Tugas 1 Praktikum Sains Data

Semester Genap Tahun Ajaran 2021/2022

Petunjuk Umum:

1. Kerjakan secara individu
2. Kerjakan tugas ini dengan bahasa pemrograman python. Anda disarankan menggunakan jupyter untuk mengerjakan tugas ini.
3. **Sertakan penjelasan untuk setiap variable yang digunakan dan setiap proses secara singkat** di samping potongan kode (dengan '#'). **Sertakan juga penjelasan program secara lengkap** (idenya apa, bagaimana cara eksekusi dalam program atau algoritma program yang digunakan, penjelasan atau analisis dari model yang dibuat) pada **cell dibawah program.**
   Contoh:

```
In [1]: a=input("Ini buat input: ") #untuk menyimpan yang akan diprint
        b=str(a) #paksa nilai dari variabel a menjadi str
        print(b)

        Ini buat input: output
        output
```

```
Program ini adalah program untuk print input dari user.
Idenya adalah menyimpan nilai input dari user kedalam suatu variabel lalu variabel yang disimpan akan diprint.
Algoritmanya:
1. Simpan input user dalam sebuah variabel a
2. Paksa variabel input menjadi sebuah string lalu simpen ke variabel baru b
3. Print variabel b
```

4. File yang harus diunggah terdiri dari:
   a. beberapa model dalam format **.pkl**. Penamaan untuk model dibebaskan, selama tidak mengandung SARA, ujaran kebencian, dll.
   b. satu file python notebook (**file berbentuk .ipynb BUKAN .py**).
5. Semua file disatukan dalam **1 (satu)** file .zip, dengan format penamaan:
   Nama_NPM_Kelas SIAK Sains Data_Tugas1PrakSainsData.zip
   Contoh penamaan yang benar:
   Itadori Yuji_190688675_A_Tugas1PrakSainsData.zip
6. Batas pengumpulan tugas ini adalah **Kamis, 21 April 2022 pukul 23.00**. Tugas dikumpulkan sesuai dengan kelas SIAK anda:
   Kelas A: Kelas A EMAS2
   Kelas B: Kelas B EMAS2
   *mohon perhatikan waktu pengumpulan yang tertera dan kumpulkan tugas secara tepat waktu.
7. **Dilarang melakukan plagiarism** atau menduplikasi dalam mengerjakan tugas ini. Apabila terdapat kesamaan program atau penjelasan pada tugas yang dikumpulkan, **NILAI TUGAS PRAKTIKUM SAINS DATA ANDA LANGSUNG MENJADI 0 TANPA PERINGATAN** bagi semua pihak yang terlibat plagiarism dalam tugas ini.
8. Module yang boleh digunakan pada tugas ini adalah pandas, scikit-learn, numpy, scipy, matplotlib, pickle, joblib. Penggunaan module selain yang disebutkan harap dikonfirmasikan ke narahubung terlebih dahulu.

9. Apabila ada yang ingin ditanyakan, silakan mengontak salah satu kontak berikut:
   Muhammad Shiqo Filla (line: mshiqofilla)
   Hanifah Sulasri (line: hanifahunt)

Nomor 1

Cobalah membuat data artifisial yang berkorelasi secara linier dengan banyak data 20, yakni $y = aX + b + Gaussian\ Noise,$ misal dengan kode seperti berikut

```
#cobalah mengganti nilai a dan b
a, b = 3, -5
X = np.random.rand(20)
y =  a*X + b + np.random.rand(20)
```

Tunjukkan scatter plotnya. Nilai koefisien dan intercept fungsi regresi dari suatu data X dan y yang berkorelasi secara linier dapat dinyatakan oleh formula berikut

$$intercept = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}, \qquad coef = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

(*Hint*: Hitung dengan python ya….)

Lalu buat model menggunakan LinearRegression dari Scikit-Learn yang telah kita pelajari dan bandingkan nilai coef dan intercept yang diperoleh. Apakah kedua metode memberikan nilai yang sama?

Gambarkan plot dari garis yang terbentuk oleh fungsi regresi tersebut bersama dengan scatter plot sebelumnya. Tentukan juga nilai RMSE (Root Mean Squared Error).

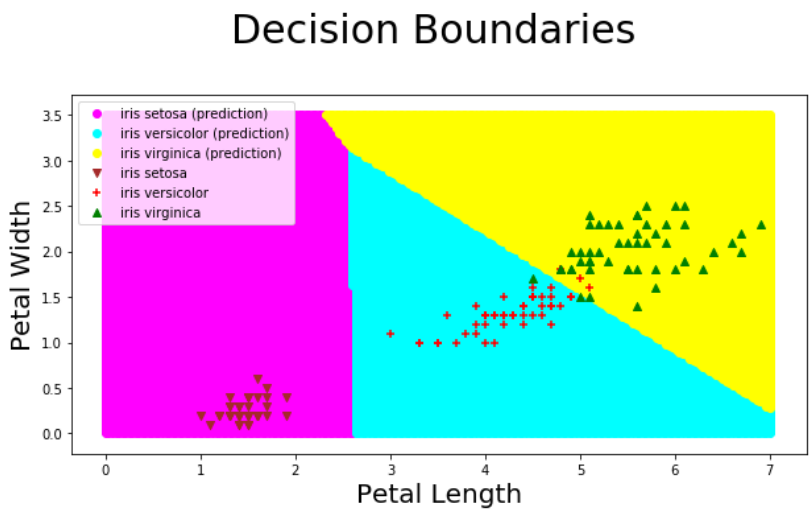
Nomor 2

Diberikan **dataset iris.**

Dataset di atas merupakan data ukuran *sepal* (helai kelopak bunga) dan ukuran *petal* (helai mahkota bunga) dari 150 bunga iris yang terdiri dari tiga spesies yaitu *Iris setosa, Iris versicolor* dan *Iris virginica*. Kita hanya akan menggunakan ukuran panjang (*length*) dan lebar (*width*) dari *petal*-nya untuk mengklasifikasi data ukuran yang diberikan kedalam tiga spesies yang telah disebutkan.

