Analisis Perbandingan Algoritma Random Forest dan Logistic Regression Dalam Memprediksi Pindah Layanan Telekomunikasi(Churn)

Pokok Bahasan

01

Pendahuluan

(Latar Belakang, Rumusan Masalah, Tujuan Penelitian, Manfaat Penelitian, Teori RF & SVM)

02

Metodologi Penelitian

(Sumber Data, Struktur Data, Kondisi Data, Langkah Penelitian)

03

Hasil dan Pembahasan

(Perbandingan hasil antara algoritma Random Forest & Logistic Regression)

PENDAHULUAN

- Latar Belakang
- Rumusan Masalah
- Tujuan Penelitian
- Manfaat Penelitian
- Teori RF & LG

LATAR BELAKANG

- Industri telekomunikasi menghadapi **tantangan besar** dalam mengurangi churn, yaitu t**ingkat pelanggan yang berhenti atau tidak memperpanjang langganan.** Mengingat pentingnya layanan komunikasi dalam kehidupan sehari-hari, pelanggan sangat sensitif terhadap kualitas layanan yang diterima.

- Dengan dilakukannya analisis ini diharapakan dapat membantu **perusahaan dalam mengidentifikasi pelanggan yang berisiko, meningkatkan kualitas layanan, dan merancang strategi untuk mempertahankan pelanggan,** sehingga mendorong loyalitas dan pertumbuhan bisnis yang berkelanjutan.

RUMUSAN MASALAH

01

Bagaimana hasil memprediksi Pelanggan yang Berisiko Churn?

02

Bagaimana tingkat akurasi yang dihasilkan berdasarkan k-Fold Cross Validation & Repeated Holdout?

03

Metode mana yang memiliki tingkat akurasi paling baik?

TUJUAN PENELITIAN

01 Menganalisis Faktor Penyebab Churn

02 Memprediksi Pelanggan yang Berisiko Churn

03 Meningkatkan Retensi Pelanggan

04 Optimalisasi Biaya Akuisisi Pelanggan

TUJUAN PENELITIAN
& MANFAAT PENELITAN

01 Mahasiswa dapat mengaplikasikan ilmu perkuliahan dan mengetahui kondisi ril

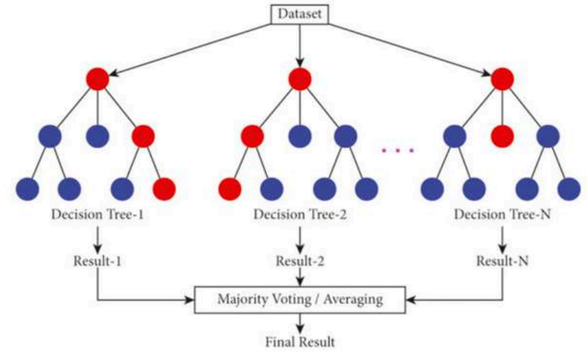
Perusahaan Telekomunikasi dapat menjadikan penelitian ini sebagai bahan evaluasi 02

03 Masyarakat dapat menambah ilmu dan wawasan seputar telekomunikasi

TEORI

RANDOM FOREST

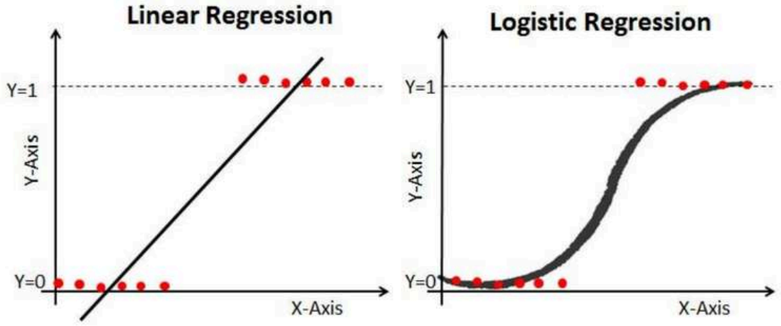
Random Forest adalah algoritma ensemble yang membangun banyak pohon keputusan (decision trees) untuk meningkatkan akurasi prediksi. Setiap pohon dibangun menggunakan subset acak dari data pelatihan dan subset acak dari fitur, kemudian hasil prediksi diambil berdasarkan mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi). Random Forest dapat menangani data yang tidak terstruktur dan memiliki kemampuan untuk mengatasi masalah overfitting pada model pohon keputusan tunggal.



Referensi:
Liaw, A., & Wiener, M. (2002). "Classification and Regression by randomForest". R News, 2(3), 18-22.

Logistic Regression

Logistic Regression adalah metode statistik untuk memodelkan hubungan antara variabel dependen biner (0 atau 1) dengan satu atau lebih variabel independen. Model ini menggunakan fungsi logistik (sigmoid) untuk memetakan output yang tidak terbatas menjadi nilai probabilitas antara 0 dan 1. Logistic Regression sangat populer dalam klasifikasi biner seperti prediksi churn atau deteksi spam.



Referensi:
Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. Wiley-Interscience.

TEORI

RANDOM FOREST

Random Forest adalah algoritma ensemble yang membangun banyak pohon keputusan (decision trees) untuk meningkatkan akurasi prediksi. Setiap pohon dibangun menggunakan subset acak dari data pelatihan dan subset acak dari fitur, kemudian hasil prediksi diambil berdasarkan mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi). Random Forest dapat menangani data yang tidak terstruktur dan memiliki kemampuan untuk mengatasi masalah overfitting pada model pohon keputusan tunggal.

Algoritma :

- Pilih subset acak dari data pelatihan untuk setiap pohon.
- Untuk setiap pohon, pilih subset acak dari fitur yang akan digunakan.
- Bangun pohon keputusan dengan cara split yang mengoptimalkan pembagian berdasarkan kriteria tertentu (misalnya, Gini impurity atau entropi).
- Prediksi menggunakan mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi) dari hasil pohon-pohon yang dibangun.
- Evaluasi hasil model menggunakan metrik seperti akurasi, precision, recall, atau F1-score.

Referensi:
Liaw, A., & Wiener, M. (2002). "Classification and Regression by randomForest". R News, 2(3), 18-22.

Logistic Regression

Logistic Regression adalah metode statistik untuk memodelkan hubungan antara variabel dependen biner (0 atau 1) dengan satu atau lebih variabel independen. Model ini menggunakan fungsi logistik (sigmoid) untuk memetakan output yang tidak terbatas menjadi nilai probabilitas antara 0 dan 1. Logistic Regression sangat populer dalam klasifikasi biner seperti prediksi churn atau deteksi spam.

Algoritma :

- Tentukan variabel dependen (target) dan variabel independen (fitur).
- Hitung log-odds (logaritma dari rasio peluang) menggunakan persamaan linear.
- Terapkan fungsi logistik untuk mengubah log-odds menjadi probabilitas.
- Tentukan threshold untuk memutuskan kelas (misalnya, jika probabilitas > 0.5, prediksi kelas 1).
- Evaluasi model menggunakan metrik seperti akurasi, precision, recall, dan AUC-ROC.

Referensi:
Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. Wiley-Interscience.

METODOLOGI PENELITIAN

- Sumber Data
- Struktur Data
- Kondisi Data
- Langkah Penelitian

SUMBER DATA

Dataset ini mencakup informasi tentang pelanggan yang berhenti berlangganan (Churn), layanan yang digunakan seperti telepon, internet, dan streaming, serta detail akun pelanggan seperti lama berlangganan, jenis kontrak, metode pembayaran, biaya bulanan, dan total biaya. Selain itu, terdapat data demografis pelanggan seperti jenis kelamin, usia, serta status pasangan dan tanggungan.

