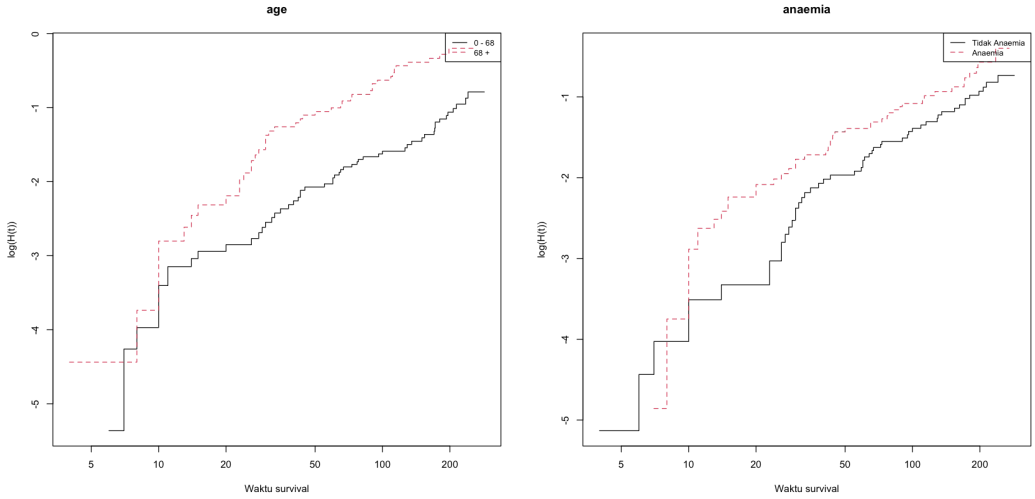


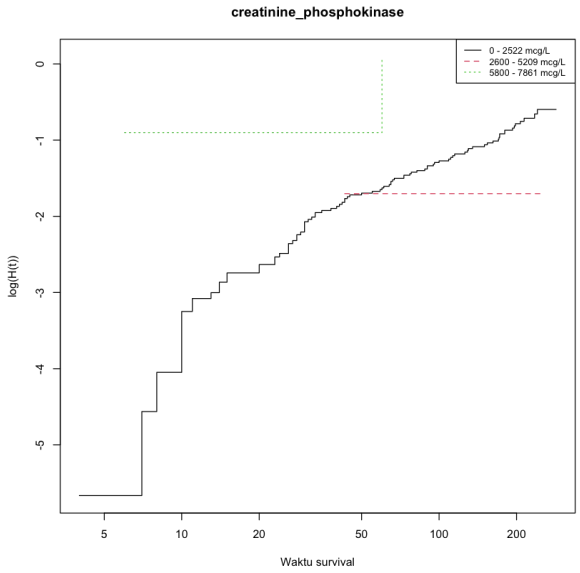
Akan dilakukan uji $K = 2$ sampel dengan $H_0: h_1(t) = h_2(t)$ dengan $\alpha = 0.05$. Karena $p - value > \alpha = 0.05$, maka H_0 gagal ditolak pada $a = 0.05$. Artinya tidak ada perbedaan survival experience antara pasien yang merokok dan tidak merokok.

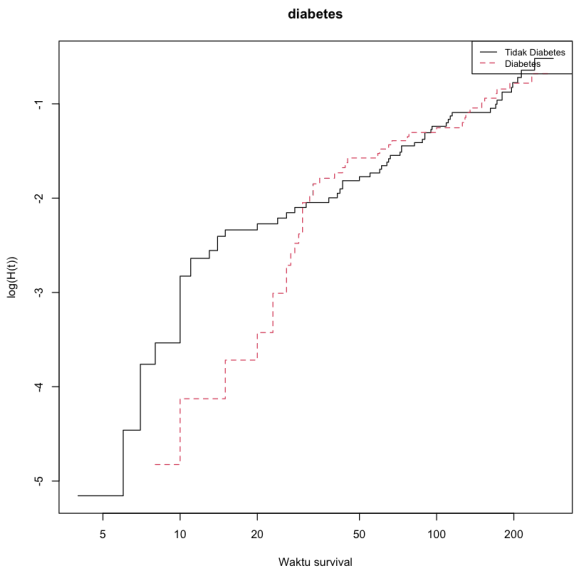
```
Call:
survdiff(formula = Surv(time, DEATH_EVENT) ~ smoking, data = df)

            N Observed Expected (O-E)^2/E (O-E)^2/V
smoking=0 203       66     65.8   0.00064   0.00204
smoking=1  96       30     30.2   0.00139   0.00204

 Chisq= 0  on 1 degrees of freedom, p= 1
```

