Predict Clicked Ads kustomer Classification by using Machine Learning

Created by:
Ahya Ramdhanitasari
ahyaramdha02@gmail.com
www.linkedin.com/in/ahya-ramdhanitasari

"Bachelor degree of Science from Geography Majoring, University of Indonesia with spatial data (geographic information system) analysis specialization and want to expand my career to data science and business analyst. Equipped with relevant trainings or courses to support my career. Good to operate python data science, SQL query, and BI tools. Likes to explore, problem solving, and analysis activities. Detailed and organized person."

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

“Sebuah perusahaan di Indonesia ingin mengetahui efektifitas sebuah iklan yang mereka tayangkan, hal ini penting bagi perusahaan agar dapat mengetahui seberapa besar ketercapainnya iklan yang dipasarkan sehingga dapat menarik kustomers untuk melihat iklan.

Dengan mengolah data historical advertisement serta menemukan insight serta pola yang terjadi, maka dapat membantu perusahaan dalam menentukan target marketing, fokus case ini adalah membuat model machine learning classification yang berfungsi menentukan target kustomers yang tepat ”

Analisis Univariat

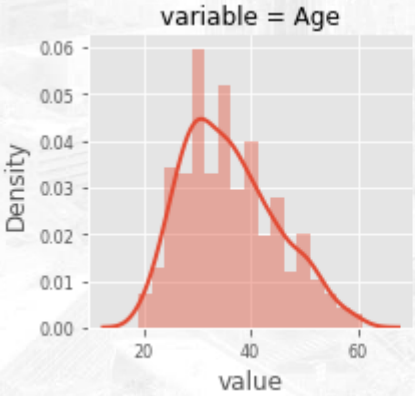
a. Unimodal

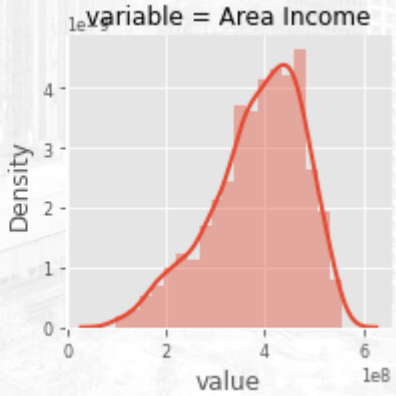
Kolom ‘Age’ memiliki distribusi positif, dimana mean>median>modus

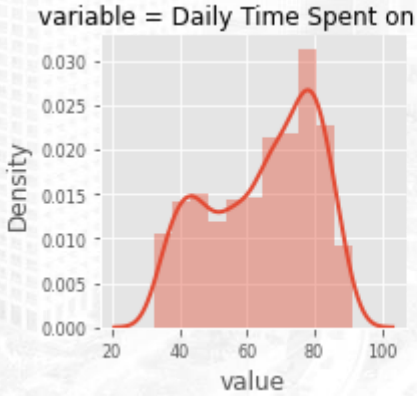
Kolom ‘Area Income’ memiliki distribusi negative, dimana mean<median<modus

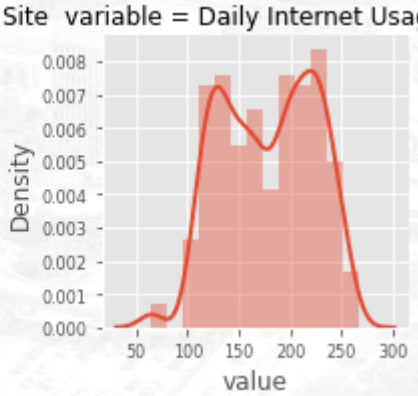
b. Bimodal

Kolom ‘Daily Time Spent on Site’ dan ‘Daily Internet Usage’ kemungkinan **memiliki 2 value dengan jumlah kemunculan yang hampir sama dan 2 bukit**. Hal ini dapat disebabkan oleh adanya kondisi dari variabel lain yang akan dibuktikan pada slide berikutnya.