Link :
https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data

VARIABEL

Identitas Pelanggan

1. ID Pelanggan: Identitas unik pelanggan
2. Gender: Jenis kelamin (Laki-laki / Perempuan)
3. SeniorCitizen: Status sebagai warga lanjut usia (Ya/Tidak)
4. Partner: Status memiliki pasangan (Ya / Tidak)
5. Dependents: Status memiliki tanggungan (Ya / Tidak)

Informasi Akun dan Layanan

6. Tenure: Durasi pelanggan menjadi anggota (dalam bulan)
7. PhoneService: Status memiliki layanan telepon (Ya / Tidak)
8. MultipleLines: Status memiliki banyak saluran (Ya / Tidak / Tidak ada layanan telepon)
9. InternetService: Jenis layanan internet (DSL / Fiber optic / Tidak ada)

VARIABEL

Layanan Tambahan

10. OnlineSecurity: Status memiliki keamanan online (Ya / Tidak / Tidak ada layanan internet)
11. OnlineBackup: Status memiliki cadangan online (Ya / Tidak / Tidak ada layanan internet)
12. DeviceProtection: Status memiliki perlindungan perangkat (Ya / Tidak / Tidak ada layanan internet)
13. TechSupport: Status memiliki dukungan teknis (Ya / Tidak / Tidak ada layanan internet)
14. StreamingTV: Status memiliki TV streaming (Ya / Tidak / Tidak ada layanan internet)
15. StreamingMovies: Status memiliki layanan streaming film (Ya / Tidak / Tidak ada layanan internet)

Kebijakan dan Pembayaran

16. Contract: Jenis kontrak (Bulan ke bulan / Satu tahun / Dua tahun)
17. PaperlessBilling: Status tagihan tanpa kertas (Ya / Tidak)
18. Metode Pembayaran: Metode pembayaran (Cek elektronik / Cek lewat pos / Transfer bank otomatis / Kartu kredit otomatis)

Biaya

19. MonthlyCharges: Biaya bulanan yang dibebankan
20. TotalCharges: Total biaya yang dibebankan

Status Churn

21. Churn: Status pelanggan berhenti (Ya / Tidak)

STRUKTUR DATA

No.	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	...	X_{21}
1	Y_1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,6}$	$X_{1,7}$	$X_{1,8}$	$X_{1,9}$	$X_{1,10}$	$X_{1,11}$	...	$X_{1,21}$
2	Y_2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	$X_{2,5}$	$X_{2,6}$	$X_{2,7}$	$X_{2,8}$	$X_{2,9}$	$X_{2,10}$	$X_{2,11}$	...	$X_{2,21}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7043	Y_{7043}	$X_{7043,1}$	$X_{7043,2}$	$X_{7043,3}$	$X_{1025,4}$	$X_{7043,5}$	$X_{7043,6}$	$X_{7043,7}$	$X_{7043,8}$	$X_{7043,9}$	$X_{7043,10}$	$X_{7043,11}$	...	$X_{7043,21}$

KONDISI DATA

JUMLAH DATA

Terdapat **7043 data** dengan **21 variabel** yang akan diolah dalam analisis ini dan data telah *balance*

MISSING VALUE

Tidak terdapat indikasi missing value pada data sehingga tidak perlu dilakukan imputasi.

KORELASI

Uji korelasi dilakukan menggunakan uji Pearson dan chi-square. Hanya **variabel 'gender', 'PhoneService'** yang **tidak berkorelasi dengan variabel Y.**

OUTLIER

Tidak terdapat indikasi outlier pada data

Langkah Penelitian

1. **Studi Literatur**
- Menggali metode Random Forest dan SVM serta mencari data
2. **Persiapan Data**
- Melakukan pengecekan dan mengatasi missing value & outlier
- Melakukan analisis eksplorasi data
- Mengecek keseimbangan frekuensi dan menguji hipotesis
3. **Preprocessing**
- Melakukan scaling data dan pembagian data (training-testing)
4. **Visualisasi**
5. **Pengujian**
- Menggali metode Random Forest dan Logistic Regression serta mencari data
6. **Analisis**
- Melakukan analisis dan prediksi menggunakan metode terpilih
7. **Evaluasi**
- Evaluasi model (K-Fold Cross Validation & Repeated Holdout)
8. **Kesimpulan**
- Menyimpulkan hasil penelitian

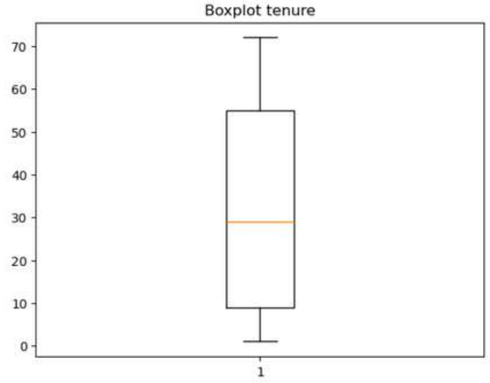
HASIL DAN PEMBAHASAN

- Preprocessing
- Summary Statistics and Visualization
- Feature Selection
- Random Forest
- Logistic Regression
- K-Fold & Repeated Holdout

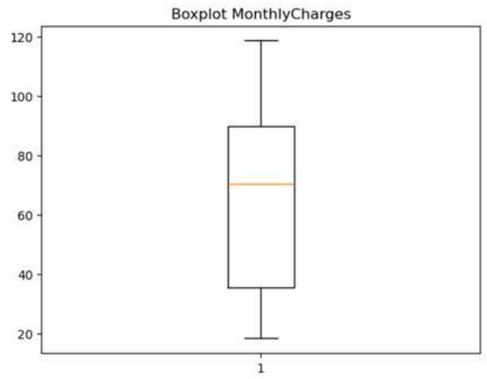
Preprocessing

```
data.isnull().sum()
```

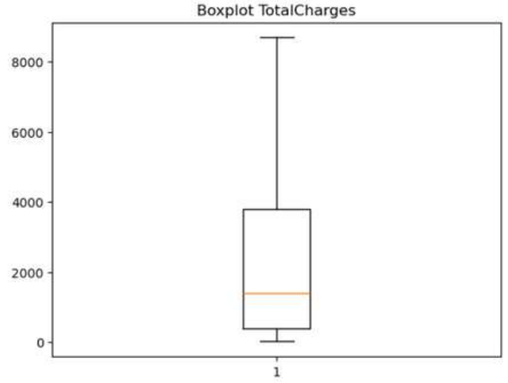
```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```




```
Variabel tenure:
Q1: 9.0
Q3: 55.0
IQR: 46.0
Outlier bawah: 0
Outlier atas: 0
```




```
Variabel MonthlyCharges:
Q1: 35.5875
Q3: 89.8625
IQR: 54.275
Outlier bawah: 0
Outlier atas: 0
```

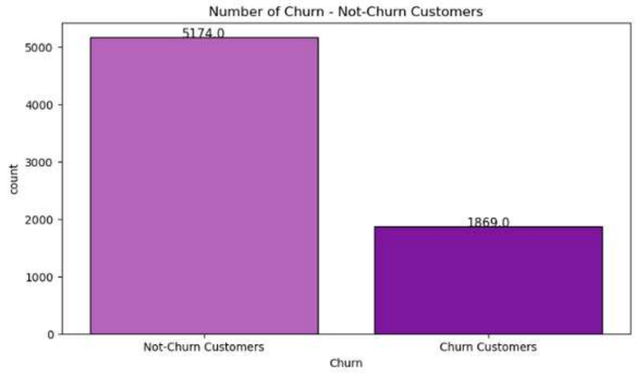


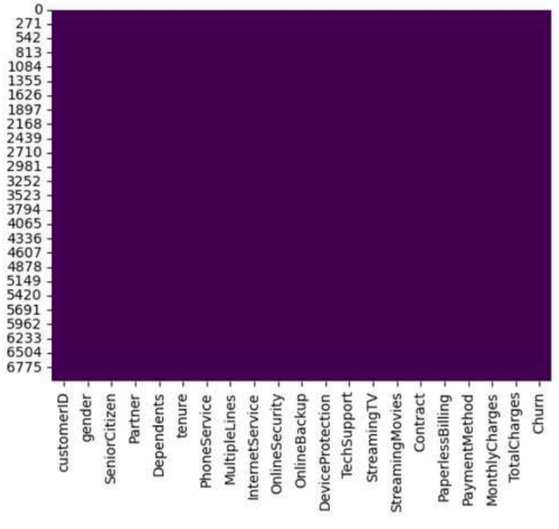