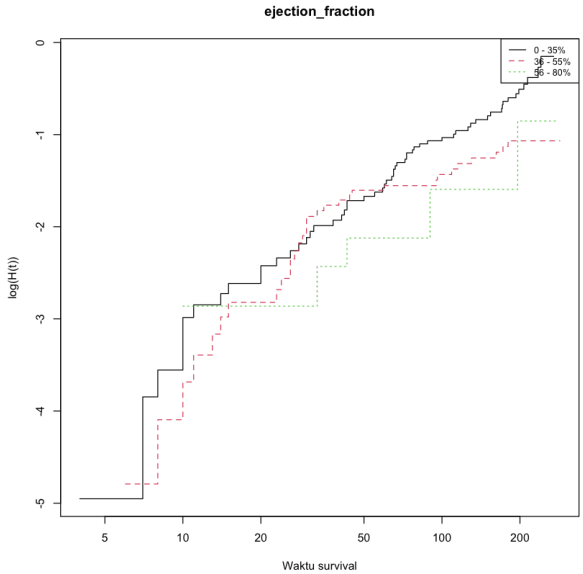
Cox PH Modelling:

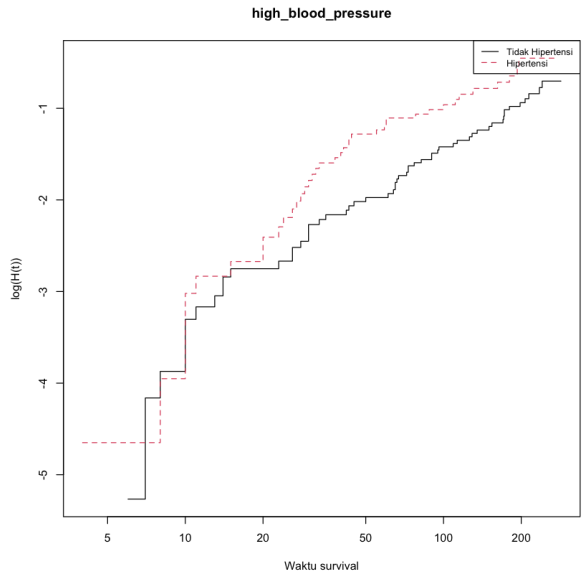
1. Cox PH Assumption

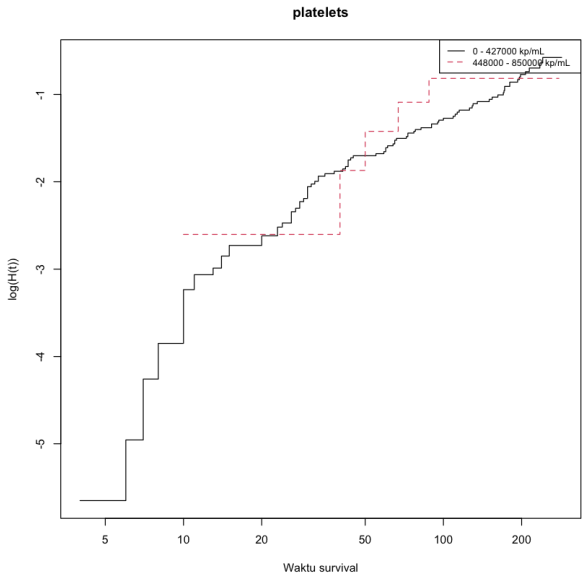


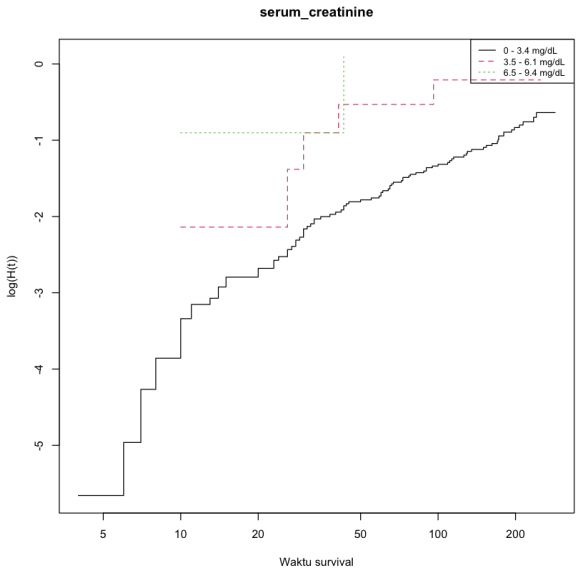

creatinine_phosphokinase
0 - 2522 mcg/L
2600 - 5209 mcg/L
5800 - 7861 mcg/L
log(H(t))
Waktu survival