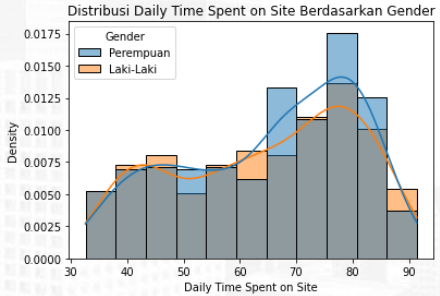


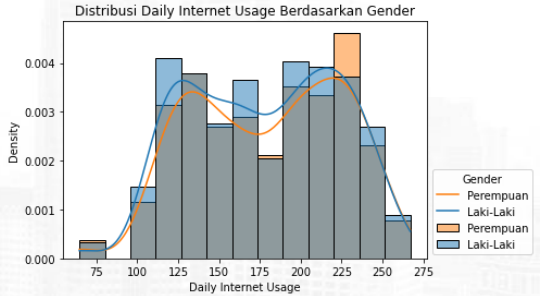


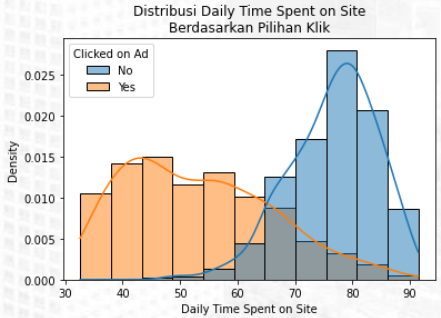
Tipe Kustomer dan Analisis Perilaku Kustomer dari Klik Iklan

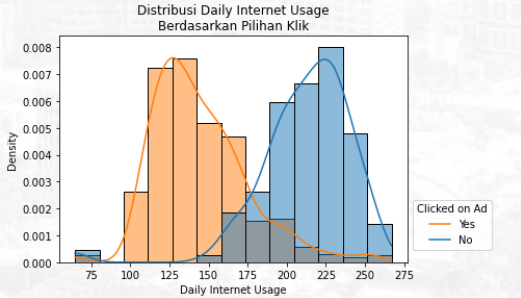
Setelah dilakukan pengecekan, ternyata **distribusi kolom ‘Daily Internet Usage’ dan ‘Daily Time Spent on Site’ mempengaruhi pilihan klik pada iklan.**

- Iklan cenderung di klik saat rata-rata lama waktu mengunjungi site lebih sedikit.
- Iklan juga cenderung di klik saat rata-rata konsumsi pemakaian internet cenderung sedikit.







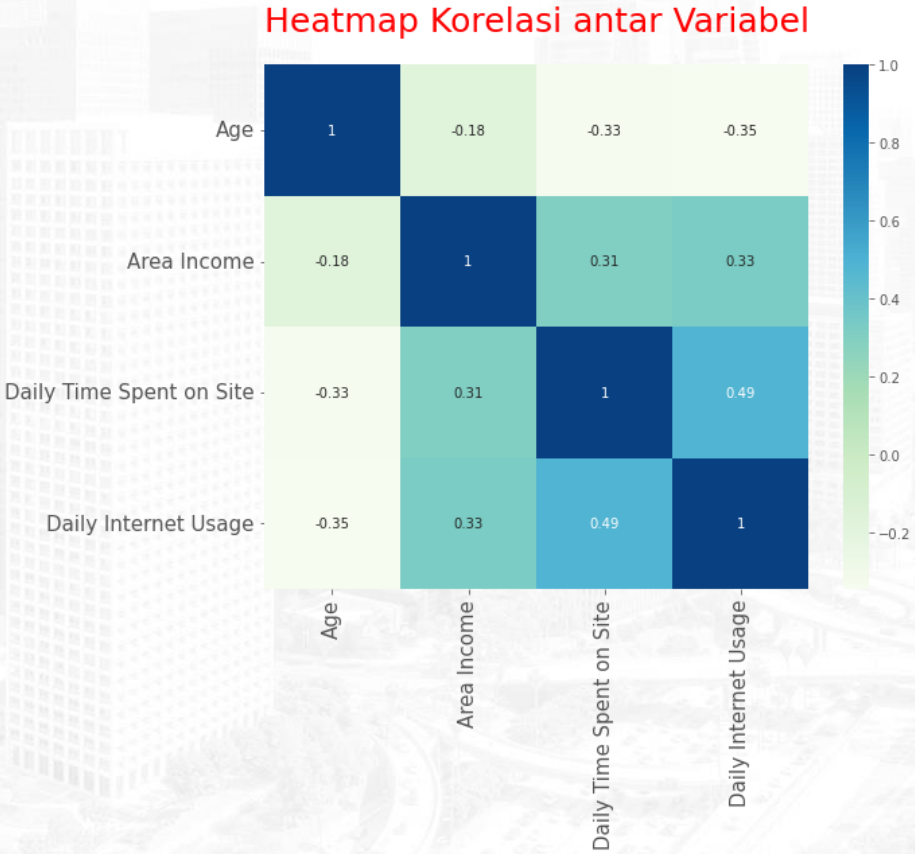


Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://drive.google.com/drive/folders/1HH5PKaFP3bBRGtL-uUFBDw0eer0r890M?usp=sharing

