Membangun Proyek Machine Learning

Walker Valentinus Simanjuntak

Institut Teknologi Del

Dicoding Indonesia

Abstrak

[The abstract should be one paragraph of between 150 and 250 words. It is not indented. Section titles, such as the word Abstract above, are not considered headings so they don’t use bold heading format. Instead, use the Section Title style. This style automatically starts your section on a new page, so you don’t have to add page breaks. Note that all of the styles for this template are available on the Home tab of the ribbon, in the Styles gallery.]

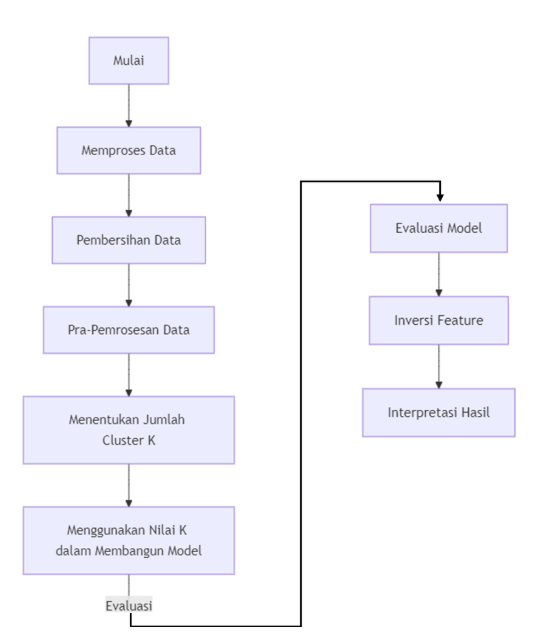
Keywords: [Click here to add keywords.]

Membangun Proyek Machine Learning

Pada proyek ini akan digunakan metode *unsupervised learning* (*clustering*) untuk menemukan pola dan secara otomatis membuat label baru untuk setiap data yang belum memiliki label. Selanjutnya, label hasil *clustering* akan digabungkan ke dalam *dataset* asli, kemudian mengubahnya menjadi data yang siap untuk klasifikasi. Terakhir, dengan dataset yang kini sudah memiliki label, akan dilatih sebuah model *supervised learning* untuk memprediksi kelas-kelas yang telah dibuat sebelumnya.

# Desain Pengerjaan

Pengerjaan proyek akan dibagi menjadi 2 metode. Metode pertama adalah menggunakan *clustering*. Tahapan pengerjaan pada metode ini dapat dilihat pada gambar x berikut.

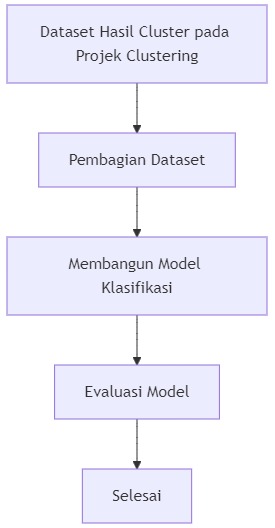


Gambar x clustering flowchart

*Flowchart* tersebut menampilkan metodologi penelitian yang diterapkan. Proses diawali dengan tahap persiapan data yang meliputi pemrosesan data mentah, pembersihan untuk menangani inkonsistensi, atau data yang hilang, serta pra-pemrosesan data seperti normalisasi atau standardisasi fitur. Langkah berikutnya adalah menentukan jumlah *cluster* optimal (K). Setelah nilai K yang paling sesuai didapatkan, nilai tersebut digunakan untuk membangun model *clustering*.

Setelah model terbentuk, alur dilanjutkan ke tahap evaluasi model untuk mengukur kualitas dan validitas *cluster* yang dihasilkan menggunakan metrik yang relevan. Untuk dapat memahami karakteristik dari setiap *cluster* secara intuitif, dilakukan proses inversi fitur (inverse transform), yang bertujuan untuk mengembalikan nilai-nilai fitur yang telah diproses ke skala aslinya. Tahapan terakhir adalah interpretasi hasil, di mana setiap *cluster* dianalisis secara mendalam untuk mendapatkan wawasan atau kesimpulan yang relevan dari pola yang ditemukan dalam data.

