**Laporan Tugas Besar Pembelajaran Mesin**

**Budi Nugroho/1301174220**

**IF41-02**

**Used\_Cars**

**Formulasi Masalah**

Dalam permasalahan ini saya mendapatkan dataset ‘used\_cars.csv’ yang memiliki banyak data sebesar 20001 row dan memiliki 26 colom yant terdiri dari atribut unnamed, ide, url, region, region\_url, price, year, manufacturer, model condition, cylinders fuel, odometer, title\_status, tranmission, vin, drive, size, type, paint\_color. Dalam mengerjakan tugas besar ini disarankan menggunakan 4 tahap problem solution menggunakan mesin learning, yaitu :

1. Cluestering

Cluestering adalah metode penganalisa data yang bertujuan untuk mengelompokkan data dengan mengelompokkan dimana dalam satu kelompok mempunyai karakteristik yang sama lainnya dan mempuyai karakteristik yang berbeda dengan data yang ada di dalam kelompok lainya. Dalam penentukan cluestering ini saya menggunakan pengelompokkan *non-Hierarchial* dengan menggunakan K-Means*(unsupervised)*. Langkah-langkah dalam menggunakan K-Means adalah pertama saya menentukkan cendroids(k), setelah itu mengklasifikasikan data berdasarkan jumlah titik data, setelah itu menghitung centroid cluster yang dilakukan berulang untuk mendapatkan centroids yang optima

1. Classification

Klasifikasi ini bertujuan untuk memetakkan satu titik data ke dalam kelas yang telah ditentukan. Pada Klasifikasi ini dilakukan secara *supervised.* Dalam hal ini saya menngunakan dua cara klasifikasi yaitu untuk model pertama saya menggunakan naive bayes dan model kedua saya menggunakan *K-Nearest Neighbors*(Knn)*.* Pada model pertama saya memilih naive bayes karena lebih sedehana, cepat serta memiliki akurasi yang tinggi. Pada model kedua saya memilih *K-Nearest Neighbors*(Knn) karena dapat menangani dengan banyak data dan akurasinya sangat baik.

1. Hasil Program

Dalam membuat program tugas besar ini, dilakukan dengan penyelesaian untuk dua model. Disini saya untuk model pertama menggunakan pemodelan dengan mengambil atribut odometer dan long untuk mentukan awal mobil digunakan sampai penggunaan yang terakhir. Sedangkan untuk model kedua saya menggunakan atribut long dan lat untuk menentukan lokasi awal sampai lokasi tujuan setiap mobil digunakan.

1. Data Preparation dan Persiapan Data

Membaca data Csv yang diberikan



Melihat describe data

Model pertama



 Model Kedua

Melihat apakah value dari setiap atribut ada yang kosong atau null atau Nan

Model pertama



Model kedua



Melihat type data dari setiap value atribut untuk menentukkan atribut yang diambil, disini saya mengambil type data float atau integer yang saya gunakan untuk pemodelan.