Analisis Multivariat

a. Korelasi

- **Kolom Age memiliki korelasi negative yang cukup besar dengan kolom ‘Daily Time Spent on Site’ dan kolom ‘Daily Internet Usage’**, sehingga dapat diketahui bahwa semakin tua kustomer, semakin sedikit waktu yang dipakai untuk bermain internet.
- **Kolom ‘Daily Time Spent on Site’ dan kolom ‘Daily Internet Usage’, memiliki korelasi positif dengan ‘Area Income’**. Sehingga semakin banyak pendapatan, semakin banyak waktu yang dihabiskan untuk bermain internet.



Tipe Kustomer dan Analisis Perilaku Kustomer dari Klik Iklan

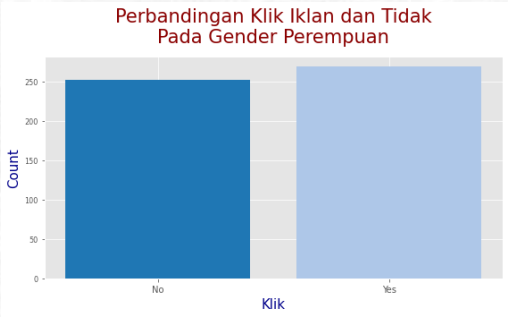
Analisis Multivariat

b. Hubungan antara pilihan klik iklan dengan gender

- **Kustomer dengan Gender Laki-laki lebih memilih untuk tidak klik iklan** di website.
- **Kustomer dengan Gender Perempuan lebih memilih untuk klik iklan** di website.

c. Hubungan antara pilihan klik iklan dengan umur

- **Semakin tua, kustomer semakin memilih untuk klik iklan** di website.



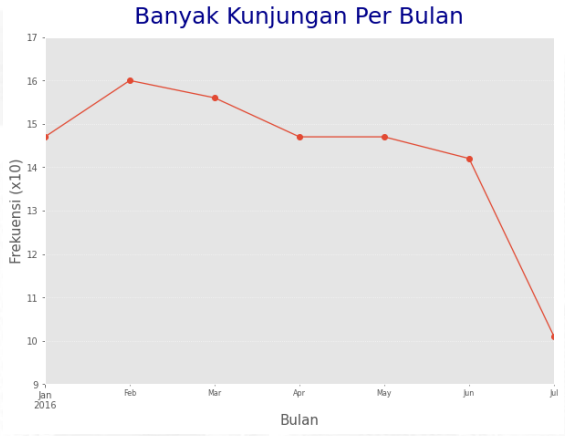
Analisis Tambahan

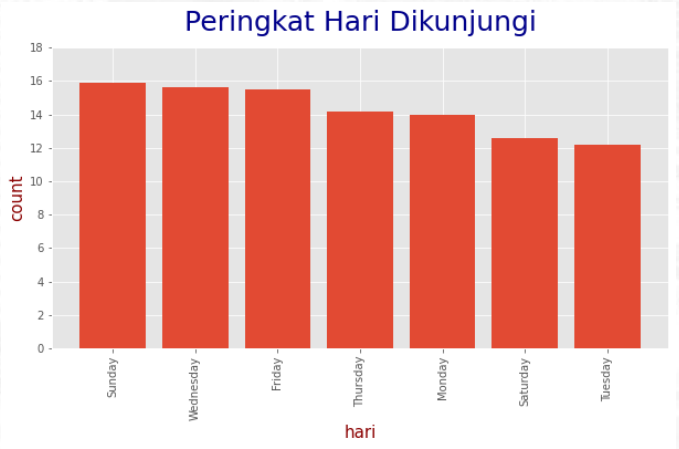
a. Banyak Kustomer berkunjung tiap bulan

Jumlah kustomer yang berkunjung tiap bulannya sejak bulan Januari 2016 hingga Juli 2016, lebih banyak **mengalami penurunan** dengan **kunjungan kustomer tertinggi pada bulan Februari**.

b. Peringkat hari dikunjungi kustomer

Jumlah kustomer paling banyak mengunjungi website pada hari minggu, sedangkan **hari selasa memiliki jumlah kustomer paling sedikit mengunjungi website.**