Metode kedua adalah menggunakan *classification*. Tahapan pengerjaan pada metode ini dapat dilihat pada gambar x berikut.



Gambar x classification flowchart

Langkah pengerjaan dimulai dari membaca dataset hasil clustering yang sudah dilakukan pada metode pertama. Setelah itu, dataset akan dibagi menjadi beberapa split yaitu train, test, dan val. Ketiga split ini akan digunakan untuk melatih (membangun) dan mengevaluasi performa model.

## Dataset

Dataset yang digunakan Adalah hasil modifikasi tim Dicoding dari dataset Kaggle “Bank Transaction Dataset for Fraud Detection” yang dilakukan untuk penilaian kriteria pemenang. Dataset terdiri dari 2537 baris dan 16 kolom.

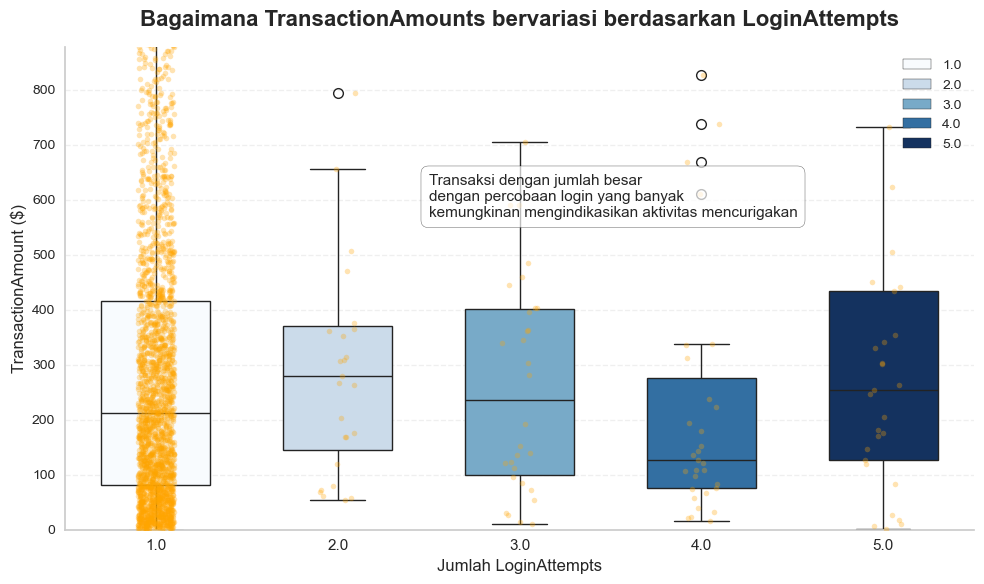
### Clustering

* **Libraries**

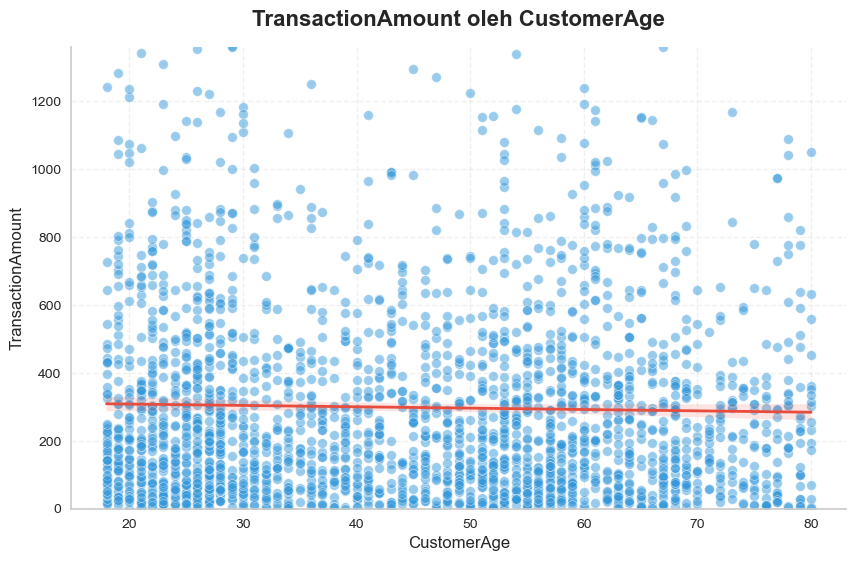
Pada file jupyter Clustering, beberapa library yang digunakan serta penjelasannya dapat dilihat pada tabel x berikut.

