Nama Kelompok :
1. Samuel Erlangga (NIM 1301180307)
2. Rizky Fauzi Ramadhani (NIM 1301184144)

Kelas : IF-42-10

Formulasi Masalah

Terdapat 2 dataset yang diberikan yaitu, salju dan kendaraan. Dataset salju diminta untuk mencari tahu apakah besok akan turun salju atau tidak, sedangkan dataset kendaraan diminta untuk mencari tahu apakah pembeli tertarik untuk membeli kendaraan baru atau tidak. Dataset tersebut digunakan dengan menggunakan classification KNN. Dataset yang akan digunakan yaitu dataset kendaraan untuk mencari tahu apakah pembeli tertarik untuk membeli kendaraan baru atau tidak.

Eksplorasi dan Persiapan Data

Kami menggunakan dataset kendaraan, kemudian eksplorasi data adalah dengan mengidentifikasi variabel.

Type of Variable :

- Predictor : Jenis_Kelamin, Umur, SIM, Kode_Daerah, Sudah_Asuransi, Umur_Kendaraan, Kendaraan_Rusak, Premi, Kanal_Penjualan, Lama_Berlangganan
- Target : **Tertarik**

Data Type :

- Character/String : Jenis_Kelamin, Kendaraan_Rusak, Umur_Kendaraan
- Numeric : Umur, SIM, Kode_Daerah, Sudah_Asuransi, Premi, Kanal_Penjualan, Lama_Berlangganan, **Tertarik**

Variable Category :

- Categorical : Jenis_Kelamin, SIM, Kode_Daerah, Sudah_Asuransi, Kendaraan_Rusak, **Tertarik**
- Discrete : Premi, Kanal_Penjualan
- Continuous : Umur, Umur_Kendaraan, Lama_Berlanggagan

Kemudian dilakukan perhitungan dataset, kemudian melihat info dari dataset, kemudian melakukan preprocessing data dengan mendrop/menghapus baris yang berisi kosong, agar nanti dataset dapat diproses dengan sangat baik. Kemudian kami mengubah data agar menjadi numerik, yang pertama diubah yaitu pria dan wanita diubah menjadi 1 dan 0 , umur kendaraan < 1 tahun diubah menjadi 0, umur kendaraan 1-2 tahun menjadi 0, dan umur kendaraan > 2 tahun diubah menjadi 2, dan yang terakhir kendaraan rusak, jika kendaraan tidak rusak diubah menjadi 0 dan kendaraan pernah rusak diubah menjadi 1. Kami juga melakukan describe pada dataset agar dapat mengetahui count, mean, std, min, 25% , 50%, 75%, dan max dataset.

Kemudian dilakukan normalisasi dataset dengan menggunakan library minmaxscaler agar tidak terjadinya tumpang tindih data dengan mengubah nilai dan range data. Kemudian melakukan korelasi heatmap untuk melihat korelasi antar kolom.

Implementasi algortima KNN dilakukan dengan *5 fold cross validation* yang menghasilkan akurasi tertinggi. Dataset baru dibagi menjadi 5 fold yang dibagi lagi dimana 1 fold berisi dataset testing dan training untuk menghasilkan rata-rata akurasi dari algoritma kNN yang digunakan.

Pemodelan

Strategi yang digunakan untuk implementasi penyelesaian masalah algoritma k-Nearest Neighbor (kNN) adalah menggunakan bahasa pemrograman Python pada platform Google Colab. Berdasarkan dataset yang diberikan pada file "Kendaraan.csv" dilakukan scaling/normalisasi pada dataset menggunakan rumus :

$$xn = x - min \max - min$$

Kemudian,setelah dataset dinormalisasi, dilakukan perhitungan jarak menggunakan formula Manhattan Distance sebagai berikut :

$$d(i,j) = |xi1 - xj1| + |xi2 - xj2| + \cdots + |xip - xjp|$$

Hasil dari perhitungan jarak menggunakan formula Manhattan distance kemudian diurutkan berdasarkan jarak paling dekat dan dipilih sebanyak k satuan. Kemudian dilakukan pengecekan perhitungan prediksi dataset testing dan training dan pemilihan k, jika output benar maka kemudian nilai tersebut ditampung ke dalam hasil prediksi. Lalu hasil dari 5 dataset tadi, digabungkan dan dibagi berdasarkan *5-fold cross-validation* untuk menghasilkan rata-rata akurasi dari algoritma kNN yang digunakan. Kemudian yang terakhir, dilakukan pengujian nilai k yang paling efisien berdasarkan akurasi paling maksimal yang menjadi k terbaik diantara nilai k lainnya. Pengujian algortima untuk mendapatkan k terbaik beserta akurasinya dengan jangkauan k dari (1-100) menghasilkan k terbaik yaitu 1 dengan rata-rata akurasi kNN berdasarkan *5-fold cross validation* adalah 100 % yang ditampilkan menggunakan grafik agar dapat terlihat dengan jelas pemilihan k terbaik berdasarkan akurasinya.

Evaluasi

Metode evaluasi yang digunakan adalah dengan menghitung nilai akurasi antara hasil prediksi dengan data *testing*. Kami juga menggunakan teknik *5-fold cross-validation* untuk merata-ratakan akurasi dari setiap fold yang diprediksi. Hasil dari prediksi pada setiap fold akan dibandingkan dengan data *testing* bila nilainya sama maka akan dihitung berapa banyak nilai yang sama lalu dibagi dengan jumlah banyaknya data.

$$acc = jumlahSama \,/\, jumlahData$$

Maka akan ditemukan akurasi untuk satu fold data, setelah itu akurasi akan dihitung untuk setiap fold yang telah dilakukan proses klasifikasi. Dari akurasi setiap fold yang dihitung akan diambil fold dengan nilai akurasi terbaik sebagai hasil terbaik untuk nilai K tersebut.

Akurasi akan dihitung untuk setiap nilai K. Setelah menghitung setiap akurasi pada setiap K maka akan diambil nilai K dengan akurasi paling tertinggi sebagai nilai K terbaik. Data akurasi setiap nilai K ditampilkan dalam bentuk diagram garis.

Eksperimen

Eksperimen yang dilakukan adalah dengan mengubah nilai K yang digunakan dalam menghitung jarak tetangganya. Nilai K yang digunakan dimulai dari 1 sampai 100, pada setiap nilai K akan dihitung nilai akurasinya dan dicari nilai K dengan akurasi paling tinggi sebagai nilai K dengan hasil terbaik.

Kesimpulan

Dari pemodelan dan eksperimen yang dilakukan dapat disimpulkan bahwa, K yang diuji dari nilai K sama dengan 1 sampai 100 dengan akurasi paling tinggi didapat pada nilai K sama dengan 1 dengan akurasi 100%.