

# Predict Clicked Ads Customer Classification by using Machine Learning

Supported by:  
**Rakamin Academy**  
Career Acceleration School  
[www.rakamin.com](http://www.rakamin.com)



**Created by:**

**Arieska Restu**

[arieskarestu02@gmail.com](mailto:arieskarestu02@gmail.com)

[linkedin.com/in/arieskarestu](https://linkedin.com/in/arieskarestu)

I am an Assistant Lecturer who has experience in the field of Data Science with a background in Informatics. Experienced in Data Analysis, Data Mining, and Machine Learning projects. Also experienced in extracting primary and secondary data, as well as developing and maintaining databases. Able to conduct in-depth data analysis to identify trends that are relevant to companies and clients, and proficient in creating analysis reports. I also have expertise in programming languages and data visualization.

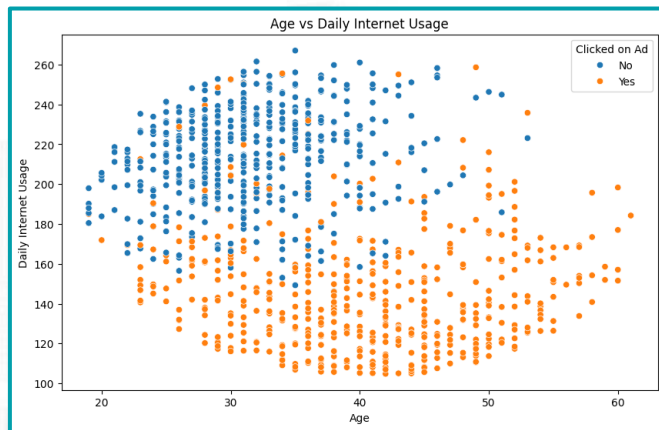
“Sebuah perusahaan di Indonesia ingin mengetahui efektifitas sebuah iklan yang mereka tayangkan, hal ini penting bagi perusahaan agar dapat mengetahui seberapa besar ketercapainnya iklan yang dipasarkan sehingga dapat menarik customers untuk melihat iklan.

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 customers yang tepat ”

- Proses EDA yang dilakukan terdiri dari tiga tahap, yakni mulai dari Quick EDA, Statistical Summaries, Univariate Analysis, Bivariate Analysis, dan Multivariate Analysis.
- Proses-proses yang dilakukan pada tahap Quick EDA yaitu pengecekan jumlah baris dan kolom, pengecekan informasi kolom dataset, pengecekan data yang hilang, dan pengecekan data yang duplikat.
- Proses-proses yang dilakukan pada tahap Statistical Summaries yaitu melihat ringkasan statistik baik fitur numerical maupun categorical.
- Pada tahap Univariate Analysis dilakukan analisis dengan menggunakan visualisasi dari persebaran data untuk setiap kolom, baik kolom numerical maupun kolom categorical.
- Pada tahap Univariate Analysis dilakukan analisis dengan menggunakan visualisasi untuk melihat hubungan antara kolom Age, Daily Internet Usage, dan Daily Time Spent on Site.
- Pada tahap Multivariate Analysis dilakukan analisis dengan menggunakan visualisasi dari correlation matrix untuk setiap fitur.

- Dari proses EDA yang dilakukan didapatkan informasi-informasi dalam dataset, yakni sebagai berikut.
  - Terdapat fitur yang tidak memiliki nama.
  - Terdapat missing value pada kolom 'Daily Time Spent on Site', 'Area Income', 'Daily Internet Usage', dan 'Male'.
  - Dalam dataset tidak ada data yang duplicate.
  - Mayoritas pengguna berada di situs antara 51.27 dan 78.46 menit per hari.
  - Pengguna situs mayoritas berusia antara 29 dan 42 tahun. Ini mengindikasikan bahwa situs tersebut cenderung menarik perhatian kelompok usia dewasa muda hingga dewasa pertengahan.
  - Pengguna berasal dari wilayah dengan tingkat ekonomi yang bervariasi, tetapi mayoritas berada di wilayah dengan pendapatan antara 328 juta hingga 458 juta.
  - Sebagian besar pengguna menggunakan internet antara 138.71 hingga 218.79 menit per hari. Ini mengindikasikan bahwa kebanyakan pengguna cukup aktif di internet dengan penggunaan internet harian yang tinggi.
  - Mayoritas pengguna adalah perempuan, sekitar 52% dari total data. Ini menunjukkan bahwa perempuan mungkin lebih banyak berinteraksi dengan situs atau iklan, yang bisa jadi pertimbangan dalam strategi marketing perusahaan.
  - Kota Surabaya dan provinsi DKI Jakarta merupakan asal pengguna yang paling banyak.
  - Kategori iklan Otomotif adalah yang paling sering dilihat atau diklik oleh pengguna.