|  |  |  |
| --- | --- | --- |
| No | Library | Kegunaan |
| 1 | Pandas |  |
| 2 | Numpy |  |
| 3 | Matplotlib, Seaborn |  |
| 4 | Sklearn (MinMaxScaler, LabelEncoder) |  |
| 5 | KElbowVisualizer |  |
| 6 | Kmeans |  |
| 7 | Silhouette score |  |
| 8 | Joblib |  |
| 9 | Pca |  |

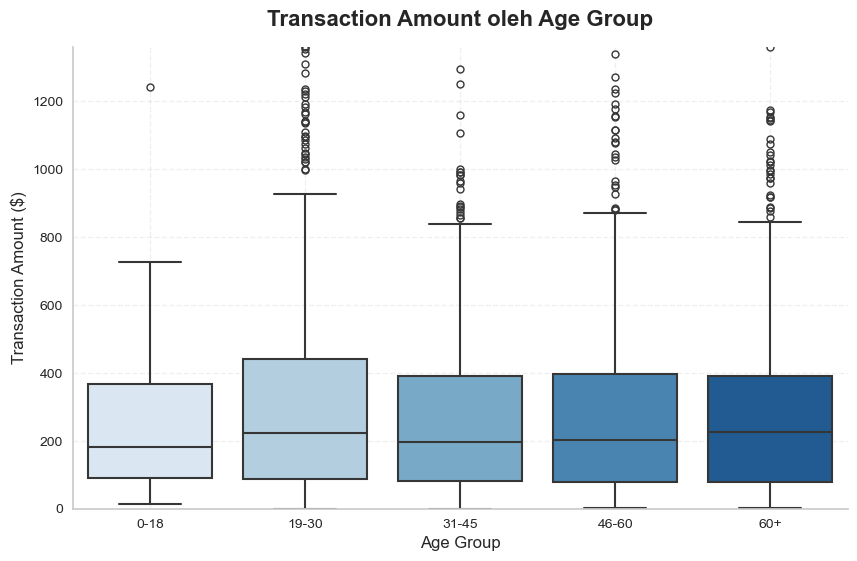
* **EDA (Exploratory Data Analysis)**

****

Transaksi dengan jumlah besar dengan percobaan login yang banyak kemungkinan mengindikasikan aktivitas mencurigakan.

****

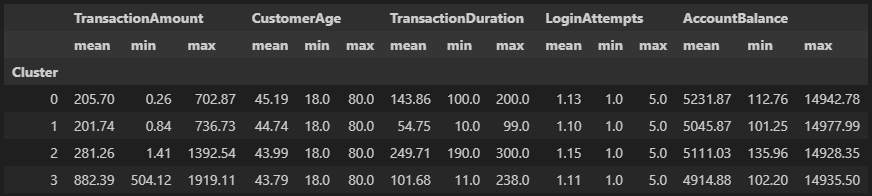
Dapat gambar x, dapat kita lihat bahwa plot regresi tidak menunjukkan adanya tren peningkatan yang mengatakan bahwa semakin tua customer, maka transaksi yang dilakukan semakin besar (ditandai dengan garis merah).