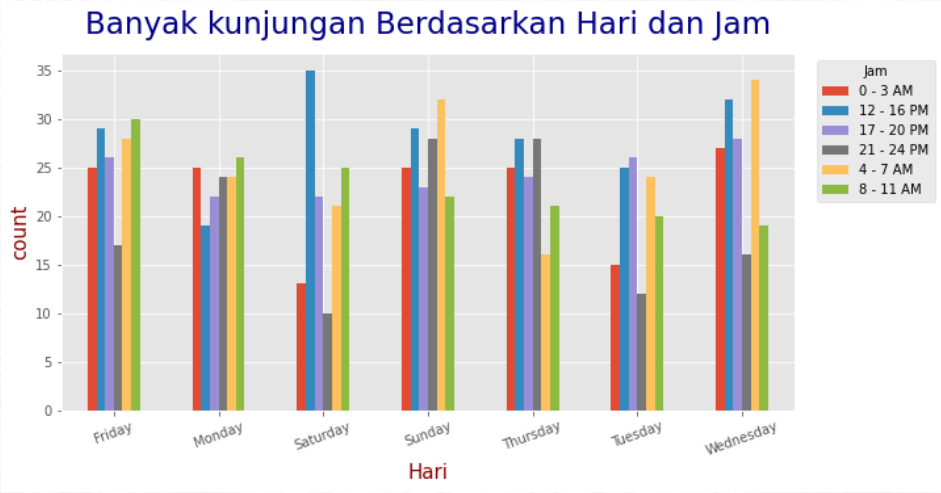




Analisis Tambahan

c. Banyak kunjungan berdasarkan waktu

- Senin : Pukul 8 – 11 AM
- **Selasa : Pukul 5 – 8 PM**
- Rabu : Pukul 4 – 7 AM
- Kamis : Pukul 0 – 4 PM & 9 – 12 PM
- Jumat : Pukul 8 – 11 AM
- Sabtu : Pukul 0 – 4 PM
- Minggu : Pukul 8 – 11 AM



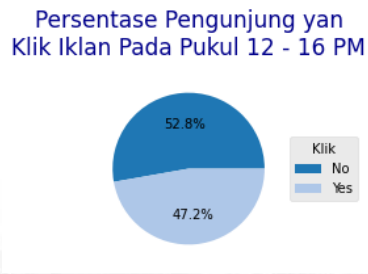
Tipe Kustomer dan Analisis Perilaku Kustomer dari Klik Iklan

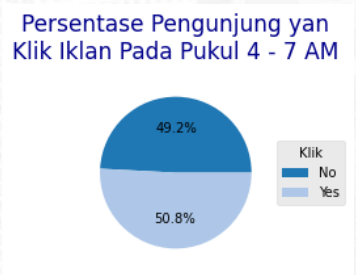
Analisis Tambahan

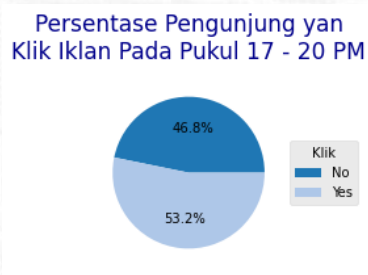
c. <u>Persentase kustomer klik iklan dan tidak menurut jam</u>

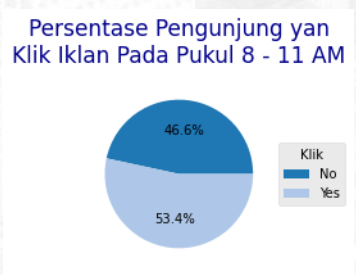
Selisih persentase kustomer lebih banyak klik iklan terjadi pada pukul:

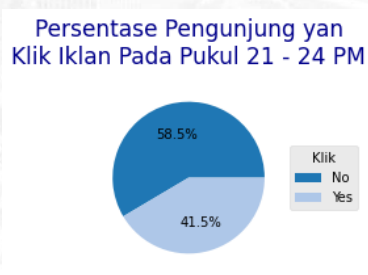
- 0 – 3 AM (cukup besar)
- 4 – 7 AM (kecil)
- **8 – 11 AM (besar)**
- **17 – 20 PM (besar)**











Untuk selengkapnya, dapat melihat jupyter notebook disini: https://drive.google.com/drive/folders/1HH5PKaFP3bBRGtL-uUFBDw0eer0r890M?usp=sharing