Model pertama



Model kedua



Menghapus value dari setiap atribut yang bernilai Null atau kosong

Model pertama

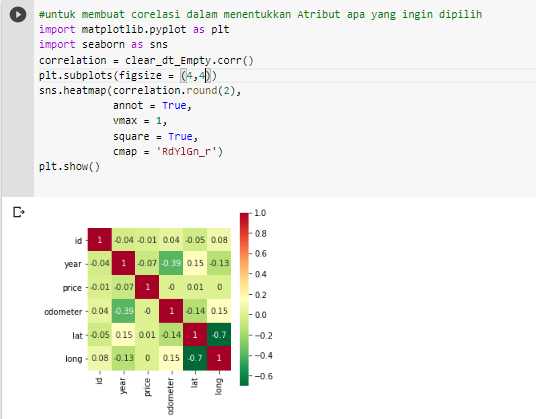


Model Kedua

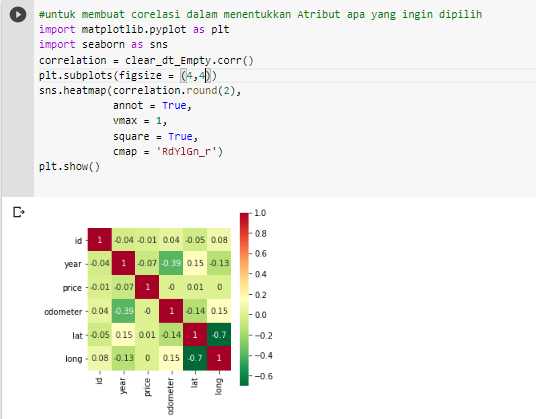


Melihat korelasi antar Atribut

Model pertama



Model kedua



Mengambil values Atribut yang ingin dilakukan pemodelan dan disini saya mengambil dua atribut pada setiap pemodelan

Model pertama



Model kedua

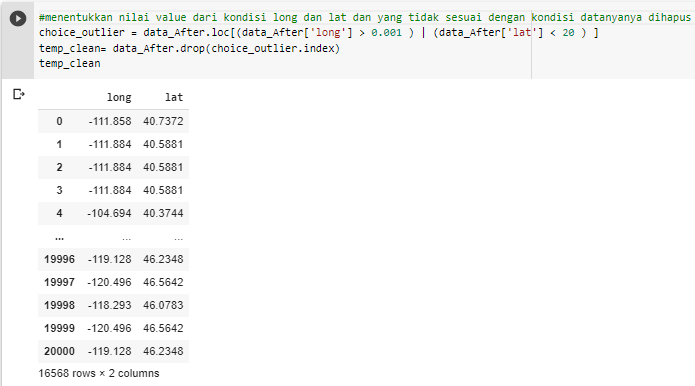


Menghapus data sesuai kondisi yang dinyatakan outlier

Model pertama

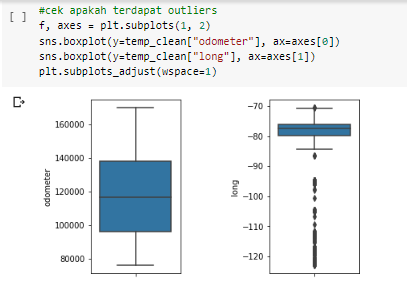


Model kedua

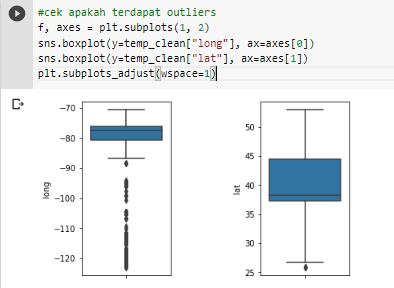


Mengecek value pada Atribut adanya outlier atau tidak

Model pertama

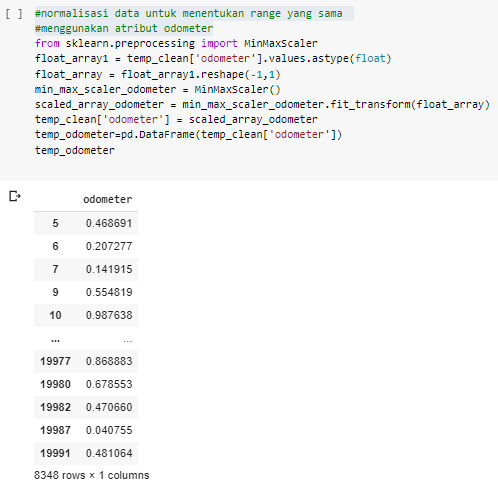
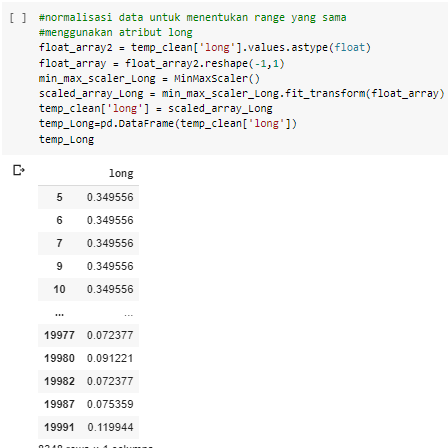