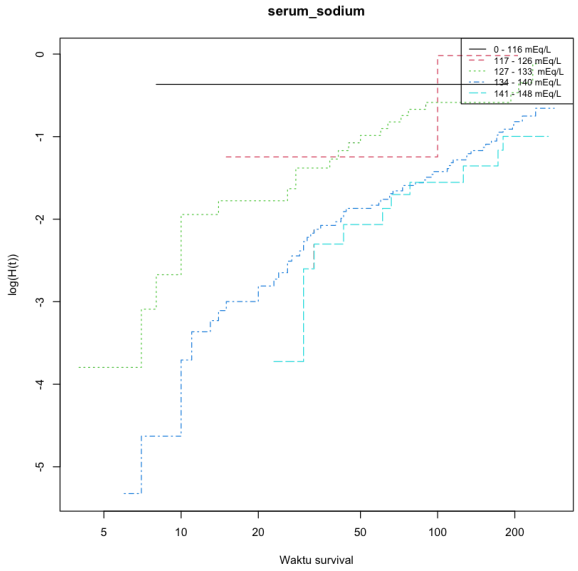

diabetes
Tidak Diabetes
Diabetes
log(H(t))
Waktu survival

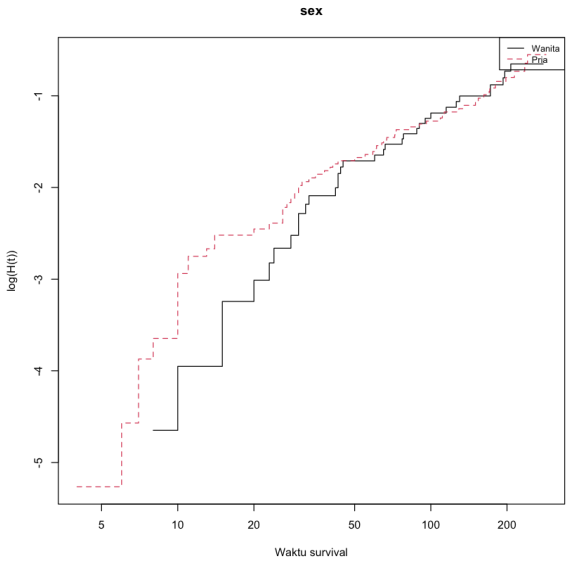

ejection_fraction
0 - 35%
36 - 55%
56 - 80%
log(H(t))
Waktu survival

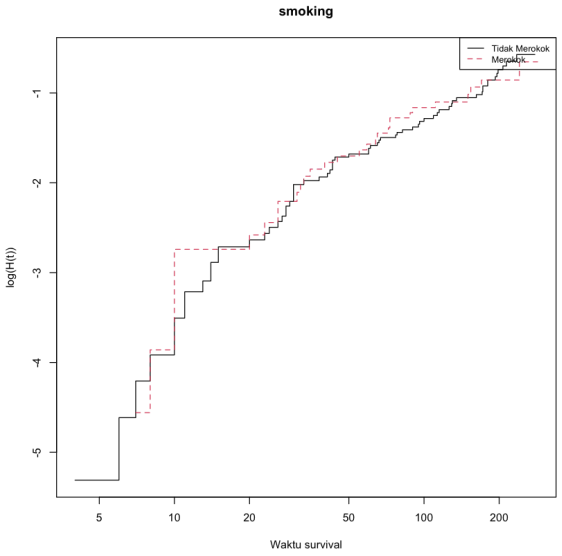

high_blood_pressure
Tidak Hipertensi
Hipertensi
log(H(t))
Waktu survival


platelets
0 - 427000 kp/mL
448000 - 850000 kp/mL
log(H(t))
-1
-2
-3
-4
-5
5
10
20
50
100
200
Waktu survival


serum_creatinine
0 - 3.4 mg/dL
3.5 - 6.1 mg/dL
6.5 - 9.4 mg/dL
log(H(t))
0
-1
-2
-3
-4
-5
5
10
20
50
100
200
Waktu survival


serum_sodium
0 - 116 mEq/L
117 - 126 mEq/L
127 - 133 mEq/L
134 - 140 mEq/L
141 - 148 mEq/L
log(H(t))
0
-1
-2
-3
-4
-5
5
10
20
50
100
200
Waktu survival


sex
Wanita
Pria
log(H(t))
-1
-2
-3
-4
-5
5
10
20
50
100
200
Waktu survival



Kesimpulan:

Berdasarkan visualisasi asumsi PH secara grafis tersebut , ada beberapa variabel yang tidak memenuhi asumsi.

