UAS Praktikum Data Mining

Topik : Machine Learning dengan Python Scikit-Learn

****

Problematika:

Pada situs UCI Repository (*https://archive.ics.uci.edu/*), terdapat sebuah dataset bernama ***Adult Dataset*** atau dikenal juga dengan sebutan ***Census Income*** dataset. Rincian informasi terkait dataset ini dijelaskan pada gambar di bawah ini.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	1962641

Pada informasi diatas, diketahui bahwa dataset ini bertipe *multivariate* (campuran numerik dan kategorik) dan memiliki *missing values.* Artinya dataset ini mengandung sejumlah data yang tidak lengkap atributnya sehingga tidak bisa langsung digunakan untuk membangun model. Anda dapat mengunduh dataset ini melalui link berikut : http://archive.ics.uci.edu/ml/machine-learning-databases/adult/.

Petunjuk Pengerjaan

Pada tugas kali ini, anda diminta untuk membangun sebuah Machine Learning dengan menggunakan bahasa Python berdasarkan dataset tersebut. Jika anda mengklik link dataset ini, anda akan menjumpai dua file bernama adult.data dan adult.test. Unduh kedua file ini, kemudian ubah ekstensi kedua file menjadi .csv terlebih dahulu, dimana file *adult.data diganti menjadi adult-training.csv dan file adult.test menjadi adult-test.csv*. Pastikan bahwa kedua file harus diberi header (nama kolom) untuk setiap kolomnya dengan mengikuti nama kolom yang tertera pada link dataset ini di UCI Repository : UCI Machine Learning Repository: Adult Data Set ***(INGAT! Header dari kedua file harus sama agar tidak terjadi error).***

Setelah anda mengubahnya menjadi file, langkah berikutnya adalah bersihkan semua missing values yang ada pada kedua file. Perlu diketahui bahwa sebuah data disebut sebagai missing values jika sekurang-kurangnya ada satu atributnya yang tidak mengandung nilai. Pada kedua dataset ini, nilai kosong disimbolkan dengan tanda tanya "?". Jika sebuah data tidak lengkap dan mengandung satu saja atribut dengan nilai "?", maka hapus data itu.

Pastikan bahwa semua data kotor ini harus bersih hingga tidak tersisa. Anda boleh menggunakan library apa saja, namun kami merekomendasikan untuk menggunakan library Pandas karena lebih efisien dan mudah untuk digunakan.

Setelah kedua file dibersihkan, buatlah sebuah file baru bernama adult-complete.csv yang berisi gabungan seluruh data dari adult-training.csv dan adult-test.csv. Jadi pada akhirnya anda akan memiliki tiga buah file .csv yang siap digunakan untuk membangun Machine Learning.

Prosedur Pembangunan Model

Pada tahapan ini, anda harus membangun model dengan menggunakan dua teknik yang berbeda. Kedua teknik tersebut yaitu:

1. Train Test Split : anda harus menggunakan file adult-complete.csv untuk membangun model. Pada metode ini, file ini harus dipisah menjadi data latih dan data training dengan menggunakan bantuan function train_test_split yang sudah ada pada python. Metodenya sama dengan yang diajarkan di praktikum. Pada function tersebut, setel parameter test_size=0.2 dan random_state=42.
2. Using Data Testing : anda harus membangun model dengan menggunakan semua data pada file adult-training.csv dan menguji akurasi model dengan data pada file adult-test.csv. Disini, file adult-test.csv menjadi data testing yang akan dipakai dalam menguji akurasi model.

Untuk jenis algoritma yang digunakan, anda hanya boleh menggunakan dua dari empat algoritma berikut : **KNN, Naive Bayes, Decision Tree, SVM, dan Neural Nets**. Pastikan bahwa code yang anda tulis mampu mencetak nilai akurasi, precision, recall, dan F-measure dari model yang dibangun.

Setelah anda berhasil membangun dua model yang berbeda tersebut, buatlah sebuah laporan yang menjelaskan tentang semua langkah yang anda lakukan, mulai dari tahapan pembersihan data, teknik membangun model, hingga penjelasan tentang hasil pengujian model. Kemudian, simpulkanlah manakah teknik yang menghasilkan model terbaik dari keduanya.

Tugas ini adalah tugas individual dan tidak dibenarkan untuk saling menyontek. Apabila ketahuan terdapat laporan yang sama persis, maka nilai dari semua yang sama tersebut adalah nol (0) untuk tugas ini.

Pengumpulan

Laporan harus dinamai dengan format **NIM-Nama-FINAL-DM.zip**. Isi file .zip itu adalah ketiga dataset yang sudah dibersihkan, semua code yang digunakan, dan sebuah laporan dengan format nama **NIM-Nama-FINAL-DM.pdf.** Laporan wajib dikumpulkan via Email paling lambat pada tanggal 04 Januari 2021 pukul 23.59 WIB. Jika laporan terlambat untuk dikumpulkan, maka nilai akan dikurangi menurut keterlambatannya.

- **Praktikum Kamis** -> (dinaluthfi@mhs.unsyiah.ac.id)
- **Praktikum Jumat** -> (tharianisa@mhs.unsyiah.ac.id)

- Selamat Bekerja. Happy Coding -