Model kedua

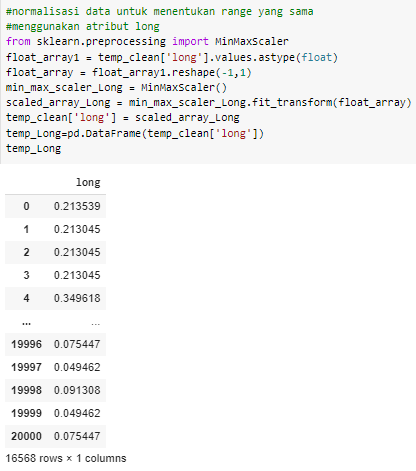
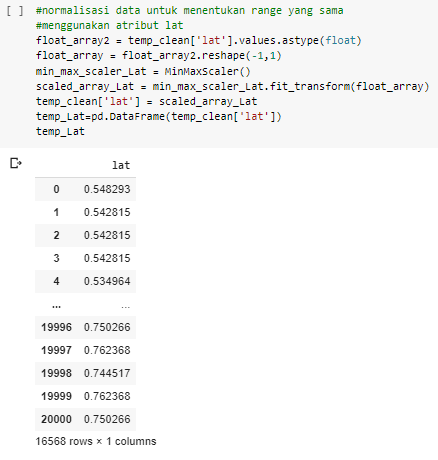


Mencari min\_max\_scaler digunakan menormalisasikan data agar jaraknya tidak terlalu jauh dan gap rangenya adalah (0 sampai 1)

Model pertama

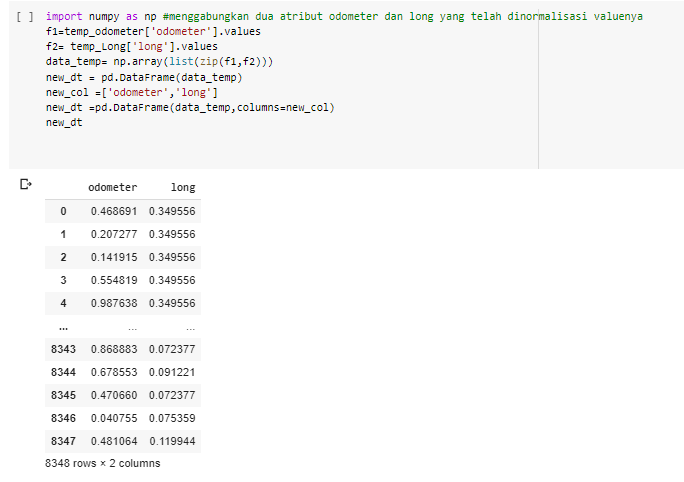


Model kedua

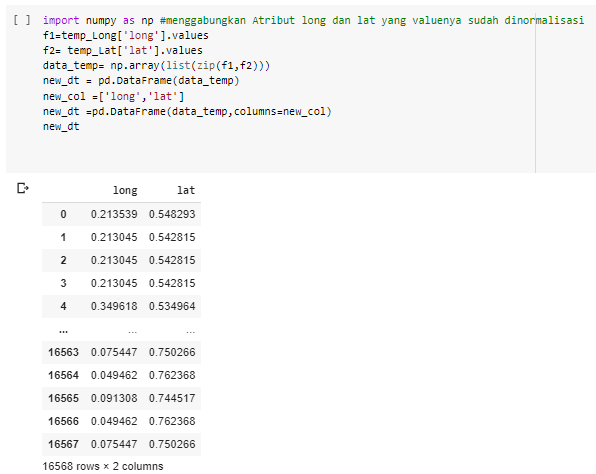


Menggabungkan Atribut yang dinormalisasi karena saya pertama memisahkan per atribut yang ingin di normalisasi agar lebih jelas dan mudah dalam mengetahui nilai dari Atributnya dan fungsi dari list(zip(f1,f2)) adalah mengabungkan nilai antar atribut yang telah dinormalisasi.