```
Variabel TotalCharges:
Q1: 401.45
Q3: 3794.7375
IQR: 3393.2875000000004
Outlier bawah: 0
Outlier atas: 0
```

```
: data.duplicated().sum()
```

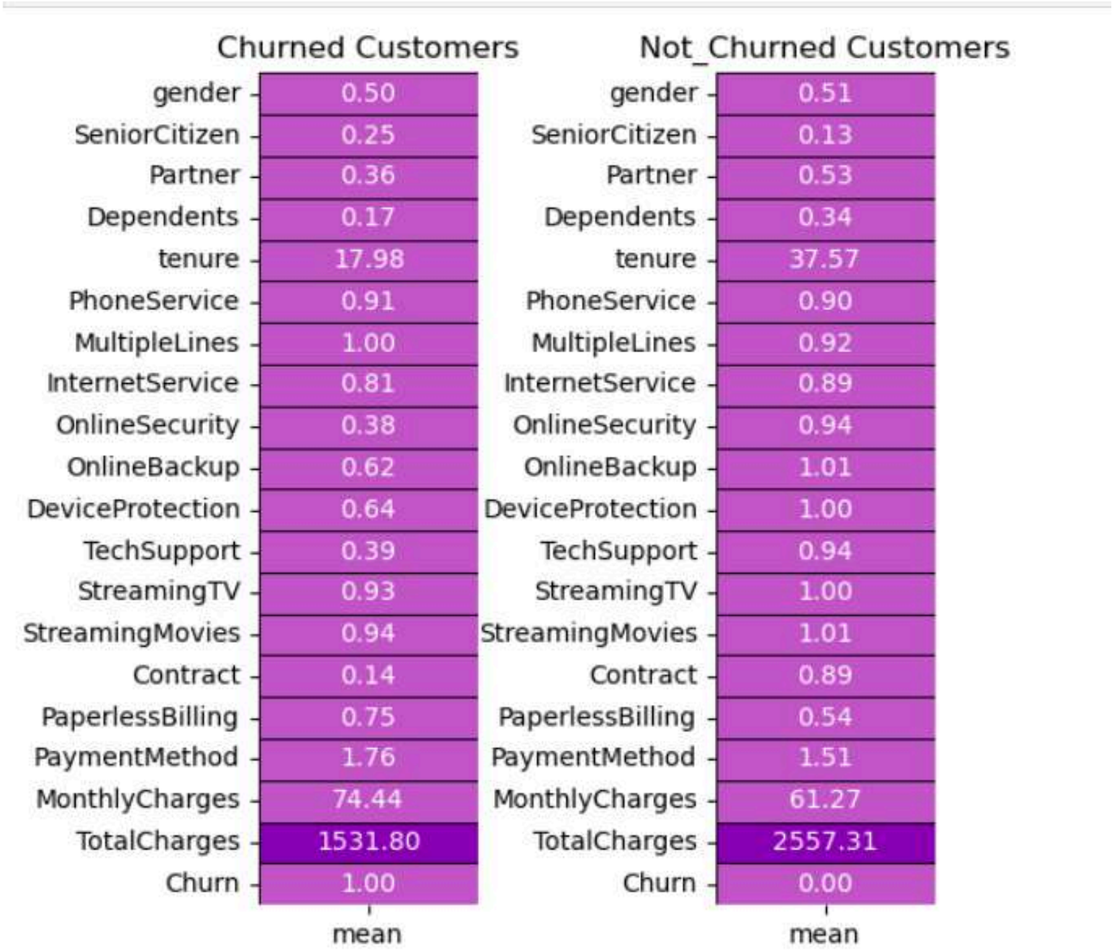
```
: 0
```

