TUGAS BESAR MACHINE LEARNING 2020

Yovan Julio Adam  
1301174236  
IF-41-01

**Implementasi Unsupervised dan Supervised Learning pada Data Mobil Bekas.**

1. PENDAHULUAN

**Machine Learning** adalah metode analisis yang membantu menangani data besar dengan cara mengembangkan algoritma komputer. Dengan menggunakan data, pembelajaran mesin memungkinkan komputer menemukan wawasan tersembunyi tanpa diprogram secara eksplisit saat mencarinya. Dengan adanya email baru, algoritma tersebut kemudian akan menghasilkan prediksi apakah email baru itu spam atau tidak.

Proses pembelajaran dimulai dengan observasi data, seperti contoh: pengelaman langsung, atau intruksi untuk mencari pola data dan membuat keputusan yang lebih baik dimasa depan berdasarkan contoh tersebut. **Tujuan utamanya** adalah membiarkan komputer belajar secara otomatis tanpa intervensi atau bantuan manusia dan menyesuaikan aktivitas yang sesuai. Karena teknologi komputasi, machine learning saat ini tidak seperti machine learning di masa lalu.

**Reduksi dimensi** adalah salah satu teknik pengurangan dimensi. Reduksi dimensi dapat dilakukan dengan seleksi fitur ataupun ekstraksi fitur. Data yang telah dibersihkan seringkali masih berupa data dengan banyak atribut. Banyaknya atribut atau dimensi ini seringnya justru menyebabkan kompleksitas dan waktu proses yang dibutuhkan menjadi lebih lama. Oleh karena itu, dalam pengerjaan tugas besar ini penting untuk melakukan reduksi dimensi.

**Clustering** merupakan salah satu metode dalam teknik unsupervised learning. Metode ini mencari kemiripan antar data untuk kemudian dikelompokkan berdasar kemiripan tersebut. Kemiripan antar data ini dilihat dari jarak antar data. Keluaran dari metode ini berupa atribut baru dalam kelas sesuai cluster.

**Klasifikasi** adalah salah satu teknik supervised learning. Klasifikasi memodelkan data-data ke dalam kelas-kelas yang telah didefinisikan. Dalam memodelkan klasifikasi, dibutuhkan masukan berupa data latih yang memiliki kelas / label. Selanjutnya data latih akan menghasilkan model klasifikasi yang akan memodelkan data uji. Dari hasil pemodelan data uji dapat dilihat akurasi model yang dibangun.

1. A Problem to ML Solution
2. Articulate Your Problem Clearly.

**Masalah** dari dataset yang diberikan adalah tidak adanya atribut kelas. Pada dataset ini terdapat 20.001 baris data dan 26 atribut, namun dari ke-26 atribut tersebut tidak ada atribut kelas. Permasalahan ini mengharuskan proses pemberian label pada dataset agar selanjutnya dapat dilakukan klasifikasi (supervised learning membutuhkan kelas data).

1. Identify Your Data Source.

Dataset yang digunakan adalah data dari situs iklan penjualan barang, **craigslist**. Dataset ini hanya berisi iklan mobil-mobil yang dijual dalam situs tersebut. Dataset ini memiliki 20.001 baris data dan 26 atribut.

1. Identify Potential Learning Problems.

Berikut adalah beberapa potensi yang dapat **menghambat** proses pembuatan model dari machine learning.

1. Data set memiliki banyak **Missing Value**.
2. Data set memiliki banyak kolom yang **tidak digunakan**.
3. Data set memiliki banyak inputan data yang **tidak sesuai**.
4. Think About Potential Bias and Ethics.

Data yang diambil sebagai input untuk melakukan prediksi dan pengelompokan data tidak menyinggung pihak manapun karena tidak terdapat unsur yang menyudutkan suatu pihak atau kelompok tertentu.

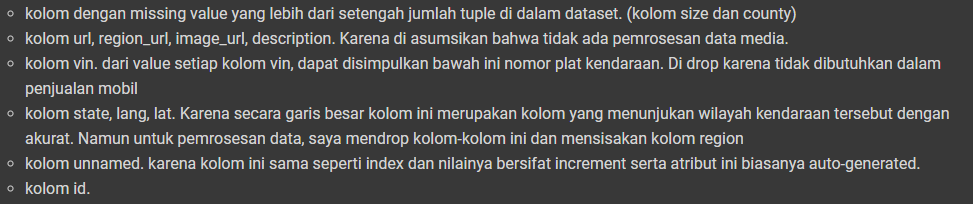
1. Eksplorasi dan Persiapan Data.
2. Data Understanding.