2. Cox Ph Modelling

```
Call:
coxph(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
    creatinine_phosphokinase + serum_creatinine + anaemia + high_blood_pressure +
    sex + platelets + serum_sodium + smoking + diabetes, data = df)

  n= 299, number of events= 96

                             coef exp(coef) se(coef)      z Pr(>|z|)
age1                      0.94126   2.56321  0.22120  4.255 2.09e-05 ***
ejection_fraction1       -0.46061   0.63090  0.23112 -1.993  0.04627 *
ejection_fraction2       -1.10094   0.33256  0.45922 -2.397  0.01651 *
creatinine_phosphokinase1 -0.80254   0.44819  1.01356 -0.792  0.42847
creatinine_phosphokinase2  1.61438   5.02479  0.63470  2.544  0.01097 *
serum_creatinine1          0.69962   2.01299  0.49384  1.417  0.15657
serum_creatinine2          2.20246   9.04725  0.69029  3.191  0.00142 **
anaemia1                   0.48079   1.61735  0.21595  2.226  0.02599 *
high_blood_pressure1       0.30809   1.36082  0.22106  1.394  0.16341
sex1                      -0.04051   0.96030  0.25657 -0.158  0.87455
platelets1                 0.05428   1.05578  0.47602  0.114  0.90921
serum_sodium1             -0.35780   0.69921  1.26339 -0.283  0.77702
serum_sodium2             -0.98361   0.37396  1.06268 -0.926  0.35466
serum_sodium3             -1.26713   0.28164  1.04143 -1.217  0.22371
serum_sodium4             -1.56176   0.20977  1.08469 -1.440  0.14992
smoking1                  -0.04652   0.95454  0.25369 -0.183  0.85449
diabetes1                  0.08431   1.08796  0.22483  0.375  0.70768
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hasil analisis Cox proportional hazards (Cox PH) menunjukkan bahwa beberapa variabel tidak signifikan, atau gagal menolak H0, yang berarti variabel tersebut tidak berpengaruh dalam model Cox PH tersebut. Namun, variabel-variabel ini mungkin tetap berdampak, meskipun nilai p-value-nya lebih dari 0.05. Berdasarkan analisis, variabel yang tidak signifikan secara statistik masih bisa memiliki arti penting dari sudut pandang klinis atau kontekstual. Oleh karena itu, akan digunakan metode seleksi stepwise untuk menentukan variabel yang berpengaruh dalam konteks analisis ini. Metode Akaike Information Criterion (AIC) akan digunakan, dimana AIC adalah metode matematis untuk mengevaluasi seberapa baik model cocok dengan data yang dihasilkan. Dalam statistik, AIC digunakan untuk membandingkan berbagai model yang mungkin dan menentukan model mana yang paling sesuai dengan data. Semakin kecil nilai AIC, semakin baik model tersebut.

3. Variable Selection

Step 1

Start: AIC=999.14

Surv(time, DEATH_EVENT) ~ age + ejection_fraction + creatinine_phosphokinase +
serum_creatinine + anaemia + high_blood_pressure + sex +
platelets + serum_sodium + smoking + diabetes

Variabel	Df	AIC
serum_sodium	4	995.62
platelets	1	997.16
sex	1	997.17