```
Jumlah Pelanggan Churn dan Tidak Churn:
Churn
No     5174
Yes    1869
Name: count, dtype: int64
```

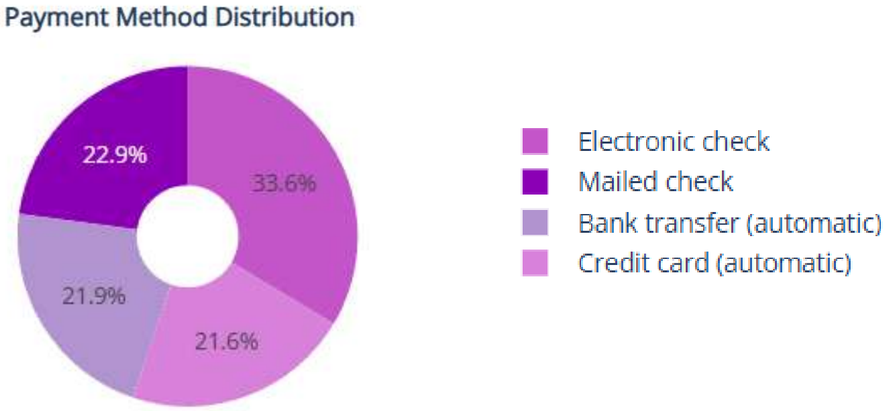


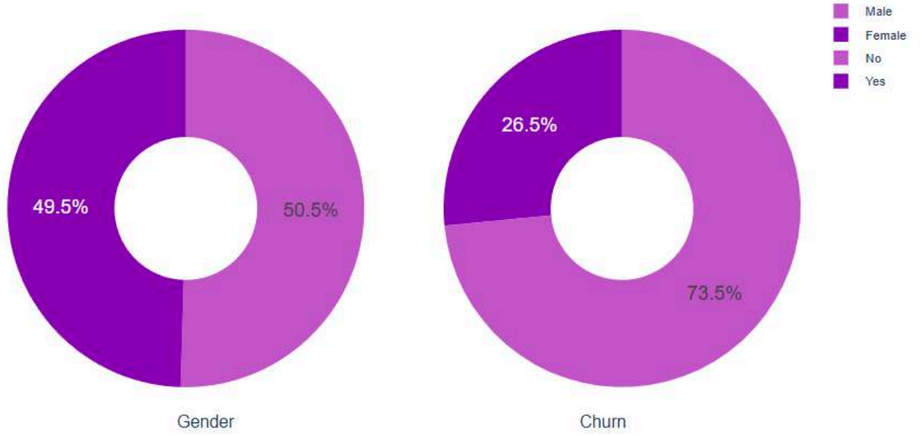


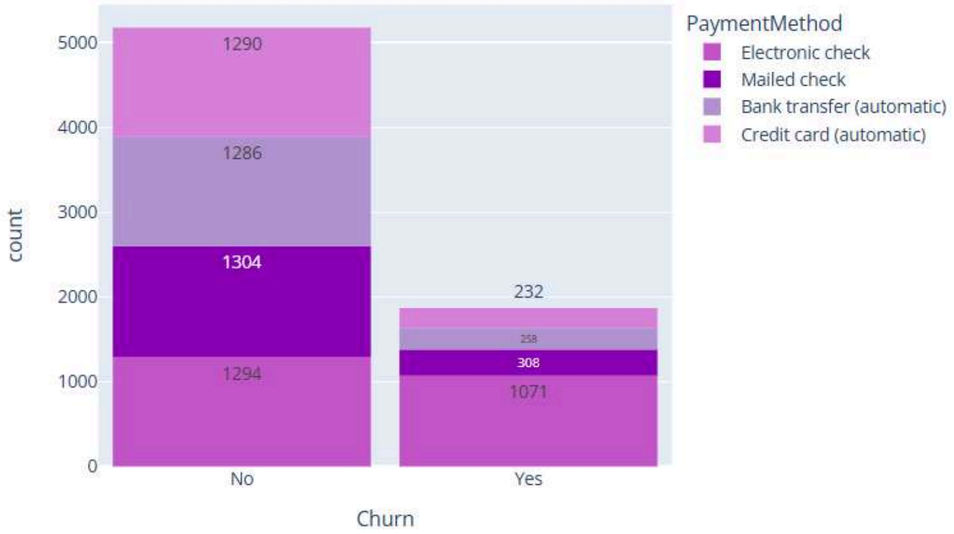
Analisis Eksplorasi Data

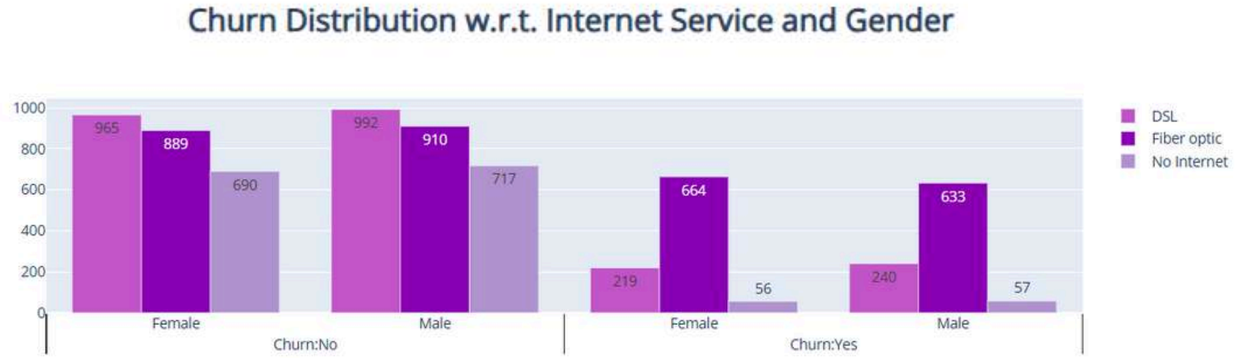


Analisis Eksplorasi Data

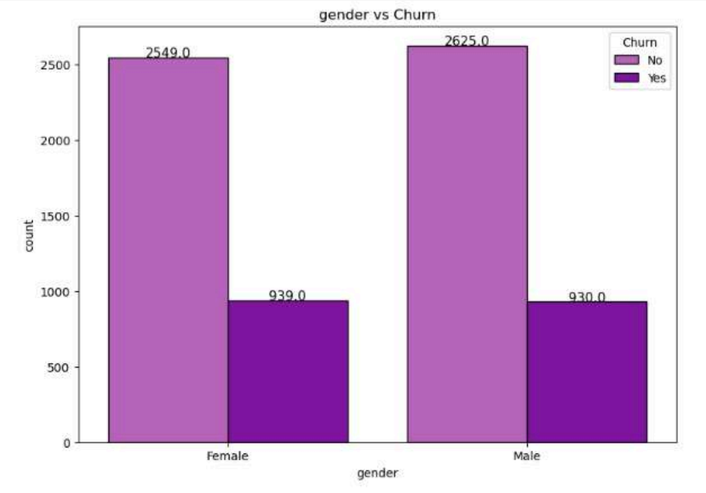


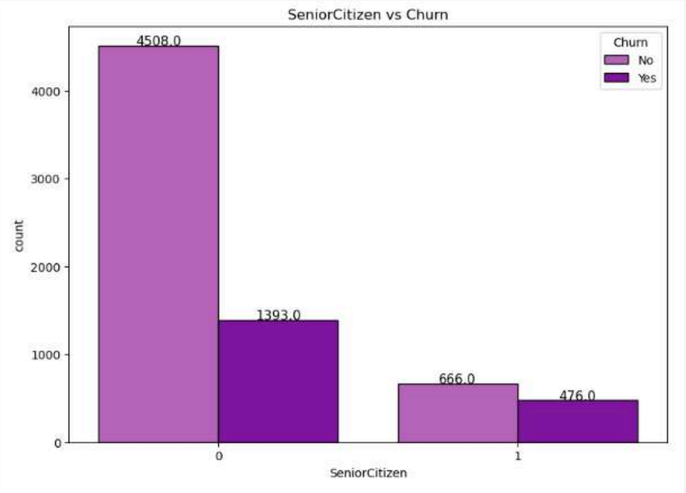


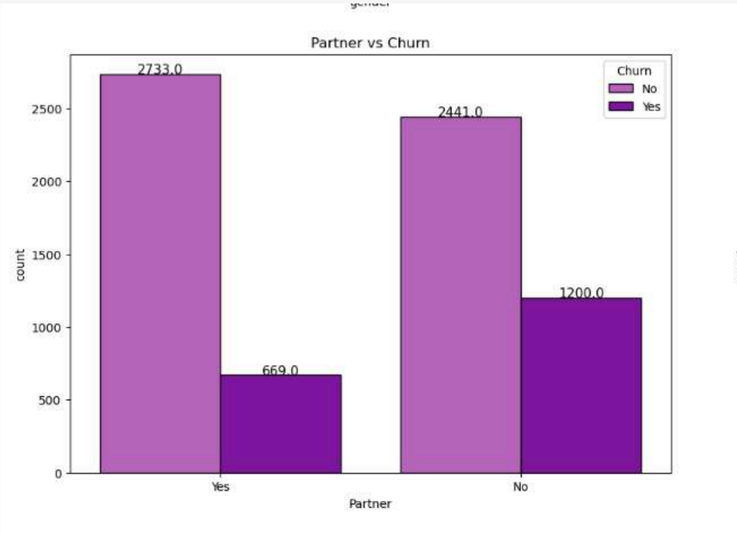


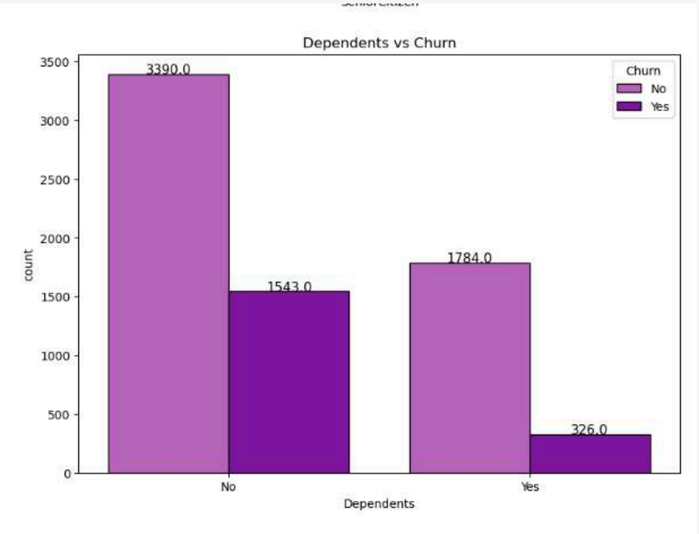


Analisis Eksplorasi Data

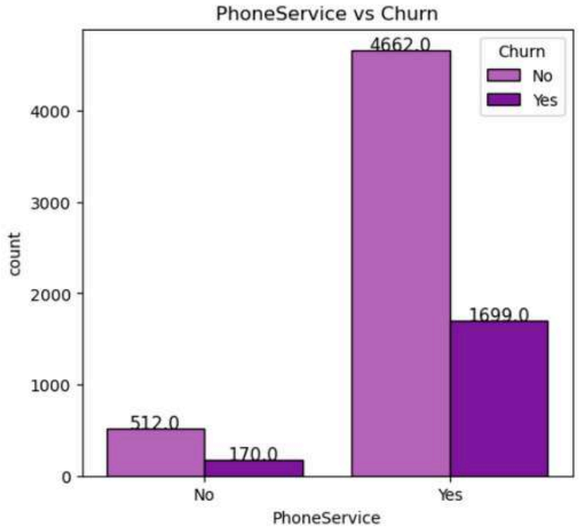


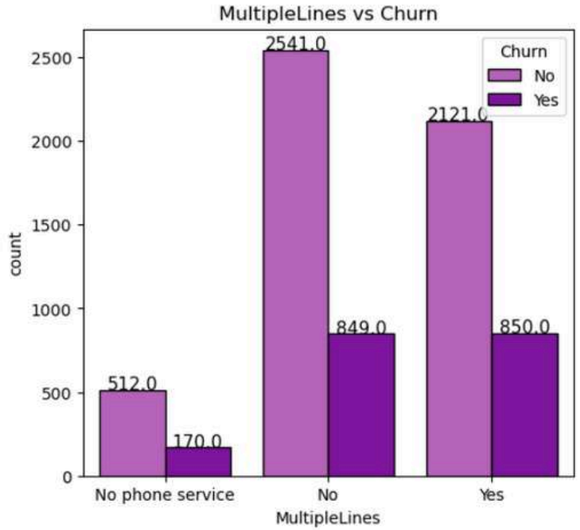


