Final Project Big Data Analytics

Buatlah analisis terhadap dataset yang Anda dapatkan dari internet.

Keterangan: File dataset bisa dicari di Kaggle.com (http://kaggle.com) atau di UCI Machine Learning (https://archive.ics.uci.edu/ml/index.php)

Prosedur analisis yang dilakukan adalah sbb:

1. Berdasarkan dataset yang dipilih, buatlah Scatter Plot untuk masing-masing features [**Score: 15**]
2. Berdasarkan pola data dari scatter plot (No. 1), pilih 2 features yang memiliki dugaan kuat ada korelasi. Narasikan alasannya! [**Score: 10**]
3. Buatlah statistik deskriptif dari beberapa 2 features yang dipilih. Berikan narasi terhadap statistik tersebut [**Score: 20**]
4. Dari 2 features yang dipilih tersebut (No. 3), buatlah analisis model regresi (linear atau nonlinear). Jenis model regresi yang dipilih harus menyesuaikan pola sebaran data dari scatter plot. Model yang telah dibuat harus ditunjukkan secara visual dg grafik, serta tunjukkan tingkat keakuratannya (bisa menggunakan sembarang metrik (suplemen/ModelEvaluation-example.ipynb)) [**Score: 25**]

   Sebagai bahan referensi:

   - Contoh analisis model regresi linear (suplemen/LinearRegression-example.ipynb)
   - Contoh analisis model regresi nonlinear (suplemen/NonLinearRegression-example.ipynb) atau polinomial (suplemen/PolynomialRegression-example.ipynb)
5. Terhadap dataset yang diperoleh, lakukan pula analisis berikutnya (pilih salah satu): [**Score: 30**]

   - Association (lihat contoh (suplemen/AssociationRule-example.ipynb))
   - Clustering (lihat contoh (https://mubaris.com/posts/kmeans-clustering/))
   - Classification (lihat contoh (https://towardsdatascience.com/solving-a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2))

     Keterangan: Problem statement yang dipilih untuk analisis di atas bisa sembarang.

Ketentuan Pengerjaan Project

- File dataset yang digunakan harus dideskripsikan (data tentang apa dsb), dan disebutkan URL sumber dari mana mendapatkannya
- Penyimpanan file dataset dan komputasi analisisnya (dengan Jupyter Notebook) harus menggunakan layanan AWS (S3 dan Sagemaker)
- Laporan project dibuat dengan Jupyter Notebook dan di setiap langkahnya harus diberikan narasi
- Untuk setiap analisis yang dilakukan harus disebutkan 'problem statement' nya

- Sebelum proses analisis data, terlebih dahulu harus dilakukan langkah pre-processing (data cleaning dan data preparing) jika diperlukan
- Durasi pengerjaan project mulai tanggal 16/11/2019 s/d 18/11/2019. Selama dalam waktu pelatihan DTS, kerjakan project ini di lab. Di luar pelatihan DTS, project bisa dilanjutkan di rumah/kost
- File project Jupyter Notebook yang dihasilkan jika sudah selesai, maka diupload ke Github (diset ke public) untuk portofolio
- Hasil project dipresentasikan pada tanggal 18-19/11/2019