Pisah data menjadi data train dan test dengan perbandingan 7:3, lakukan klasifikasi dengan Multi Logistic Regression. Kemudian buatlah plot dari *decision boundary* dan *scatter plot* dari data-data tes, sumbu x sebagai petal length, sumbu y sebagai petal width.

Berikut contoh plot yang dihasilkan



Cobalah memprediksi apa jenis dari bunga iris yang memiliki ukuran:

1. Petal length = 4 cm, petal width = 2 cm
2. Petal length = 2.4 cm, petal width = 3 cm
3. Petal length = 5.6 cm, petal width = 3.3 cm

Nomor 3

Diberikan **dataset heart attack.**

Kerjakan data secara *end-to-end* dari pengolahan data hingga tentukan model terbaik (error kecil tapi tidak overfitting) untuk memprediksi kemungkinan resiko sakit jantung pada pasien, gunakan metode yang telah dipelajari di praktikum (Logistic Regression, SVM, Decision Tree).

(*Hint*: Periksa apakah data sudah siap digunakan, tentukan *feature-feature* mana yang paling berpengaruh dan untuk kemudian di fit ke dalam model, lakukan transformasi jika diperlukan, kemudian simpanlah model kalian menggunakan pickle atau joblib)

Nomor 4

Diberikan **dataset credit card.**

Keterangan:

Client ID : ID klien

Gender (1=Male, 0=Female)

Own_car (1=Yes, 0=No)

Own_property (1=Yes, 0=No)

Work_phone (1=Yes, 0=No)

Phone (1=Yes, 0=No)

Email (1=Yes, 0=No)

Unemployed (1=Yes, 0=No)

Num_children : jumlah anak

Num_family : jumlah anggota keluarga

Target : label, (1 = berisiko tinggi gagal bayar, 0= berisiko rendah gagal bayar)

Kerjakan data secara *end-to-end* dari pengolahan data hingga tentukan model terbaik (error kecil, tetapi tidak overfitting) untuk memprediksi kemungkinan resiko gagal bayar pada aplikan kartu kredit, gunakan metode yang telah dipelajari di praktikum (Logistic Regression, SVM, Decision Tree).

(*Hint*: Periksa apakah data sudah siap digunakan, tentukan *feature-feature* mana yang paling berpengaruh dan untuk kemudian di fit ke dalam model, lakukan transformasi jika diperlukan, kemudian simpanlah model kalian menggunakan pickle atau joblib.)

Nomor 5

Sebuah perusahaan mobil lokal asal Bandung bercita-cita untuk memasuki pasar AS dengan mendirikan pabrik mereka di sana dan memproduksi mobil secara lokal untuk menjadi pendatang baru pada pasar di AS dan Eropa.

Mereka telah menghubungi sebuah perusahaan konsultan mobil untuk memahami faktor-faktor yang menjadi dasar penentuan harga mobil. Berikut adalah **data yang telah dikumpulkan.** Secara khusus, mereka ingin memahami faktor-faktor yang mempengaruhi harga mobil di pasar AS yang mungkin sangat berbeda dari pasar di Bandung.

Buatlah tiga model berbeda untuk memprediksi harga mobil. Gunakan minimal lima *feature* untuk membuat sebuah model. Pisah data menjadi data *train* dan *test* dengan perbandingan 8:2. Tentukan satu model terbaik dari model yang telah anda buat dan jelaskan alasannya. Kemudian simpanlah model kalian menggunakan pickle atau joblib.

Catatan :

Symboling : +3 = mobil berisiko, -3 = mobil cukup aman.

Nomor 6

Cauchy dan Riemann bermain tic tac toe. Cauchy sebagai “x” mendapat giliran pertama, sedangkan Riemann sebagai “o” mendapat giliran berikutnya. Agar bisa menang melawan Riemann, Cauchy membuat kemungkinan konfigurasi papan tic-tac-toe. Berikut adalah **data yang dibuat oleh Cauchy** dengan targetnya adalah "menang untuk x" (yaitu, bernilai *positive* ketika "x" berhasil membentuk salah satu dari 8 cara yang mungkin untuk membuat "*three in a row*").

Keterangan : x = Cauchy, o = Riemann, b = kosong

V1 = kotak kiri atas

V2 = kotak tengah-atas

V3 = kotak kanan atas

V4 = kotak kiri tengah

V5 = kotak tengah-tengah

V6 = kotak tengah-kanan

V7 = kotak kiri bawah

V8 = kotak tengah-bawah

V9 = kotak kanan bawah

Kelas = menang untuk x {positif,negatif}

Buatlah sebuah model *Decision Tree* berdasarkan dataset yang telah diberikan dengan f1-score minimal 90%. Visualisasikan juga *Decision Tree* yang telah anda buat.