Dalam pengerjaan tugas besar Machine Learning, saya diberikan dataset untuk melakukan unsupervised dan supervised learning. Dataset yang akan diolah adalah used\_car. Dataset yang diberikan harus dapat memodelkan clustering dan klasifikasi. Mahasiswa dibebaskan untuk mengeksplorasi proses menyiapkan data dan mahasiswa harus melakukan evaluasi terhadap model yang dibangun.

1. Data Cleansing.

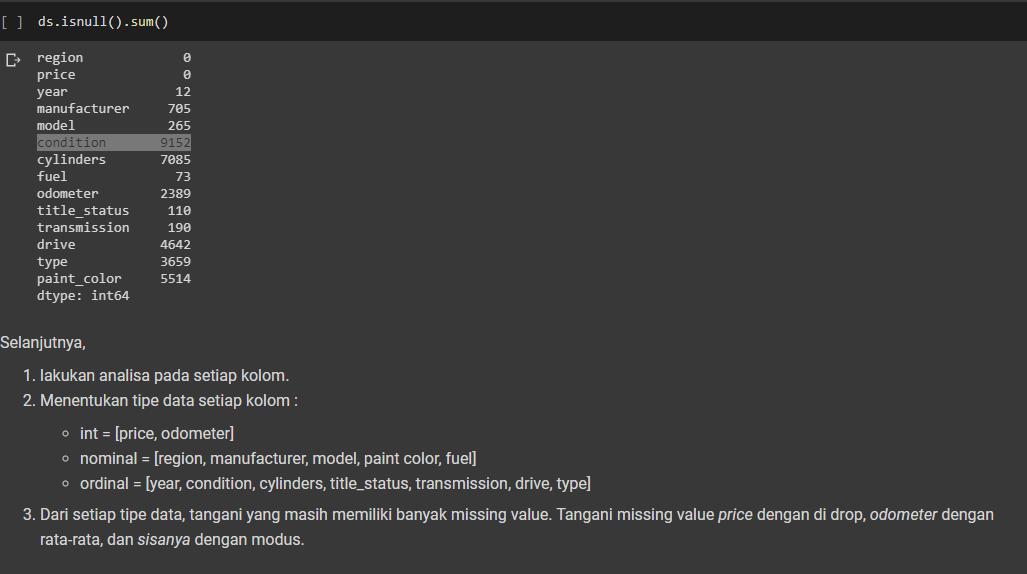
Pada data cleansing ini dilakukan cleaning pada data-data yang tidak digunakan. Kolom-kolom yang tidak digunakan akan di drop. Data yang memiliki missing value sedikit akan di-drop, sisanya diisi dengan rata-rata dan modus, serta mendefinisikan tipe data dari kolom dengan di-replace value angka, nominal, dan ordinal.

* 1. Kolom-kolom yang menurut saya tidak terlalu penting, beserta alasannya :



Gambar 1 Screenshot dari colab untuk keterangan a

* 1. Setelah melakukan analisa pada setiap kolom :

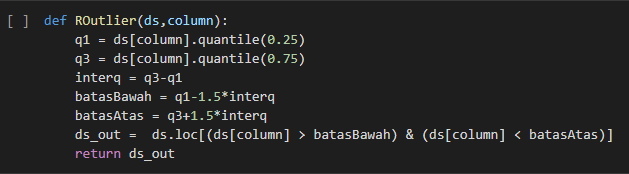


Gambar 2 Screenshot dari colab untuk keterangan b

* 1. Asumsikan, sangat sedikit jumlah orang yang ingin membeli kendaraan dalam keadaan cacat. Maka, status kendaraan yang cacat akan di drop (missing, parts only, salvage).
  2. Asumsikan, tidak ada mobil dengan harga dibawah $1600.
  3. Penanganan outlier dari atribut ‘price’, ‘odometer’ dan ‘year’. Berikut kolom-kolom yang memiliki outlier :