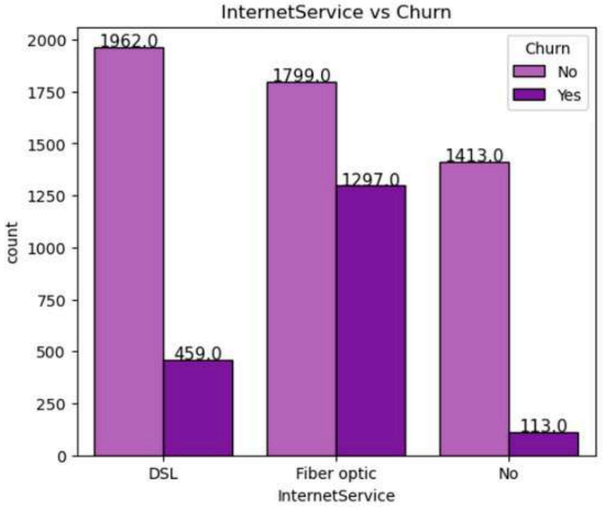


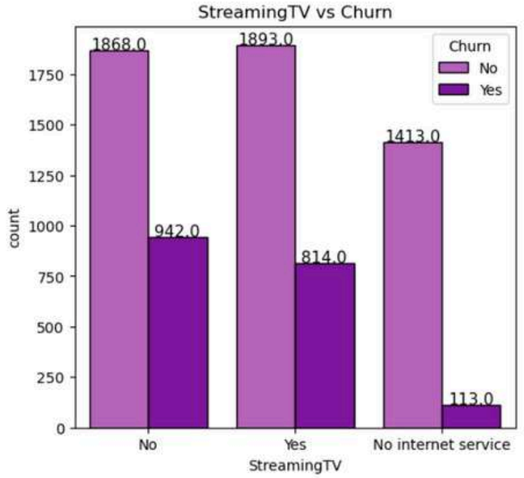


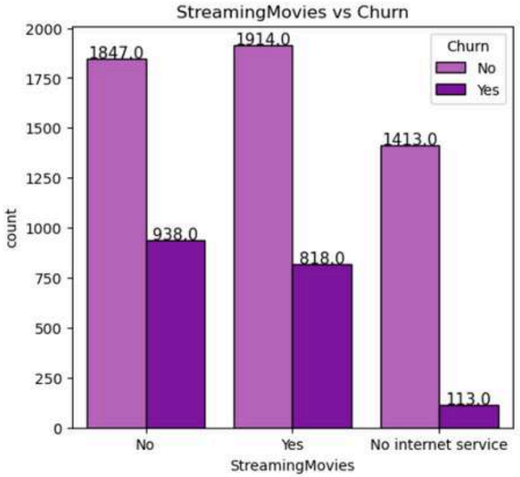
Analisis Eksplorasi Data



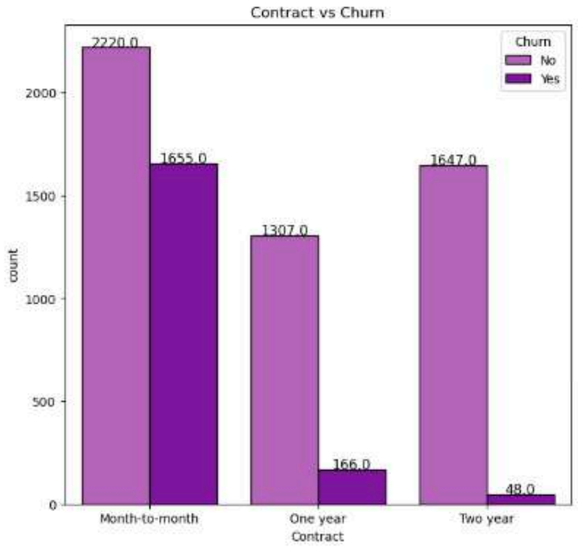


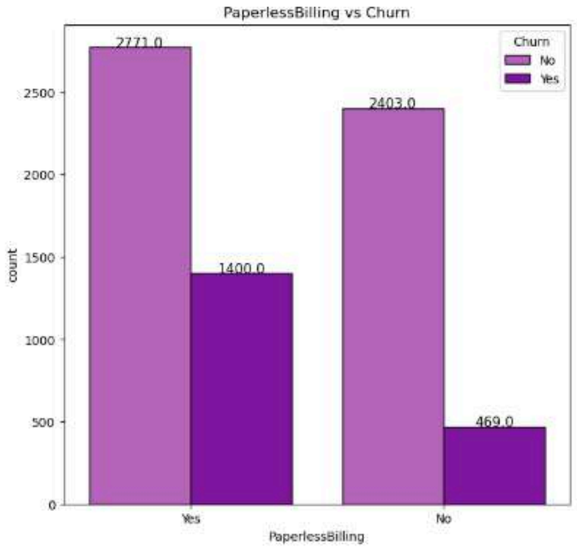


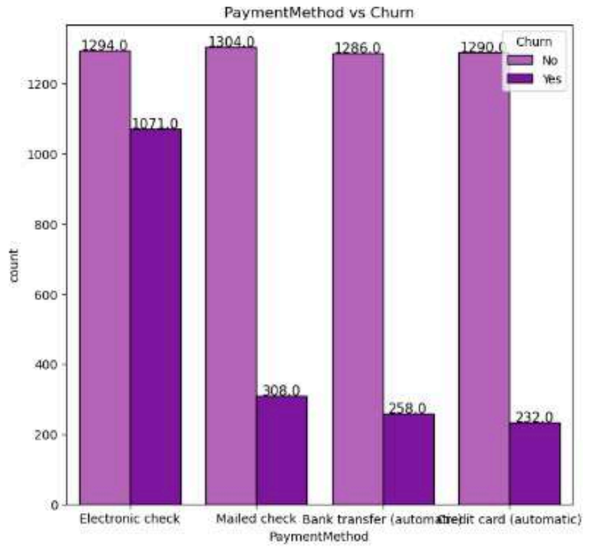




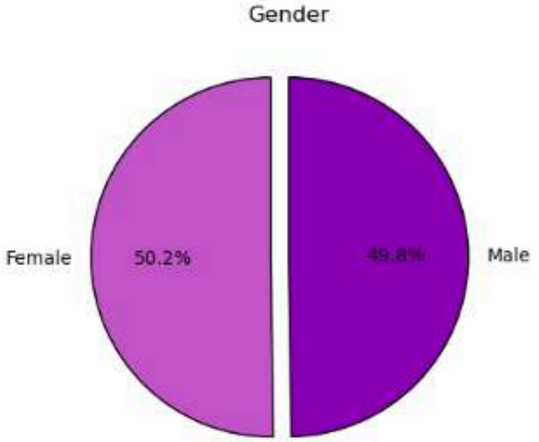
Analisis Eksplorasi Data

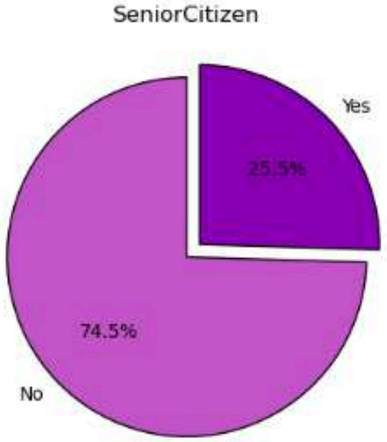


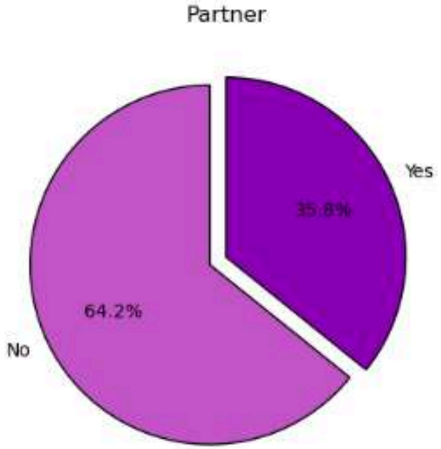


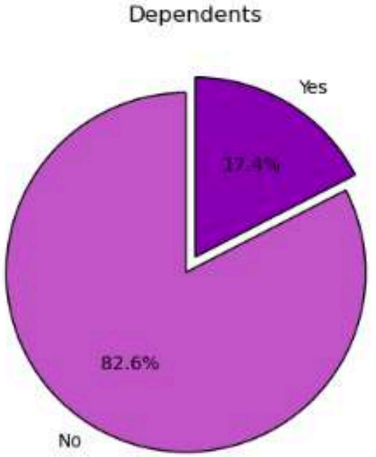


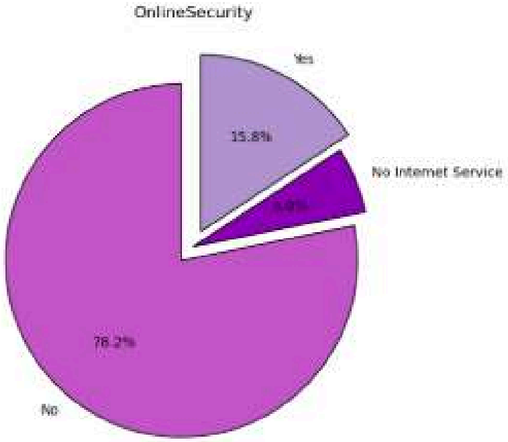
Analisis Eksplorasi Data