Data Cleaning & Preprocessing

Rakamin Academy

Fix Data Type	• Membetulkan tipe data yang memiliki kesalahan tipe data, contohnya kolom ‘Timespent’ memiliki tipe object yang seharusnya bertipe Datetime.
Fill Missing Value	• Mengisi kolom yang memiliki nilai NaN yang terdapat pada kolom ‘Daily time Spent on Site’ dan ‘Daily Internet Usage’ dengan nilai rata-rata dari kolom tersebut berdasarkan gender dan kolom ‘Area Income’ dengan nilai rata-rata dari kolom tersebut berdasarkan ‘city’.
Feature Extraction	• Membuat kolom baru dari kolom ‘Timestamp’ yang telah ada dengan menjadi kolom nama hari, bulan, tahun, dan jam.
Feature Encoding	• Label Encoding: ‘Agegroup’ • One-Hot Encoding : 'Gender', 'category', 'city', 'hari', 'monthtime', 'jambaru', dan 'Clicked on Ad'

Splitting data

Dilakukan splitting data dari library sklearn dengan jumlah test data sebanyak 243 baris.

```
:
from sklearn.model_selection import train_test_split

total_data = np.arange(1000)
num_test_data = 243
train, test= train_test_split(total_data, test_size=num_test_data, random_state=42)

print(f"Train data size: {len(train)}")
print(f"Test data size: {len(test)}")
```

```
Train data size: 757
Test data size: 243
```

```
# Pisahkan fitur (X) dan label (y)
X = dfnew3.drop('Clicked on Ad_Yes', axis=1)
y = dfnew3['Clicked on Ad_Yes']
X.head(2)
```