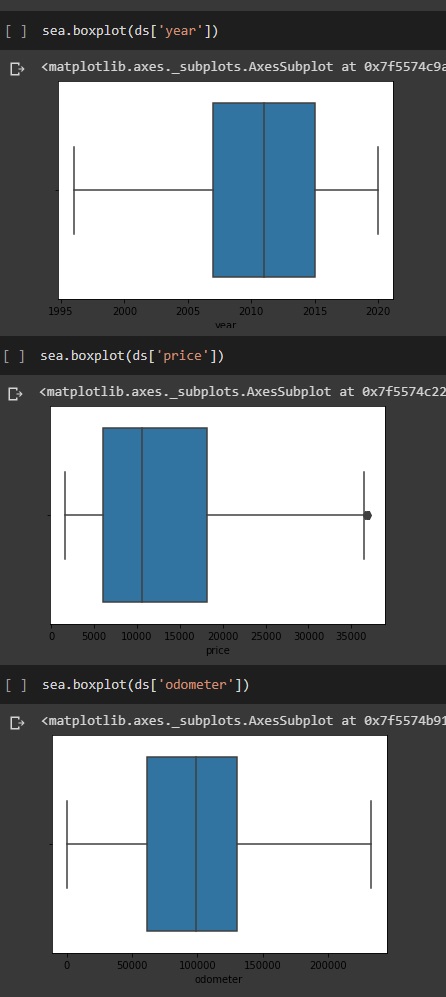


Gambar 3 Visualisasi Boxplot dari kolom pada keterangan e

Lalu, membuat fungsi untuk menghilangkan outlier, sebagai berikut : 

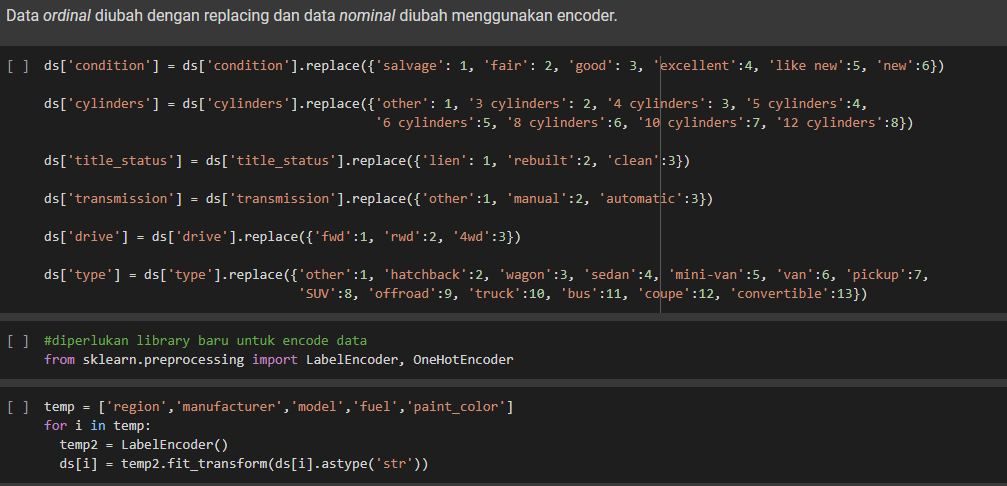
Gambar 4 Algoritma untuk menghilangkan outlier

Berikut visualisasi Boxplot yang sudah dihilangkan ouliernya :



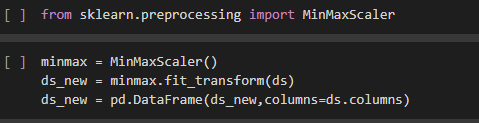
Gambar 5Visualisasi Boxplot dari kolom tanpa outlier pada keterangan e

* 1. Data yang sudah bersih dapat disimpan terlebih dahulu ke dalam file csv ‘clean\_data.csv’ agar diakhir dapat ditambahkan atribut kelas.
  2. Selanjutnya adalah mengganti nilai dalam kolom ordinal menjadi angka tersusun dan kolom nominal di encode, berikut contohnya :



Gambar 6 Algoritma Replacing data dan encode data

* 1. Lalu, dilakukan scalling dengan **MinMaxScaler**, **scaling** adalah tahap menyamakan skala antar atribut dalam dataframe :



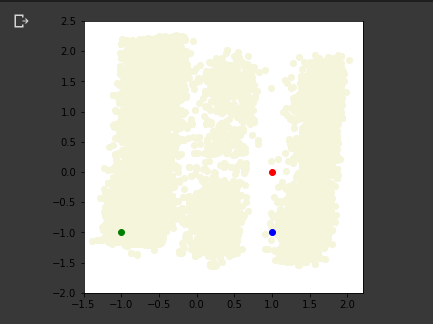
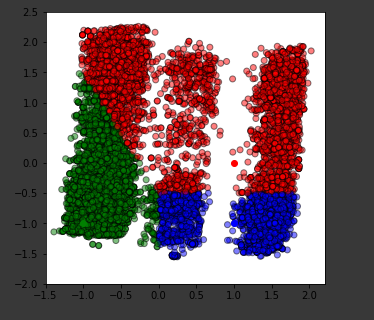
Gambar 7 Implementasi MinMaxScaler

* 1. Selanjutnya **Reduksi Dimensi**.