Model pertama

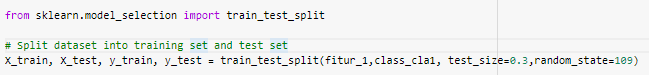


Model kedua



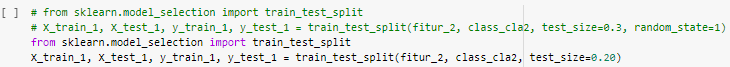
Data Split digunakan pada saat klasifikasi untuk membagi data\_train dengan data\_size

Model pertama



Artinya digunakan 70 persen menggunakan data\_train dan 30 persen menggunakan data\_size.

Model kedua



Artinya digunakan 20 persen untuk data train size dan 80 persen mengguanakan data\_train.

1. **Cluestering (pengelompokan)**

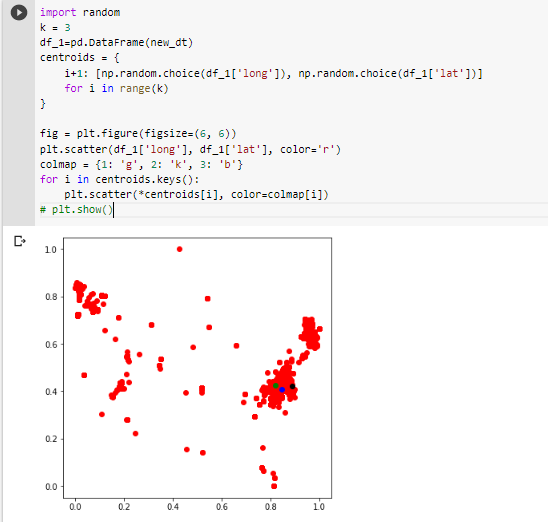
Dalam mencari cluestering ini menggunakan dua pemodelan.

Pada hal ini saya memilih centroids acak/random pada persebaran value atribut. Untuk pemodelan pertama saya menggunakan atribut odometer dan long dan sedangkan atribut kedua saya menggunakan atribut long dan lat.