smoking	1	997.18
diabetes	1	997.28
high_blood_pressure	1	999.04
<none>		999.14
creatinine_phosphokinase	2	1000.54
anemia	1	1002.04
serum_creatinine	2	1003.56
ejection_fraction	2	1004.15
age	1	1014.3

Step2

Step: AIC=995.62

Surv(time, DEATH_EVENT) ~ age + ejection_fraction + creatinine_phosphokinase +
serum_creatinine + anemia + high_blood_pressure + sex +
platelets + smoking + diabetes

Variabel	Df	AIC
sex	1	993.63
platelets	1	993.64
smoking	1	993.66
diabetes	1	994.04
high_blood_pressure	1	995.57
<none>		995.62
anemia	1	997.77
creatinine_phosphokinase	2	998.03
+ serum_sodium	4	999.14
serum_creatinine	2	1001.57
ejection_fraction	2	1001.68
age	1	1011.31

Step 3

Step: AIC=993.63

Surv(time, DEATH_EVENT) ~ age + ejection_fraction + creatinine_phosphokinase +
serum_creatinine + anaemia + high_blood_pressure + platelets +
smoking + diabetes

Variabel	Df	AIC
platelets	1	991.65
smoking	1	991.69
diabetes	1	992.05
<none>		993.63
high_blood_pressure	1	993.65
+sex	1	995.62
anaemia	1	995.77
creatinine_phosphokinase	2	996.04

+serum_sodium	4	997.17
serum_creatinine	2	999.59
ejection_fraction	2	999.74
age	1	1009.33

Step 4

Step: AIC=991.65

Surv(time, DEATH_EVENT) ~ age + ejection_fraction + creatinine_phosphokinase + serum_creatinine + anaemia + high_blood_pressure + smoking + diabetes

Variabel	Df	AIC
smoking	1	989.71
diabetes	1	990.09
<none>		991.65
high_blood_pressure	1	991.68
+platelets	1	993.63
+sex	1	993.64
anaemia	1	993.77
creatinine_phosphokinase	2	994.04
+serum_sodium	4	995.19
serum_creatinine	2	997.59
ejection_fraction	2	997.76
age	1	1007.75

Step 5

Step: AIC=989.71

Surv(time, DEATH_EVENT) ~ age + ejection_fraction + creatinine_phosphokinase + serum_creatinine + anaemia + high_blood_pressure + diabetes

Variabel	Df	AIC
diabetes	1	988.19
<none>		989.71

high_blood_pressure	1	989.81
+smoking	1	991.65
+sex	1	991.67
+platelets	1	991.69
anaemia	1	991.86
creatinine_phosphokinase	2	992.05
+serum_sodiun	4	993.26
serum_creatinine	2	995.64
ejection_fraction	2	995.81
age	1	1005.77

Step 6

Step: AIC=988.19

Surv(time, DEATH_EVENT) ~ age + ejection_fraction + creatinine_phosphokinase + serum_creatinine + anaemia + high_blood_pressure

Variabel	Df	AIC
<none>		988.19
high_blood_pressure	1	988.4
+diabetes	1	989.71
+smoking	1	990.09
+sex	1	990.12
+platelets	1	990.16
anaemia	1	990.32
creatinine_phosphokinase	2	990.46
+serum_sodium	4	991.44
serum_creatinine	2	993.73
ejection_fraction	2	994.58
age	1	1003.8

3. Final Model Cox PH

Dengan hasil akhir model terbaik dari proses stepwise diperoleh:
Surv(time, DEATH_EVENT) ~ age + ejection_fraction + creatinine_phosphokinase + serum_creatinine + anaemia + high_blood_pressure)

$$h(t,x) = h_0 e^{(\beta_1 age + \beta_2 ejection\ fraction + \beta_3 creatinine\ phosphokinase + \beta_4 serum\ creatinine + \beta_5 anaemia + \beta_6 high\ blood\ pressure)}$$