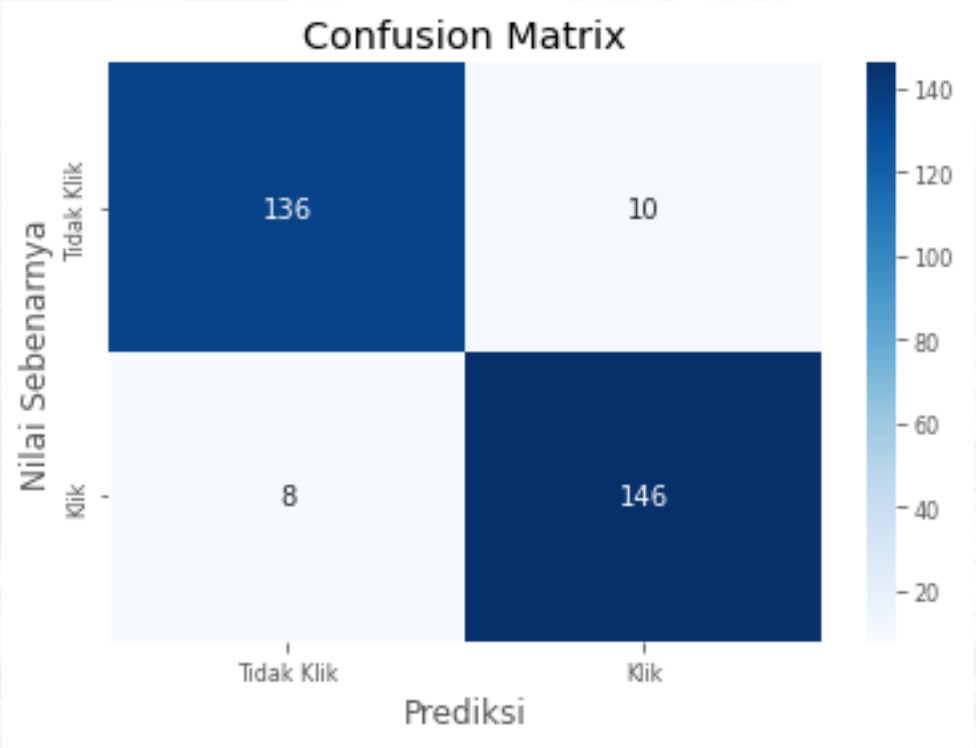
```
# split to train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Data Modeling (Non Normalisasi)

Random Forest

(RMSE = 0,24)

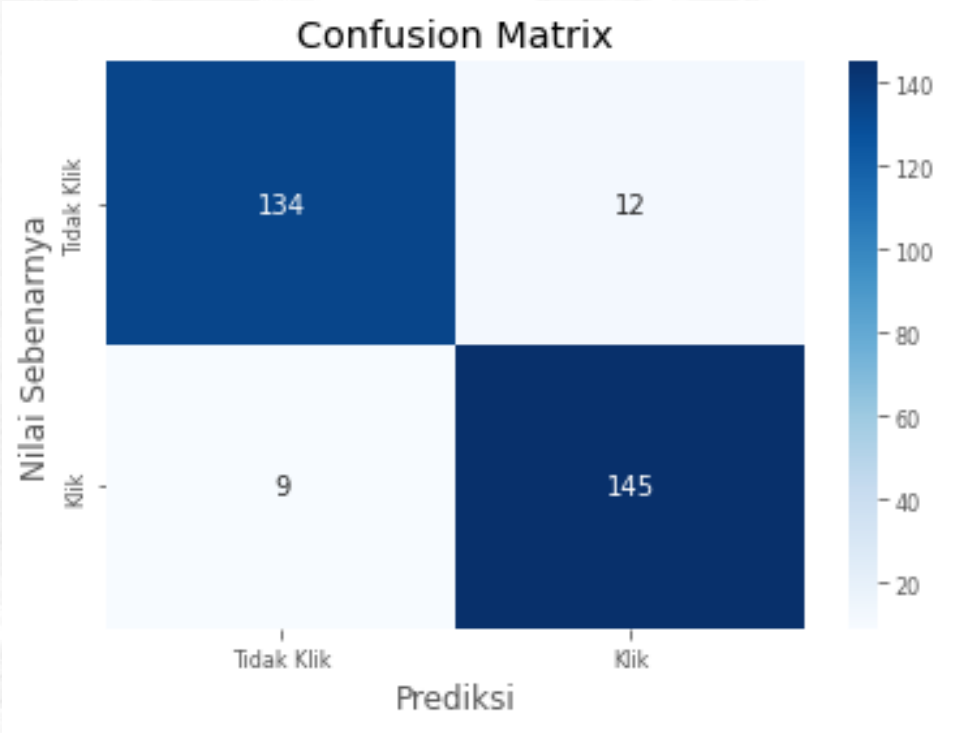
Metric	Train	Test
Accuracy	1	0,94
Precision	1	0,94
Recall	1	0,95
AUC	1	0,98
AUC (5-fold cv)	1	0,9904



Decision Tree

(RMSE = 0,26)

Metric	Train	Test
Accuracy	1	0,93
Precision	1	0,92
Recall	1	0,94
AUC	1	0,93
AUC (5-fold cv)	1	0,93007

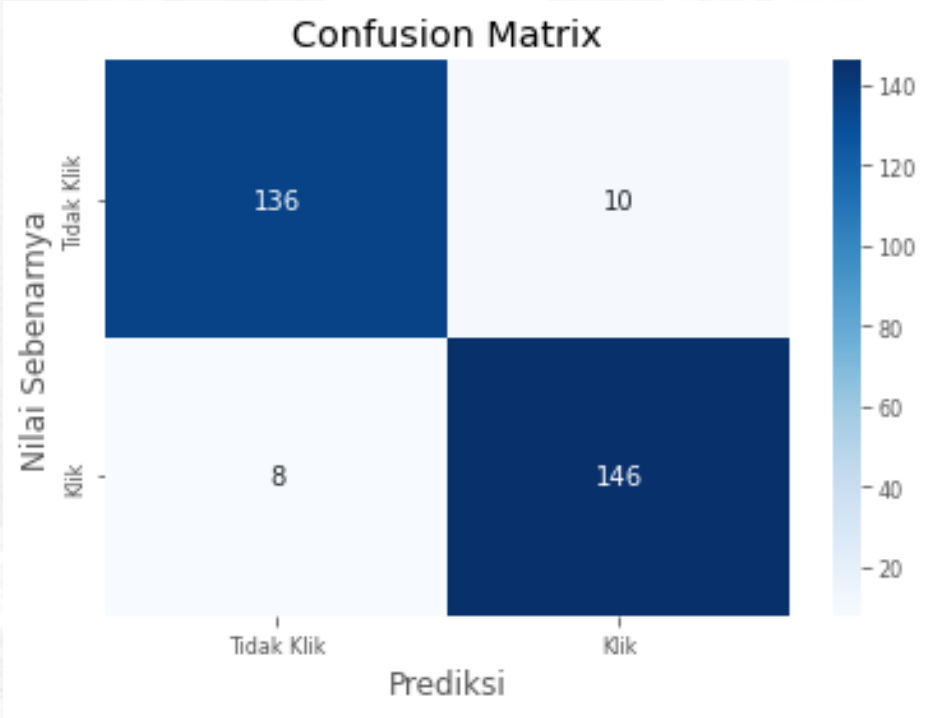


Data Modeling (Normalisasi)

Random Forest

(RMSE = 0,24)

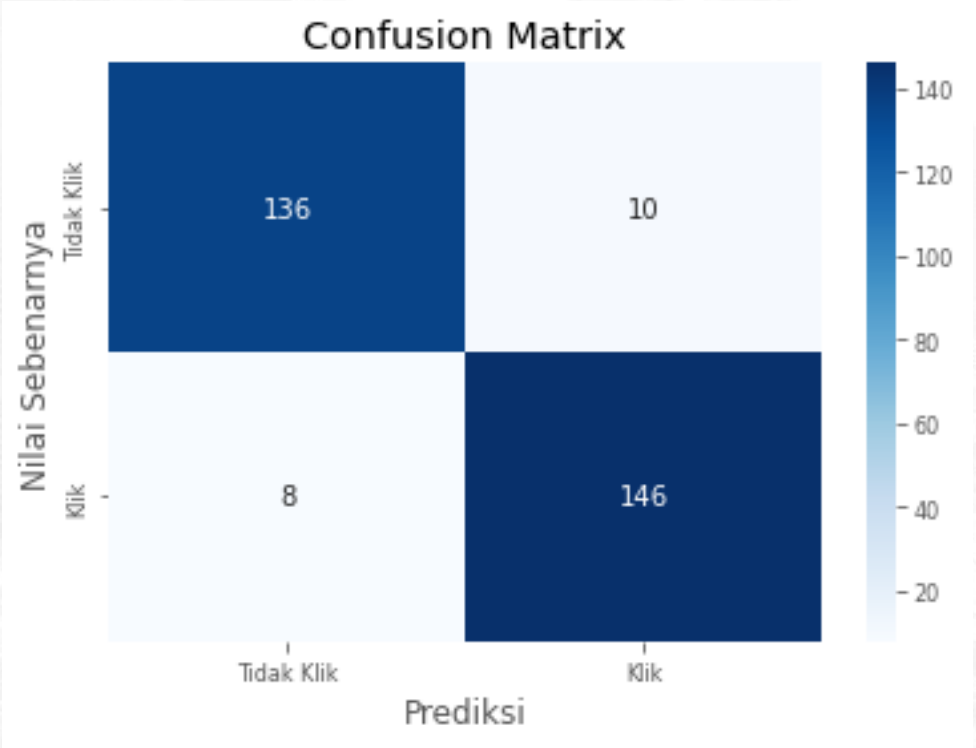
Metric	Train	Test
Accuracy	1	0,94
Precision	1	0,94
Recall	1	0,95
AUC	1	0,98
AUC (5-fold cv)	1	0,9903



Decision Tree

(RMSE = 0,26)

Metric	Train	Test
Accuracy	1	0,93
Precision	1	0,92
Recall	1	0,94
AUC	1	0,93
AUC (5-fold cv)	1	0,93007

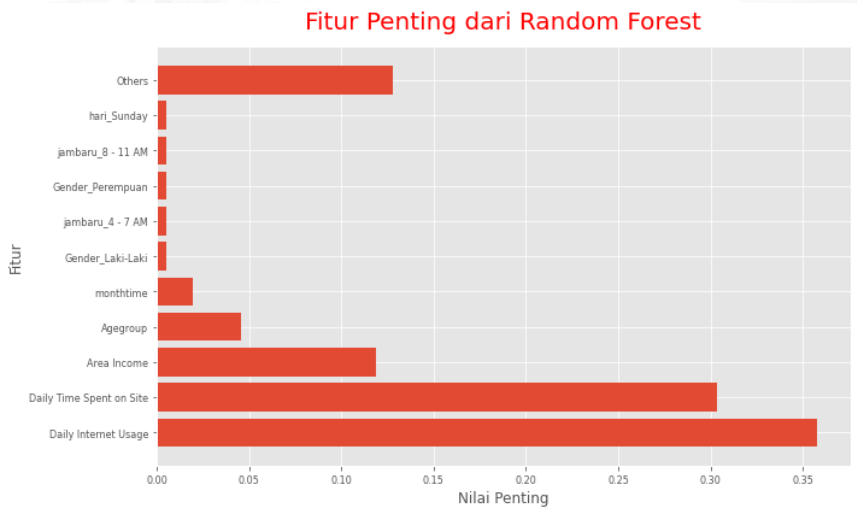


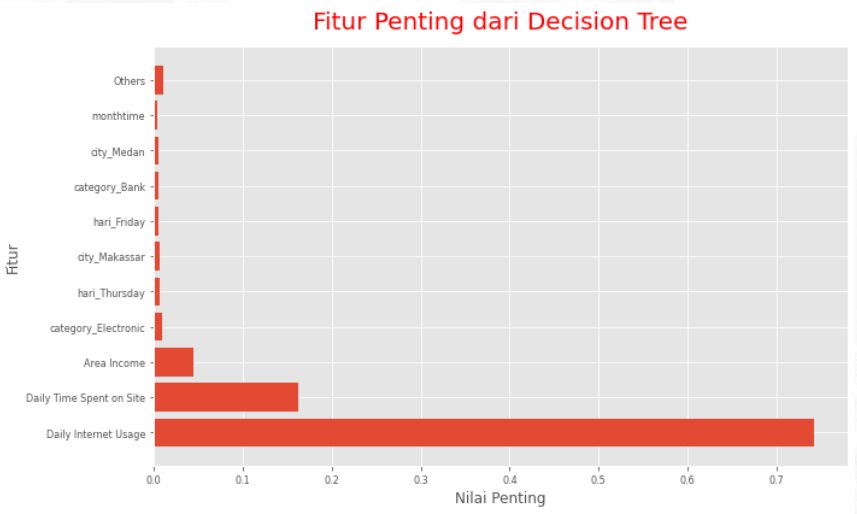
Model

Nilai dari *confusion matrix* sudah cukup relevan. Nilai RMSE dari model *random forest* tanpa dan dengan normalisasi lebih baik dari pada model *decision tree* dan performa model *random forest* terlihat