# Age vs Daily Internet Usage

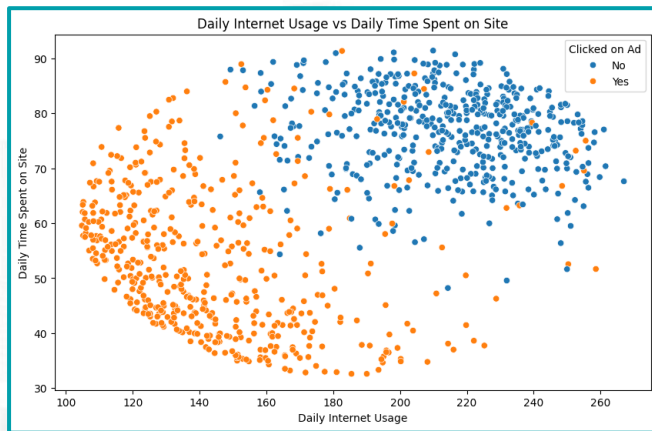


Berdasarkan grafik tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

- Titik-titik data tersebar cukup merata tanpa membentuk pola garis atau kurva yang jelas. Ini mengindikasikan bahwa tidak ada hubungan linier yang kuat antara usia dan penggunaan internet harian dalam menentukan apakah seseorang akan mengklik iklan atau tidak.
- Kedua kelompok pengguna (yang mengklik dan tidak mengklik iklan) memiliki rentang usia dan penggunaan internet harian yang sangat mirip. Terdapat banyak tumpang tindih antara kedua kelompok ini.
- Baik pengguna yang mengklik maupun tidak mengklik iklan cenderung memiliki distribusi usia dan penggunaan internet harian yang relatif seragam dalam rentang tertentu.



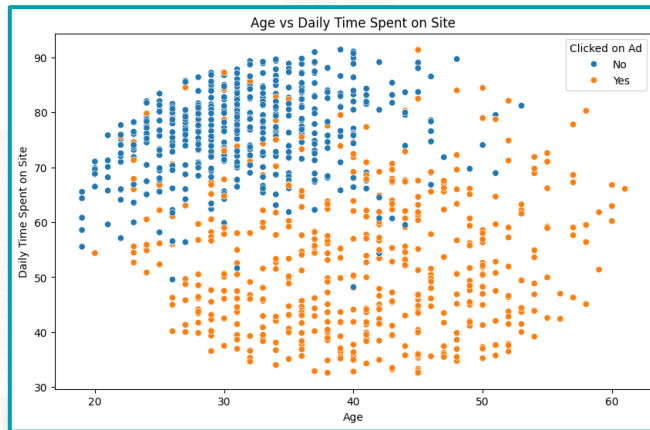
# Daily Internet Usage vs Daily Time Spent on Site



Berdasarkan grafik tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

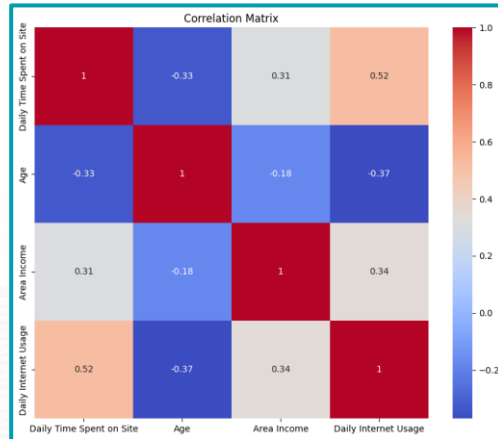
- Titik-titik data tersebar cukup merata tanpa membentuk pola garis atau kurva yang jelas. Ini mengindikasikan bahwa tidak ada hubungan linier yang kuat antara penggunaan internet harian dan waktu yang dihabiskan di situs dalam menentukan apakah seseorang akan mengklik iklan atau tidak.
- Kedua kelompok pengguna (yang mengklik dan tidak mengklik iklan) memiliki rentang penggunaan internet harian dan waktu di situs yang sangat mirip. Terdapat banyak tumpang tindih antara kedua kelompok ini.
- Baik pengguna yang mengklik maupun tidak mengklik iklan cenderung memiliki distribusi penggunaan internet harian dan waktu di situs yang relatif seragam dalam rentang tertentu.