****

Pada boxplot gambar x juga dapat kita lihat bahwa transaksi pada setiap age group hanya meningkat sedikit saja, bahkan dari umur 19-30 tahun ke umur 31-45 tahun, besar transaksinya menurun. Informasi penting yang bisa didapat adalah,

1. Anomali transaksi besar pada umur 60+ patut diwaspadai, hal ini dapat terlihat dari outlier yang sangat banyak pada boxplot di atas. Kelompok usia ini sering menjadi target penipuan seperti phishing (email palsu dari bank) atau penipuan melalui telepon.
2. Anomali juga terdapat pada umur 0-18 tahun, dengan adanya sebuah outlier. Secara logika, anak-anak atau remaja jarang memiliki sumber dana atau limit kartu kredit untuk melakukan transaksi besar. Jika ada transaksi bernilai tinggi, kemungkinannya adalah akun tersebut telah dicuri (di-hack).

* **Clustering Output**

****

Cluster 3 Transaksi "Mencurigakan" dengan rata rata TransactionAmount sebesar 0.46 ($882.39). Analisisnya adalah rata-rata transaksi yang ekstrem $882.39 dan nilai maksimum tertinggi $1919.11 menjadikannya kelompok paling mencurigakan. Penipu seringkali berusaha memaksimalkan keuntungan dalam satu kali aksi. Durasi transaksi yang relatif cepat 101.68 detik bisa menandakan penggunaan skrip otomatis atau tindakan terburu-buru untuk menghindari deteksi.

Cluster 2 Transaksi dengan “Potensi Penipuan Terencana" dengan rata-rata TransactionAmount sebesar 0.15 ($281.26). Analisisnya adalah durasi transaksi yang paling lama dengan 249.71 detik merupakan sinyal bahaya. Penipu mungkin memerlukan waktu lebih lama untuk mengeksplorasi akun yang berhasil dibobol guna mencari limit atau saldo atau melakukan beberapa percobaan sebelum berhasil. Meskipun nilai transaksinya tidak seekstrem Cluster 3, kombinasi nilai menengah ke atas dan durasi tidak wajar menjadikannya kelompok berisiko kedua yang perlu diawasi.

Cluster 1 dengan “Perilaku Normal Berisiko Rendah” dengan rata-rata TransactionAmount 0.11 ($0.84). Analisisnya adalah cluster ini mewakili transaksi sah dan aman. Nilai transaksi yang rendah $201.74 dan durasi yang sangat cepat 54.75 detik menunjukkan pembelian rutin oleh pengguna asli. Transaksi dalam kelompok ini memiliki probabilitas fraud sangat rendah dan menjadi acuan untuk perilaku normal pengguna.

Terakhir, cluster 0 merupakan "Perilaku Standar Berisiko Rendah" dengan rata-rata TransactionAmount 0.10 ($0.26). Analisisnya mirip dengan Cluster 1, kelompok ini juga mencerminkan aktivitas pengguna sah. Durasi yang rendah 143.86 detik bisa menandakan belanja biasa tanpa urgensi tinggi. Dari perspektif deteksi penipuan, cluster ini berisiko rendah dan merupakan bagian dari aktivitas transaksi normal.

#### Classification

* **Data Splitting**
* **Scaler dan Encoder**
* **Model Klasifikasi Baseline**
* **Model Klasifikasi Pembanding**
* **Evaluation**
* **Hyperparameter Tuning**

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Footnotes

1[Add footnotes, if any, on their own page following references. For APA formatting requirements, it’s easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. (Note: If you delete this sample footnote, don’t forget to delete its in-text reference as well. That’s at the end of the sample Heading 2 paragraph on the first page of body content in this template.)]

Tables

Table 1

[Table Title]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Head | Column Head | Column Head | Column Head | Column Head |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures title:

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the APA Style Manual, 6th Edition.