Model pertama

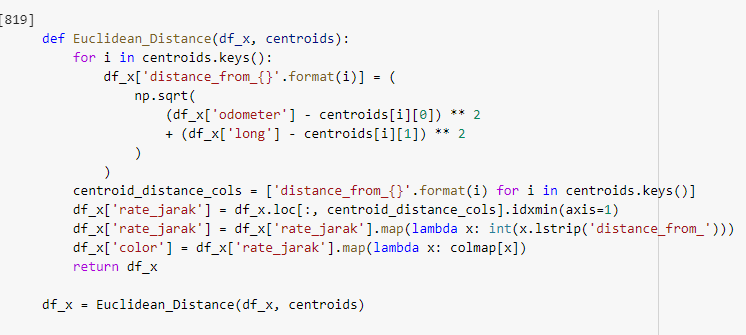


Model kedua

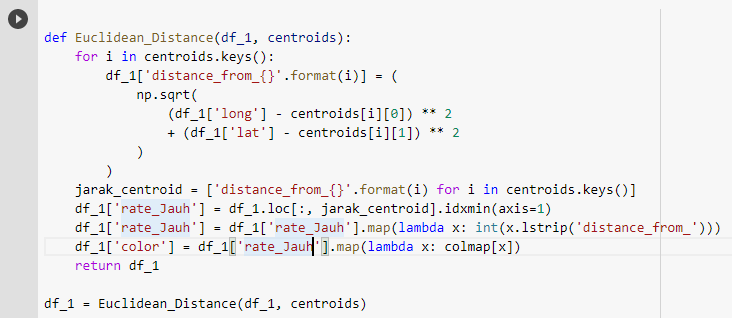


Menentukan jarak agar bisa dikelompokan, pada hal ini saya membuat atribut baru bernama rate\_jarak. Hal ini untuk mengelompokkan tentang value dari setiap atribut jaraknya dekat dengan clustering apa.

Model pertama

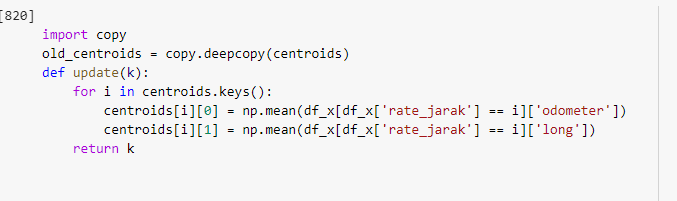


Model kedua



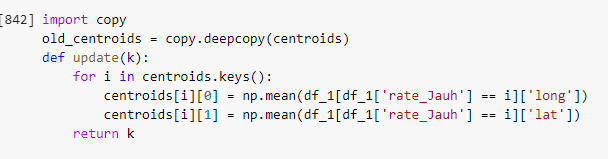
Setelah itu hal selajutnya adalah mengupdate centroid dan menentukkan cluster agar posisi dari centroid tidak berubah-ubah.

Model pertama

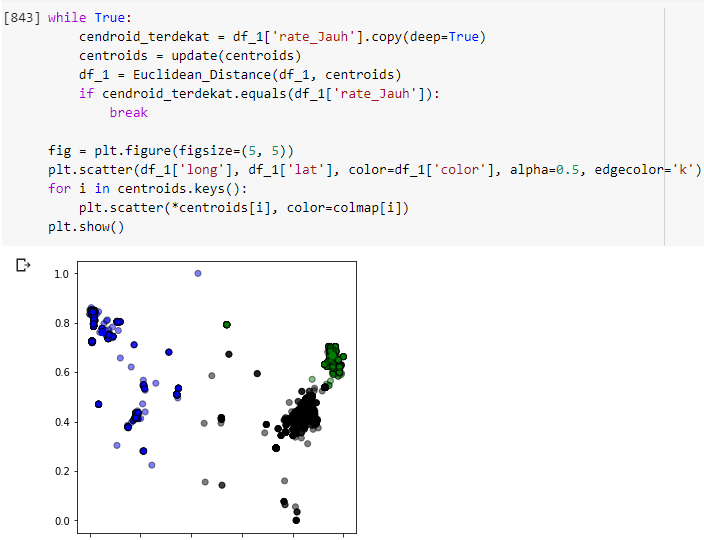


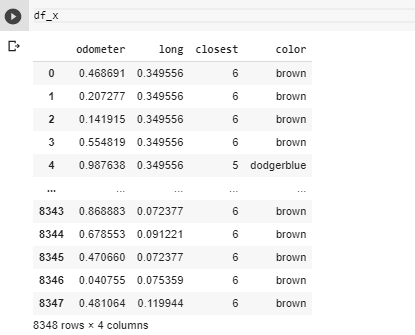


Model kedua

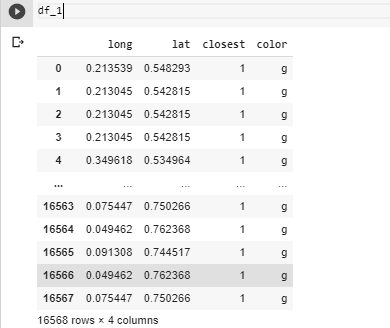


Setelah itu data yang telah dicluestering datanya akan terbagi berdasarkan rate\_jauh terbagi menjadi tiga yaitu 1,2,3