# Age vs Daily Time Spent on Site



Berdasarkan grafik tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

- Titik-titik data tersebar cukup merata tanpa membentuk pola garis atau kurva yang jelas. Ini mengindikasikan bahwa tidak ada hubungan linier yang kuat antara usia dan waktu yang dihabiskan di situs dalam menentukan apakah seseorang akan mengklik iklan atau tidak.
- Kedua kelompok pengguna (yang mengklik dan tidak mengklik iklan) memiliki rentang usia dan waktu di situs yang sangat mirip. Terdapat banyak tumpang tindih antara kedua kelompok ini.
- Baik pengguna yang mengklik maupun tidak mengklik iklan cenderung memiliki distribusi usia dan waktu di situs yang relatif seragam dalam rentang tertentu.



Berdasarkan correlation matrix tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

- Daily Time Spent on Site dan Daily Internet Usage memiliki korelasi positif yang kuat (0.52). Ini menunjukkan bahwa semakin banyak waktu yang dihabiskan seseorang di situs, semakin tinggi pula penggunaan internet harian mereka.
- Age dan Daily Internet Usage memiliki korelasi negatif (-0.37). Ini berarti bahwa seiring bertambahnya usia, cenderung ada penurunan dalam penggunaan internet harian. Ini mungkin menunjukkan perubahan kebiasaan penggunaan internet seiring bertambahnya usia.
- Age dan Daily Time Spent on Site juga memiliki korelasi negatif (-0.33). Ini menunjukkan tren yang serupa, yaitu semakin tua seseorang, semakin sedikit waktu yang mereka habiskan di situs tertentu.
- Area Income memiliki korelasi yang relatif lemah dengan fitur lainnya. Ini menunjukkan bahwa pendapatan tidak memiliki hubungan yang kuat dengan fitur-fitur numerical lainnya.



- Berdasarkan hasil dari tahap Exploratory Data Analysis, Terdapat missing value pada kolom 'Daily Time Spent on Site', 'Area Income', 'Daily Internet Usage', dan 'Male'. Selain itu, tidak terdapat data yang duplicate.
- Missing value pada fitur Daily Time Spent on Site, Area Income, Daily Internet Usage akan diisi menggunakan mean. Sedangkan pada fitur Male akan dihapus data yang missing value.

Before

Daily Time Spent on Site	Age	Area Income
59.05	57	NaN
69.62	20	NaN
69.90	43	NaN
80.46	27	NaN
72.04	22	NaN
51.68	49	NaN
66.17	33	NaN

After

Daily Time Spent on Site	Age	Area Income
59.05	57	384938568.13
69.62	20	384938568.13
69.90	43	384938568.13
80.46	27	384938568.13
72.04	22	384938568.13
51.68	49	384938568.13
66.17	33	384938568.13

- Pada tahap data preprocessing, dilakukan beberapa proses yakni proses drop features, feature encoding, dan split data feature, dan extraction datetime.
- Untuk proses drop features dilakukan untuk menghapus fitur-fitur yang tidak diperlukan seperti 'Unnamed: 0'.
- Pada proses feature encoding dilakukan dengan metode one-hot encoding menggunakan get dummy.
- Lalu untuk proses split data feature dilakukan dengan cara memisahkan attribute target dengan feature. Attribute targetnya yakni 'Clicked on Ad'.

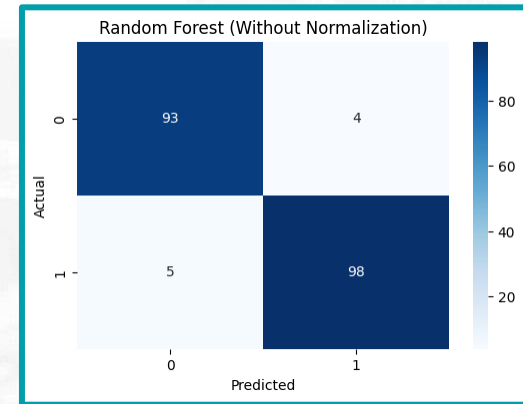
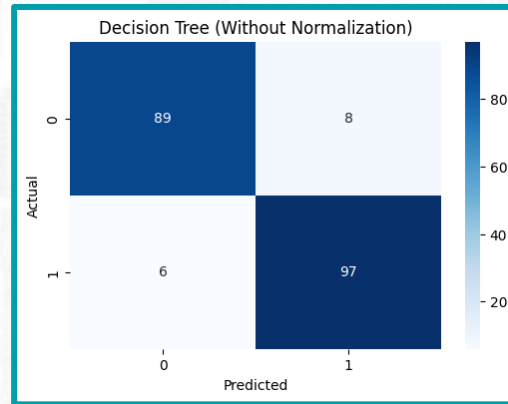
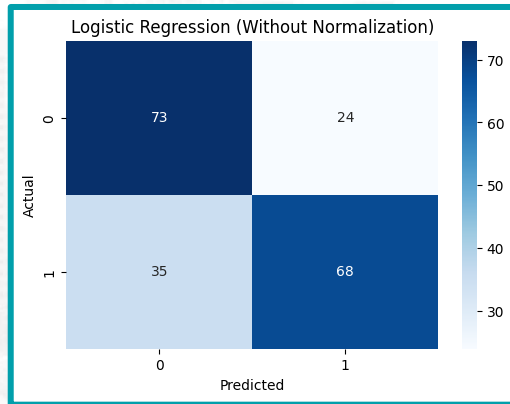
- Pada proses extraction datetime dilakukan proses ekstraksi dari fitur 'Timestamp' menjadi fitur 'Year', 'Month', 'Week', 'Day'.
- Lalu setelah diekstraksi, maka fitur 'Timestamp' akan dihapus.