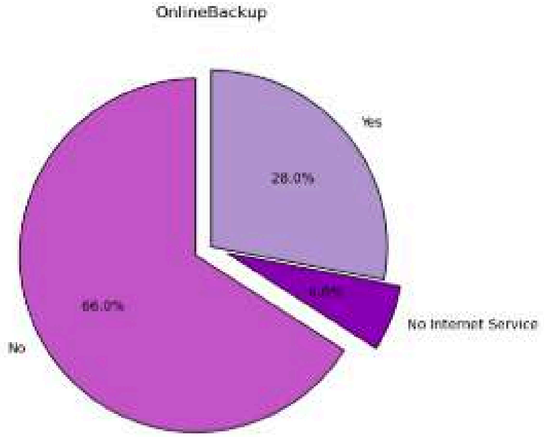


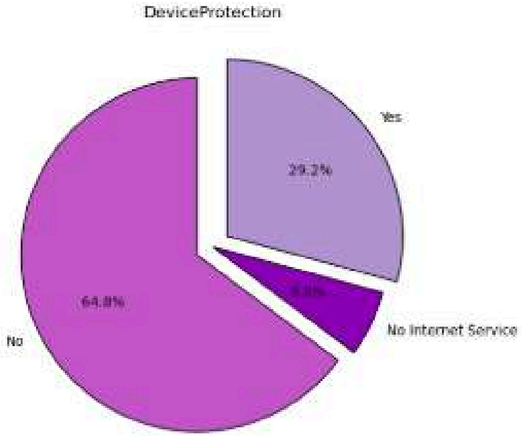


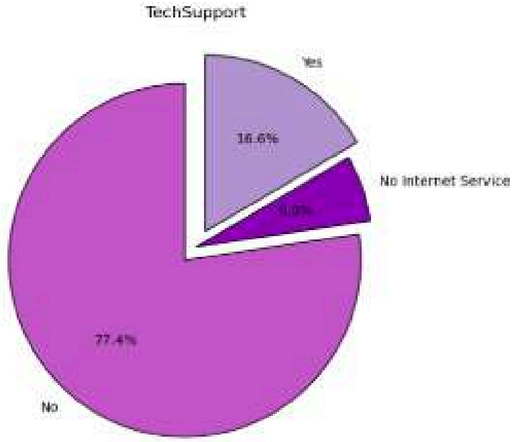




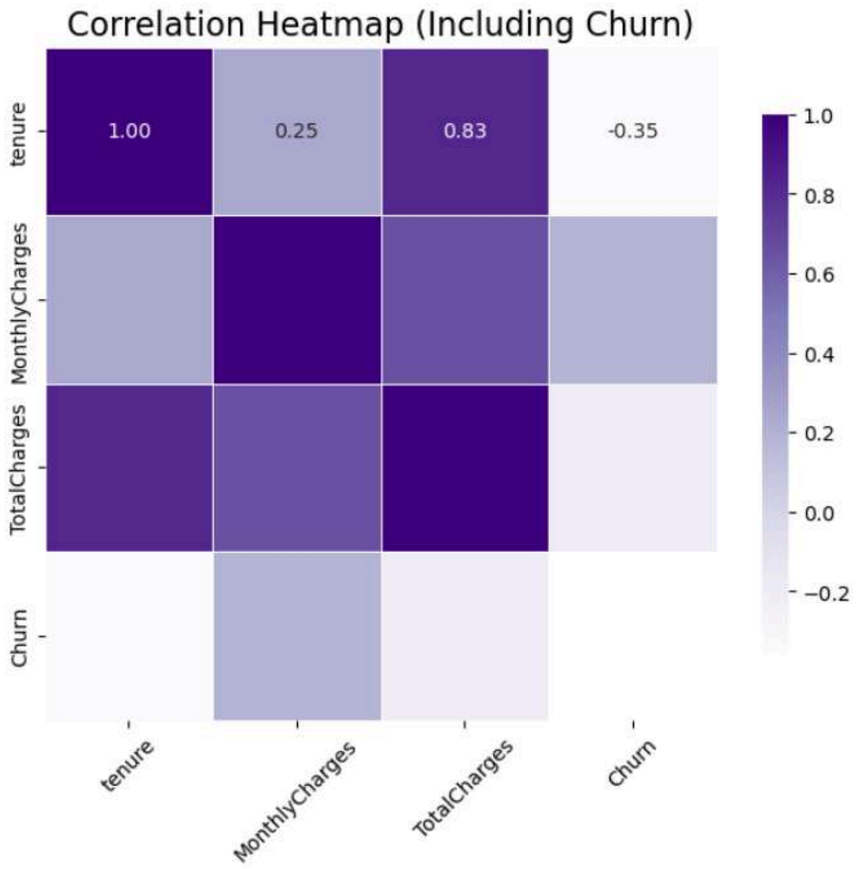








Feature Selection



Hipotesis :
H0: Variabel x dan variabel y tidak saling berkorelasi
H1: Variabel x dan y saling berkorelasi

Dengan tingkat signifikansi 5% (alpha = 0,05), maka:

Variabel yang memiliki hubungan signifikan dengan Churn:
['SeniorCitizen', 'Partner', 'Dependents', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod']

Variabel yang tidak memiliki hubungan signifikan dengan Churn:
['gender', 'PhoneService']