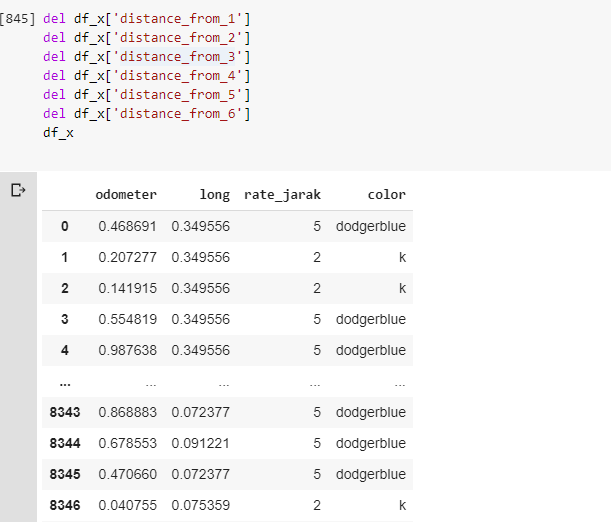


Setelah itu data yang telah dicluestering datanya akan terbagi berdasarkan rate\_Jauh terbagi menjadi tiga yaitu 1,2,3

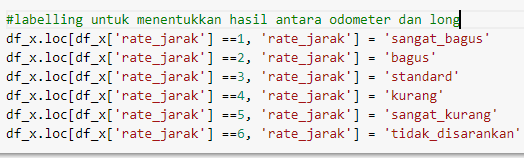


1. **Clasification**

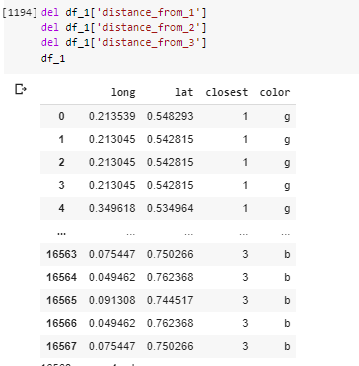
Pada model pertama saya menentukan ratae\_jarak ada 1,2,3,4, 5,6. Yang artinya apakah value gabungan dari odometer dan long masuk pada rate\_jarak yang mana.



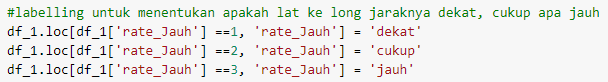
Dan melakukan labelling dengan mengubah hasil pada rate\_jarak dengan string



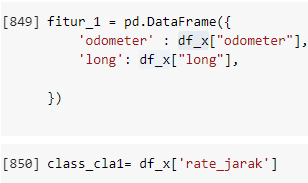
Pada model kedua saya menentukan rate\_Jauh ada 1,2,3. Yang artinya apakah value gabungan dari odometer dan long masuk pada rate\_Jauh yang mana.



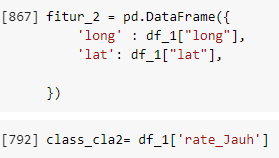
Dan melakukan labelling dengan mengubah angka ke string agar lebih mudah mengetahuinya.



Menentukkan kelas yang digunakan agar bisa di split valuenya yaitu kelas fitur\_1 yang berisi atribut long dan lat, sedangkan kelas satunya adalah class\_cla1[‘rate\_jarak’]

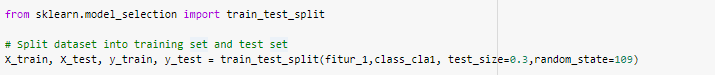


Menentukkan kelas yang digunakan agar bisa di split valuenya yaitu kelas fitur\_2 yang berisi atribut long dan lat, sedangkan kelas satunya adalah class\_cla2[‘rate\_Jauh’]