Data ini akan melakukan clustering, dan metode yang saya pilih adalah K-Means Clustering. Metode ini menerima masukan data 2 kolom untuk kemudian diberi label. Maka dari itu saya harus melakukan transformasi pada data yang telah di-scaling. Saya memutuskan untuk mengurangi jumlah kolom dengan ekstraksi fitur menggunakan PCA. Proses ekstraksi fitur dengan PCA, saya lakukan dengan library karena tingkat kesulitan yang tinggi. Pada penentuan n komponen PCA, saya memilih nilai 0.3. Hal ini karena setelah dicoba dengan nilai yang berbeda akan menghasilkan atribut dengan jumlah lebih selain 2. Saya mengasumsikan nilai ini adalah ini optimal jika ingin mereduksi data menjadi 2 dimensi dari data sebelumnya. Hasil dari transformasi berupa array, maka hasil PCA perlu diubah ke dataframe.

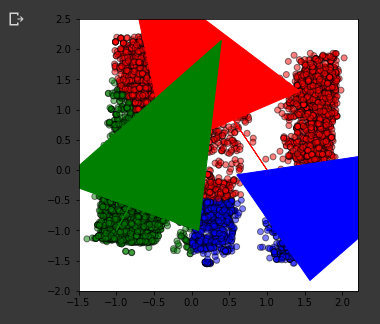
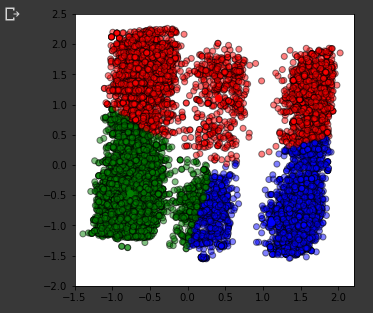
1. Clustering.

Dengan dataframe hasil PCA, maka selanjutnya dibangun model K-Means Clustering dengan k = 3. Kode dan alur clustering dapat dilihat dalam dokumentasi kode. Di bawah ini saya akan menampilkan hasil clustering 3 kelas. Warna merah menandakan kelas ‘1’, warna hijau mendakan kelas ‘2’, dan warna biru menandakan kelas ‘3’. Berikut adalah ilustrasi per-tahap K-Means Clustering :



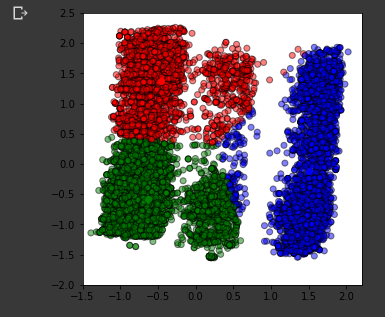
Gambar 8Langkah 2, Pengelompokan setiap instance

Gambar 9 Langkah 1. Pencarian 3 titik steroids secara acak



Gambar 10 Langkah 4. Melakukan assignment sekali lagi untuk setiap cluster

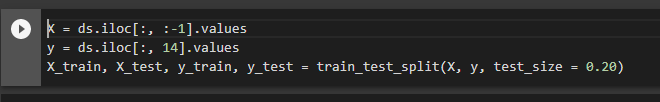
Gambar 11 Langkah 3. Mencari arah dimensi clustering



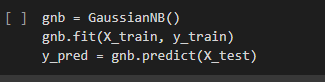
Gambar 12Hasil akhir, Tiap instance mencari cluster terdekat hingga stabil

1. Klasifikasi.

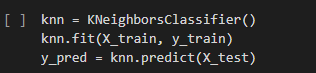
Pada tahap klasifikasi, akan dilakukan eksperimen dengan menggunakan algoritma Naïve Bayes, KNN, dan Random Forest. Ketiga algoritma tersebut membutuhkan masukan berupa data numerikal, maka data yang digunakan adalah data setelah scaling yang telah ditempelkan label dari hasil clustering. Berikut screenshot kode :