Hasil extraction datetime

Timestamp	Year	Month	Week	Day
2016-04-18 11:23:00	2016	4	16	18
2016-04-23 03:46:00	2016	4	16	23
2016-07-13 21:31:00	2016	7	28	13
2016-01-26 02:47:00	2016	1	4	26
2016-01-06 13:20:00	2016	1	1	6
2016-02-26 09:18:00	2016	2	8	26
2016-06-18 05:17:00	2016	6	24	18

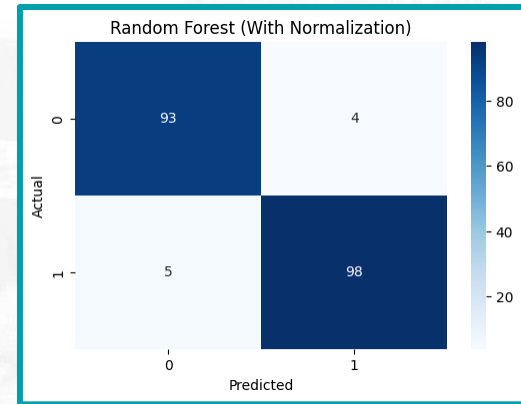
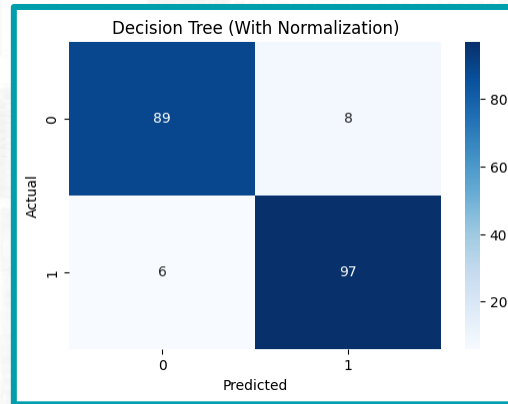
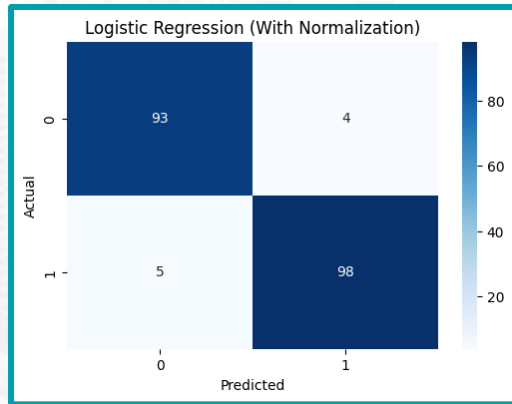
- Pada proses data modeling, akan dilakukan 2 eksperimen untuk model machine learning. Dimana eksperimen 1 dengan menggunakan data tanpa normalisasi, sedangkan eksperimen 2 akan memakai normalisasi.
- Model machine learning yang akan digunakan yaitu Logistic Regression, Decision Tree, dan Random Forest.
- Untuk proses normalisasi akan menggunakan metode min-max normalization.

- Berikut merupakan hasil confusion matrix dari model-model pada eksperimen 1.