```
Korelasi Pearson antara tenure dan Churn:
Nilai Korelasi: -0.354
P-value: 9.438e-207
Kesimpulan: tenure memiliki hubungan signifikan dengan Churn.

Korelasi Pearson antara MonthlyCharges dan Churn:
Nilai Korelasi: 0.193
P-value: 6.761e-60
Kesimpulan: MonthlyCharges memiliki hubungan signifikan dengan Churn.

Korelasi Pearson antara TotalCharges dan Churn:
Nilai Korelasi: -0.199
P-value: 4.877e-64
Kesimpulan: TotalCharges memiliki hubungan signifikan dengan Churn.
```

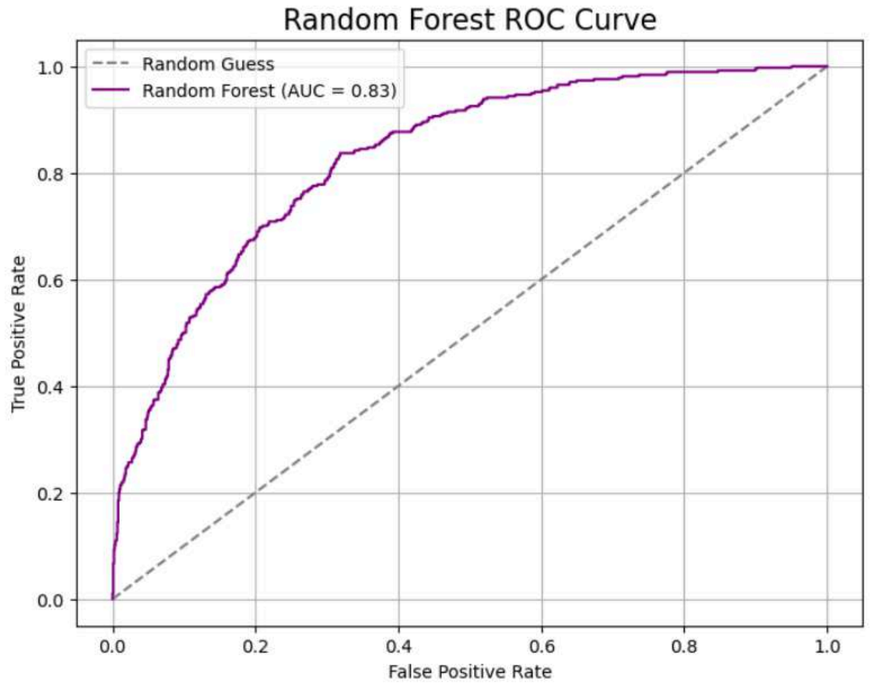
Korelasi Variabel Kategorik

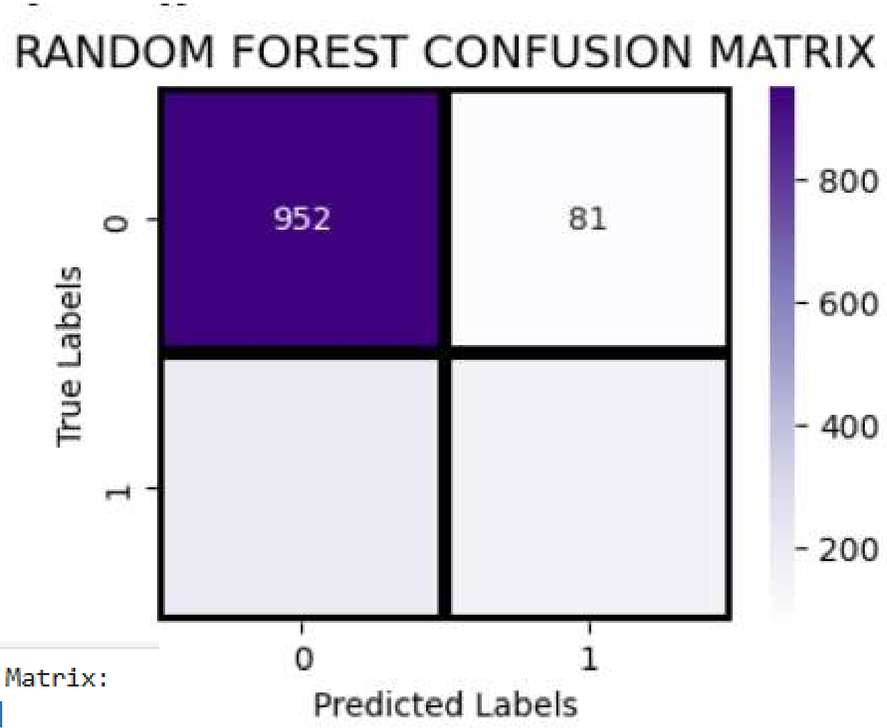
```python
from scipy import stats
import pandas as pd

# Daftar variabel kategorikal
categorical_variables = ['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService',
                         'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup',
                         'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies',
                         'Contract', 'PaperlessBilling', 'PaymentMethod']

# Variabel untuk menampung hasil
significant_vars = []  # Variabel yang memiliki hubungan signifikan dengan Churn
non_significant_vars = []  # Variabel yang tidak memiliki hubungan signifikan dengan Churn

# Melakukan uji Chi-Square
for var in categorical_variables:
    contingency_table = pd.crosstab(data['Churn'], data[var])
    stat, p, dof, expected = stats.chi2_contingency(contingency_table)

    # Menyusun hasil
    if p <= 0.05:
        significant_vars.append(var)
    else:
        non_significant_vars.append(var)