lebih stabil daripada model *decision tree*. Performa model tanpa normalisasi dan dengan normalisasi cenderung sama atau hamper sama, hanya saja dari nilai AUC (5-fold cv) pada model *random forest* tanpa normalisasi memiliki perbedaan 0,0001 lebih besar daripada model *random forest* dengan normalisasi yang berarti performa model random forest tanpa normalisasi lebih baik/stabil. Dengan ini dapat disimpulkan bahwa **model *random forest* tanpa normalisasi adalah model terbaik** yang dapat digunakan jika dibandingkan dengan model *decision tree* tanpa atau dengan normalisasi.

https://drive.google.com/drive/folders/1g1wXwxP4SEnbX3wZ6hJV5fMcVPM0klj8?usp=sharing

Feature Importance





Top 3 Fitur penting dari kedua model memiliki kesamaan dan selanjutnya mengalami perubahan fitur penting, hanya saja tingkat kepentingan dari ketiga fitur tersebut pada model *decision tree* sangat tinggi dan cenderung memiliki perbedaan nilai yang sangat signifikan dibandingkan dengan model *random forest*, sehingga variabel atau fitur lain memiliki nilai fitur penting yang sangat kecil.

Rekomendasi Bisnis

Setelah dilakukan pengecekan dengan EDA, *feature importance* dari model *random forest* lah yang cukup relevan. Untuk meningkakan *revenue* dan meminimalisir biaya, lebih baik dilakukan pengiklanan pada customer terpilih saja, yaitu dengan ciri sebagai berikut:

- Rata-rata lama waktu mengunjungi *site* toko lebih sedikit,
- Rata-rata konsumsi pemakaian internet yang cenderung lebih sedikit,
- Kustomer yang berusia lebih tua,
- Kustomer dengan jumlah pendapatan yang lebih sedikit, dan
- Kustomer dengan Gender Perempuan.

Selain itu, penyebaran iklan atau pengiklanan dilakukan pada hari selasa, pukul 5 – 8 PM agar lebih banyak customer yang melihat atau klik.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Simulasi Bisnis

Jika …

Cost pengiklanan ke 300 target adalah 30000 Rupiah

Rata-rata Revenue per konversi adalah 1500 Rupiah

Maka …

Simulasi	Tanpa Machine Learning	Dengan Machine Learning
Conversion rate	15.4%	15.6%
Total Conversion	46.2	46.8
Total Revenue	Rp 69300	Rp 70200
Total Profit	Rp 39300	Rp 40200

Thank You!
Let’s Connect!

Portfolio Link : https://bit.ly/portoahya
Github project : https://github.com/ahyaramdha/Predict_click_ads