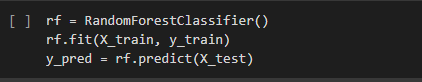
Gambar 13 Split Data



Gambar 14 Naive Bayes



Gambar 15 KNN

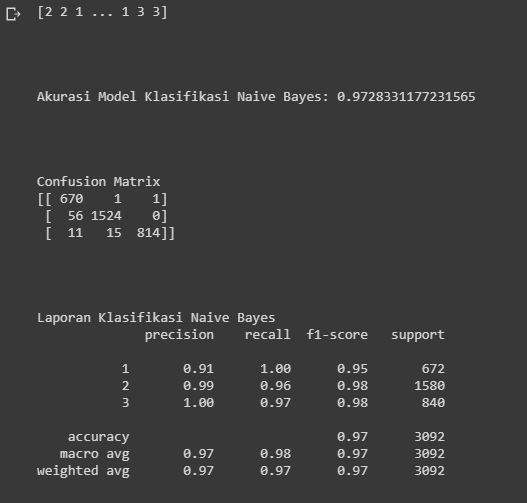


Gambar 16 Random Forest

1. Evaluasi.

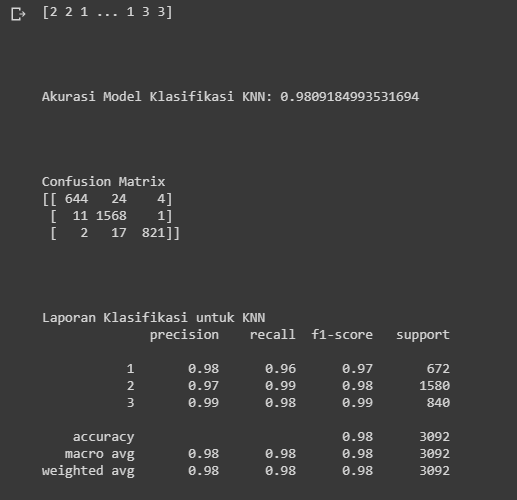
Evaluasi yang dilakukan adalah dengan menggunakan Confusion Matrix karena ini adalah salah satu cara evaluasi yang paling umum terhadap model klasifikasi yang saya bangun. Berikut adalah evaluasi terhadap model klasifikasi yang telah diimplementasikan :

1. Akurasi model Naïve Bayes.



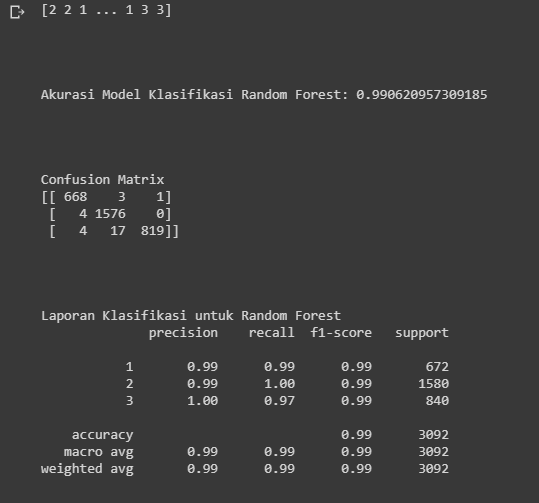
Gambar 17 Akurasi Model Naive Bayes

1. Akurasi model KNN



Gambar 18 Akurasi model KNN

1. Akurasi model Random Forest



Gambar 19 Akurasi model Random Forest

Berdasarkan hasil evaluasi pada model klasifikasi, algoritma Random Forest memiliki tingkat akurasi paling tinggi.

1. Kesimpulan.

Berdasarkan semua proses yang dilakukan kesimpulan yang dapat diambil adalah dengan algoritma dan eksplorasi data yang berbeda maka akan menghasilkan model yang berbeda pula. Selain itu pelajaran yang didapat dari semua proses yang dilakukan yaitu data eksploration atau data understanding merupakan tahapan yang sangat penting karena dapat memberi pengaruh yang sangat besar bagi model yang akan dibangun. Selain data eksploration dan algoritma terdapat faktor yang juga tidak kalah penting yaitu penentuan jumlah cluster karena berdasarkan eksperimen yang dilakukan hal tersebut dapat mempengaruhi besarnya akurasi dari model yang dibangun.

Saran untuk improvement kedepan akurasi dapat ditingkatkan dengan menambahkan fitur atau dimensi yang mendukung proses clustering dan klasifikasi sehingga dalam membangun moden dan prediksi memiliki semakin banyak pertimbangan dan mungkin dapat menambah jumlah data yang digunakan untuk data train dan data uji.