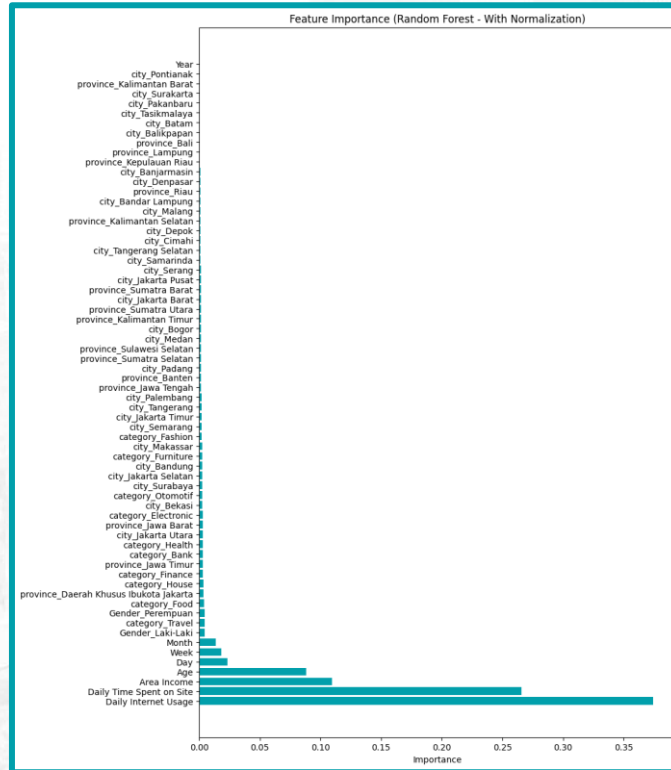


- Berikut merupakan hasil confusion matrix dari model-model pada eksperimen 2.



- Akurasi tertinggi didapatkan oleh model Random Forest, baik dengan normalisasi maupun tanpa normalisasi. Dimana akurasinya yakni sebesar 95.5%.
- Model Random Forest ini menunjukkan performa yang sangat baik dengan akurasi tinggi. Ini berarti model jarang melakukan kesalahan baik dalam mendeteksi klik maupun tidak klik.
- Dengan precision 96.1% dan recall 95.1%, model Random Forest ini memiliki keseimbangan yang baik. Artinya, ketika model memprediksi pengguna akan mengklik, prediksi tersebut sangat andal (tinggi precision). Pada saat yang sama, model juga berhasil menangkap sebagian besar dari semua klik yang benar-benar terjadi (tinggi recall).

- Berikut merupakan hasil feature importance untuk model Random Forest pada eksperimen 2.



- Rekomendasi bisnis berdasarkan Feature Importance yakni sebagai berikut.
  - 'Daily Internet Usage' dan 'Daily Time Spent on Site' adalah fitur terpenting dalam menentukan keberhasilan iklan. Saya merekomendasikan untuk menargetkan pengguna yang memiliki waktu penggunaan internet yang lebih tinggi, karena mereka lebih cenderung untuk terlibat dengan iklan.
  - 'Area Income' juga merupakan faktor penting. Pengguna dengan pendapatan yang lebih tinggi kemungkinan besar akan lebih tertarik pada produk atau layanan premium.
  - Faktor 'Age' menunjukkan bahwa usia pengguna juga berpengaruh dalam menargetkan iklan. Iklan dapat lebih disesuaikan untuk kelompok usia tertentu, terutama mereka yang lebih sering mengklik iklan.
  - Penggunaan 'Week' dan 'Month' sebagai fitur menunjukkan bahwa waktu dalam seminggu dan bulan tertentu bisa menjadi waktu terbaik untuk meningkatkan konversi iklan.

- Simulasi bisnis dilakukan dengan 2 cara yakni tanpa menggunakan hasil dari model machine learning dan dengan menggunakan hasil dari model machine learning.
- Data yang digunakan untuk simulasi yakni sebagai berikut.
  - Jumlah user = 10000
  - Biaya per user = 10
  - Conversion rate tanpa ml = 0.05
  - Harga produk = 100
- Conversion rate yang dihasilkan apabila menggunakan model machine learning yang terbaik yakni sebesar 0.49.
- Profit yang dihasilkan dari simulasi tanpa model machine learning yakni sebesar \$-50000. Sedangkan profit yang dihasilkan apabila menggunakan model machine learning yakni sebesar \$390000.
- Berdasarkan kedua simulasi ini dapat disimpulkan bahwa menggunakan model machine learning meningkatkan profit dan efisiensi marketing secara signifikan. Selain itu, keuntungan yang didapat juga meningkat dari \$-50000 menjadi \$390000.



A faded, light-colored background image of a city skyline with various skyscrapers and buildings.

# THANK YOU

Have a nice day!