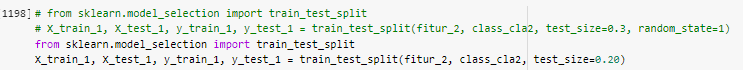


Split data berdasarkan kelas

Model pertama



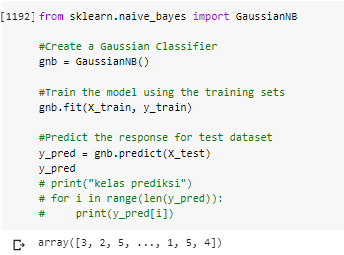
Model kedua



Klasifikasi algoritma dan hasil dengan plot

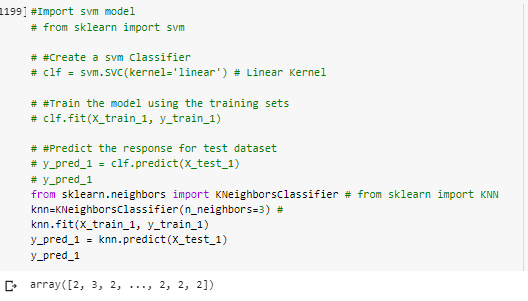
Model pertama

Menggunakan naive bayes karena pada model pertama ini datanya berkurang cukup besar dan naive bayes bisa menangani data yang tidak terlalu besar dengan baik.



Model kedua

Menggunakan knn untuk model kedua ini karena Knn dapat menangani proses banyak data dengan baik dan efektif.

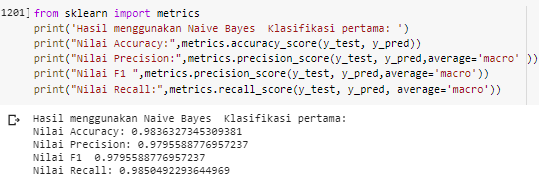


1. **Evaluation**

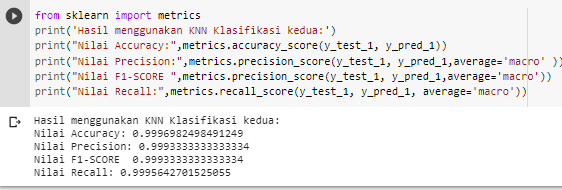
Pada hal ini saya menggunakan import library dari metric yang berguna untuk mengevaluasi hasil klasifikasi algoritma dari naive bayes dan svm. Pada evaluation metric ini terdapat akurasi, recall, presisi dan f1-score. Akurasi adalah

Untuk menunjukkan keakuratan algorima. Recall adalah sistem untuk menemukan informasi yang baik. Presisi adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. F1-score adalah perbandingan rata-rata presisi dengan recall.

Menggunakan naive bayes pada model pertama



Menggunakan KNN pada model kedua



1. **Kesimpulan**

Dari hasil yang telah didapatkan, dapat dilihat bahwa pemodelan kedua yaitu dengan mengambil atribut long dan lat memiliku acurasi, presisi, recalldan f1-score yang baik dikarenakan dengan menggunakan algoritma knn yang bisa memproses data yang sangat banyak karena pada pemodelan kedua datanya cukup banyak yaitu 16568 row.

1. **Saran**

Untuk kedepannya agar datase yang diberikan lebih variatif datanya agar lebih baik saat proses cluestering dan untuk penyelesaian masalahnya bukan hanya tentang data tapi juga ditambahkan menggunakan penyelesaian klasifikasi yang baik agar menghasilkan data keluaran yang baik pula.

Referensi

<https://www.youtube.com/watch?v=1XqG0kaJVHY&t=1380s>

<https://medium.com/bosbouw/k-nearest-neighbors-menggunakan-python-bd3652ba1e70>

<https://www.geeksforgeeks.org/ml-feature-scaling-part-2/?ref=rp>

ppt pembelajaran mesin (slide dosen)