```
coxph(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
    creatinine_phosphokinase + serum_creatinine + anaemia + high_blood_pressure,
    data = df)

  n= 299, number of events= 96

                            coef exp(coef) se(coef)      z Pr(>|z|)
age1                      0.9094    2.4828   0.2103  4.325 1.53e-05 ***
ejection_fraction1       -0.5092    0.6010   0.2265 -2.248  0.02456 *
ejection_fraction2       -1.0985    0.3334   0.4469 -2.458  0.01397 *
creatinine_phosphokinase1 -0.8863    0.4122   1.0102 -0.877  0.38028
creatinine_phosphokinase2  1.7085    5.5208   0.6114  2.794  0.00520 **
serum_creatinine1          0.8198    2.2700   0.4658  1.760  0.07839 .
serum_creatinine2          2.2188    9.1964   0.6778  3.274  0.00106 **
anaemia1                   0.4358    1.5462   0.2131  2.045  0.04084 *
high_blood_pressure1       0.3247    1.3837   0.2157  1.505  0.13223
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Meskipun terdapat variabel dengan p-value > 0.05 (tidak signifikan) yaitu high_blood_pressure, kami memutuskan untuk tetap menyertakannya. Hal ini dikarenakan pada pengujian sebelumnya, variabel high_blood_pressure terbukti menunjukkan perbedaan signifikan dalam pengalaman survival. Namun, kami tetap akan melakukan uji Schoenfeld untuk memastikan bahwa asumsi proportional hazards (PH) terpenuhi.