# Menampilkan hasil
print("Variabel yang memiliki hubungan signifikan dengan Churn:")
print(significant_vars)

print("\nVariabel yang tidak memiliki hubungan signifikan dengan Churn:")
print(non_significant_vars)
```

Random Forest

Akurasi : 0.7931769722814499

	precision	recall	f1-score	support
0	0.82	0.92	0.87	1033
1	0.67	0.44	0.53	374
accuracy			0.79	1407
macro avg	0.74	0.68	0.70	1407
weighted avg	0.78	0.79	0.78	1407



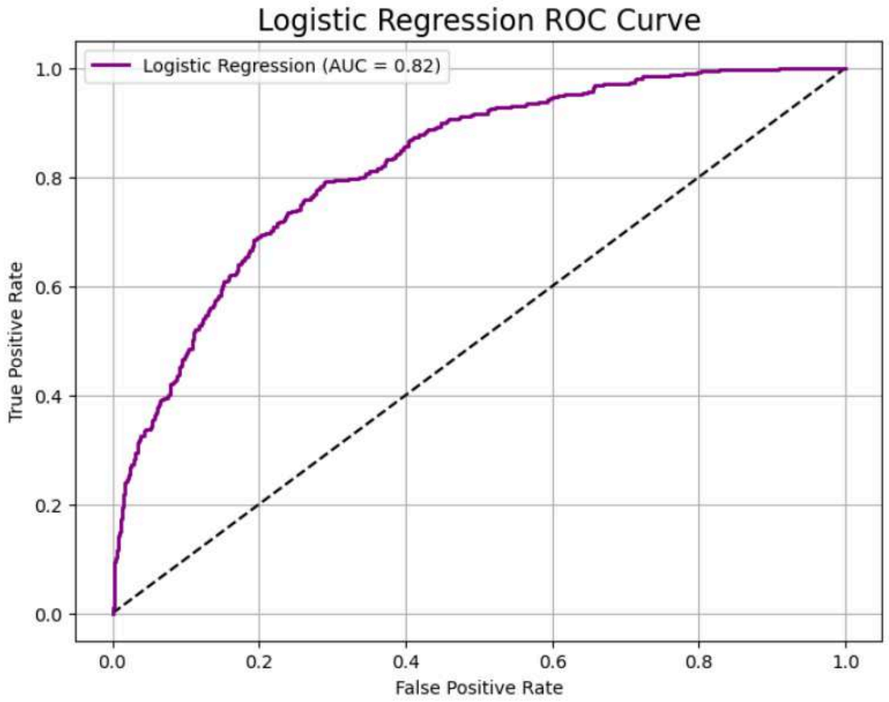


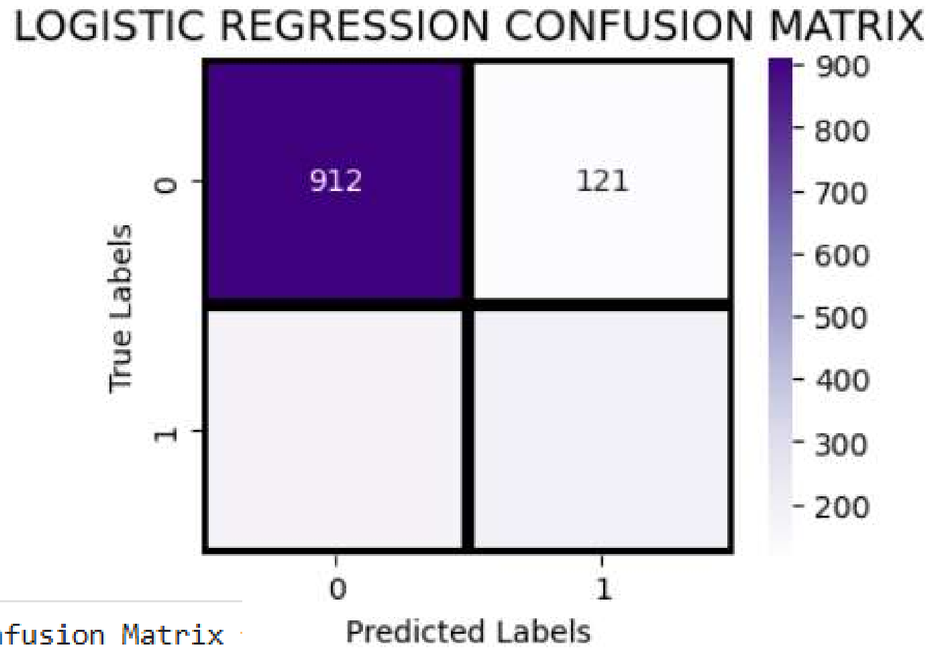
Confusion Matrix:
[[952 81]
[210 164]]

Logistic Regression

Akurasi : 0.7341862117981521

	precision	recall	f1-score	support
0	0.90	0.72	0.80	1033
1	0.50	0.78	0.61	374
accuracy			0.73	1407
macro avg	0.70	0.75	0.70	1407
weighted avg	0.79	0.73	0.75	1407





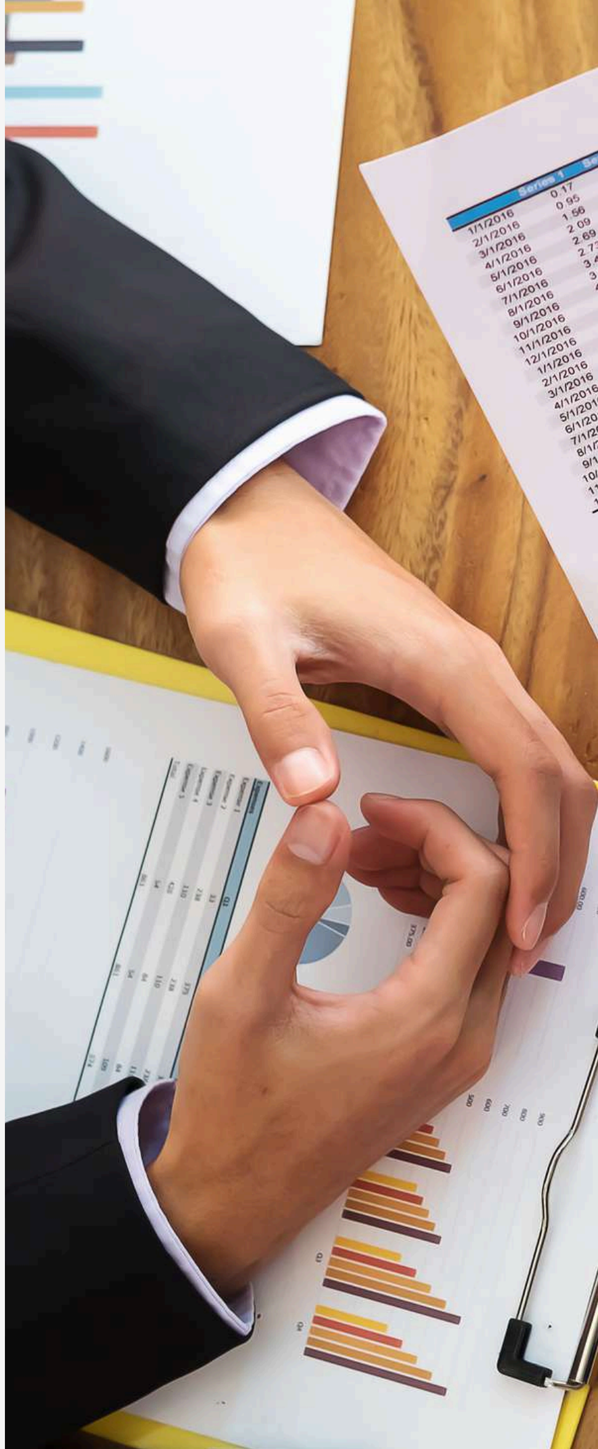

```
Confusion Matrix
[[912 121]
 [178 196]]
```

Repeated Holdout

Metric	Random Forest	Logistic Regression
Accuracy	0.7989 ± 0.0069	0.7994 ± 0.0081
F1-Score	0.536 ± 0.015	0.592 ± 0.015
Sensitivity	0.441 ± 0.019	0.540 ± 0.022
Specificity	0.927 ± 0.007	0.895 ± 0.011
AUC	0.84 ± 0.01	0.84 ± 0.01

k-Fold Cross Validation

Metric	Random Forest	Logistic Regression
Accuracy	0.8016 ± 0.0112	0.7989 ± 0.0118
F1-Score	0.558 ± 0.030	0.583 ± 0.025
Sensitivity	0.473 ± 0.032	0.530 ± 0.026
Specificity	0.920 ± 0.006	0.896 ± 0.011
AUC	0.84 ± 0.01	0.84 ± 0.01



Radom FOREST

Akurasi : 0.7931769722814499

	precision	recall	f1-score	support
0	0.82	0.92	0.87	1033
1	0.67	0.44	0.53	374
accuracy			0.79	1407
macro avg	0.74	0.68	0.70	1407
weighted avg	0.78	0.79	0.78	1407

Logistik REGRESSION

Akurasi : 0.7341862117981521

	precision	recall	f1-score	support
0	0.90	0.72	0.80	1033
1	0.50	0.78	0.61	374
accuracy			0.73	1407
macro avg	0.70	0.75	0.70	1407
weighted avg	0.79	0.73	0.75	1407


THANK YOU