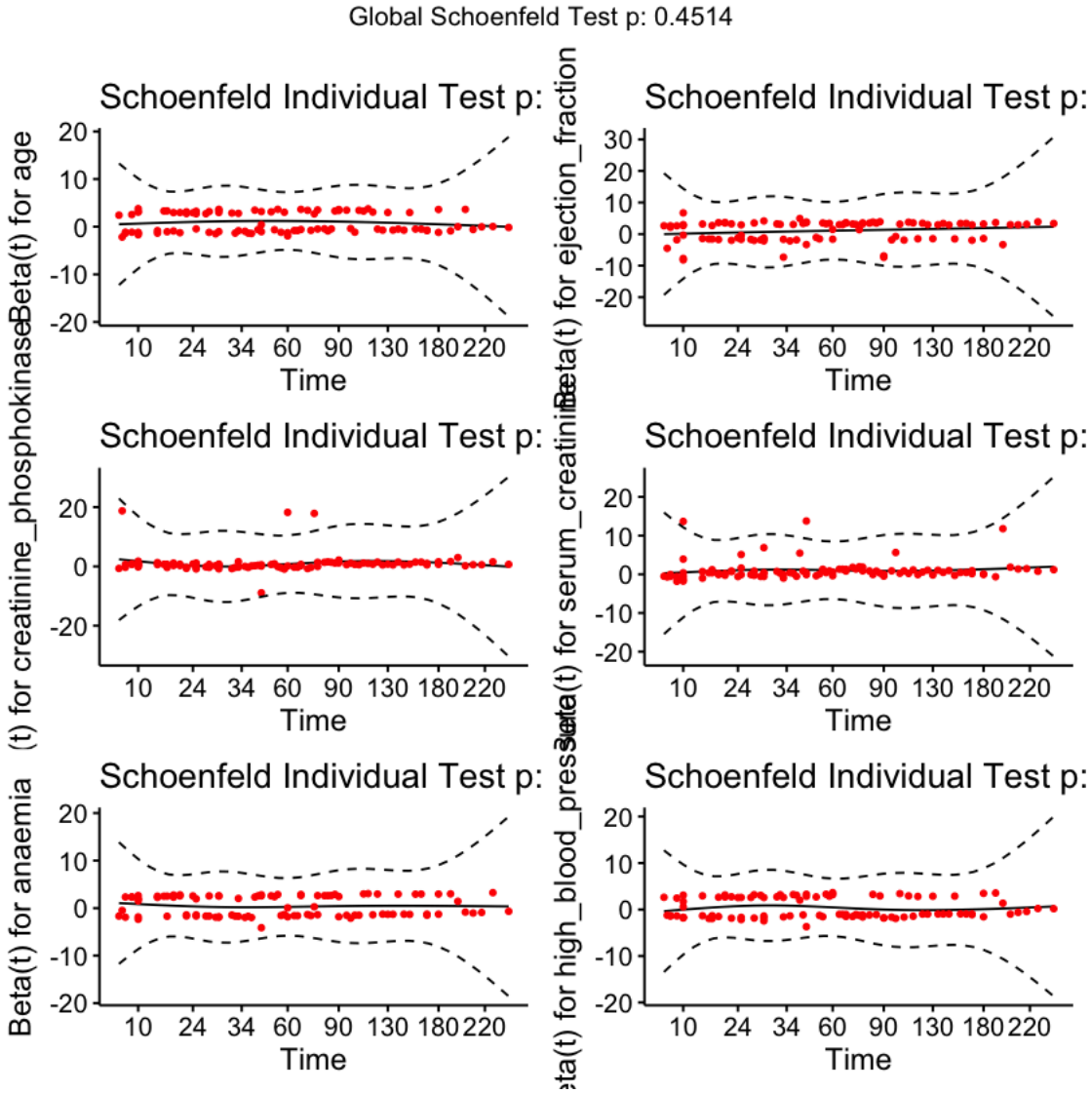
5. Model Diagnostic

- **Schoenfeld Test**

Variabel	Chi-Square	df	p-value
age	0.226	1	0.635
ejection_fraction	6.356	2	0.042

creatinine_phosphokinase	0.624	2	0.732
serum_creatinine	0.675	2	0.713
anaemia	0.268	1	0.605
high_blood_pressure	0.272	1	0.602
GLOBAL	8.848	9	0.451

Visualisasi:



Berdasarkan hasil uji Schoenfeld individual dan global, berikut kesimpulan yang dapat diambil:

- Uji Schoenfeld Global:

Nilai chi-square global adalah 8.848 dengan df (degree of freedom) 9 dan p-value sebesar 0.451.

Karena p-value global jauh lebih besar dari 0.05, kita tidak menolak hipotesis nol bahwa asumsi proportional hazards (PH) terpenuhi secara keseluruhan untuk model.

- Uji Schoenfeld Individual:

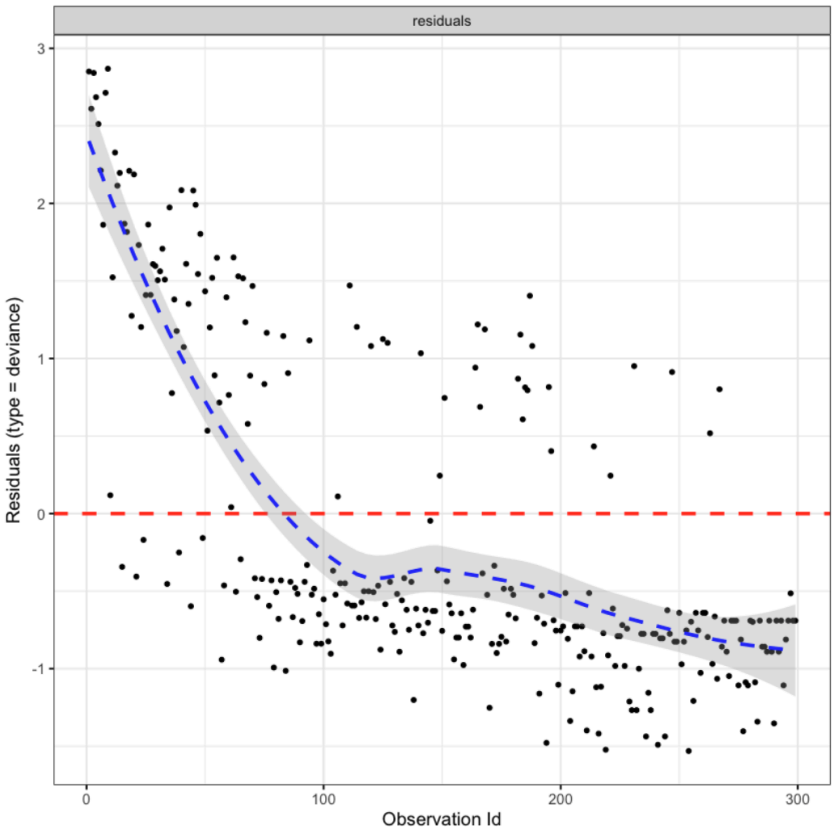
1. **age**: Nilai chi-square sebesar 0.226 dengan p-value 0.635, menunjukkan bahwa asumsi PH terpenuhi untuk variabel ini.
2. **ejection_fraction**: Nilai chi-square sebesar 6.356 dengan p-value 0.042, menunjukkan bahwa asumsi PH mungkin tidak terpenuhi untuk variabel ini (karena p-value < 0.05).
3. **creatinine_phosphokinase**: Nilai chi-square sebesar 0.624 dengan p-value 0.732, menunjukkan bahwa asumsi PH terpenuhi untuk variabel ini.
4. **serum_creatinine**: Nilai chi-square sebesar 0.675 dengan p-value 0.713, menunjukkan bahwa asumsi PH terpenuhi untuk variabel ini.
5. **anaemia**: Nilai chi-square sebesar 0.268 dengan p-value 0.605, menunjukkan bahwa asumsi PH terpenuhi untuk variabel ini.
6. **high_blood_pressure**: Nilai chi-square sebesar 0.272 dengan p-value 0.602, menunjukkan bahwa asumsi PH terpenuhi untuk variabel ini.

Kesimpulan Umum:
Secara keseluruhan, hasil uji Schoenfeld menunjukkan bahwa asumsi proportional hazards terpenuhi untuk sebagian besar variabel kecuali ejection_fraction. P-value global yang signifikan (0.451) menunjukkan bahwa model secara keseluruhan memenuhi asumsi PH, meskipun perlu perhatian lebih pada ejection_fraction.

Saran:

1. Model Lanjutan: Jika asumsi PH penting untuk semua variabel, Mungkin perlu melakukan penyesuaian model untuk memperhitungkan pelanggaran ini, atau menggunakan pendekatan alternatif untuk variabel yang tidak memenuhi asumsi PH.

- **Asumsi Normalitas Error (Normality Error Assumption)**



Dari grafik residual untuk setiap observasi, terlihat bahwa secara keseluruhan error dari setiap observasi berada pada nilai 0, hanya pada observasi dengan interval 0 sampai 100 yang tidak berada dekat dengan nilai 0. Karena secara mayoritas setiap error sudah terletak dekat dengan nilai 0, maka asumsi kenormalan dapat diterima.

Referensi:

Ivo, E. Mengukur Performa Machine Learning: Metrik untuk Mengukur Performa [Artikel].Diakses dari Metrik Untuk Mengukur Peforma Machine Learning.

The Ultimate Guide to Encoding Numerical Features in Machine Learning [Artikel]. Diakses dari The ultimate guide to Encoding Numerical Features in Machine Learning. | by Paresh Patil | Medium

Nahhas, R. W. Survival - PH Assumption [Artikel]. Diakses dari 7.16 Proportional hazards assumption | Introduction to Regression Methods for Public Health Using R

Akaike Information Criterion [Artikel]. Diakses dari Akaike Information Criterion - an overview | ScienceDirect Topics).