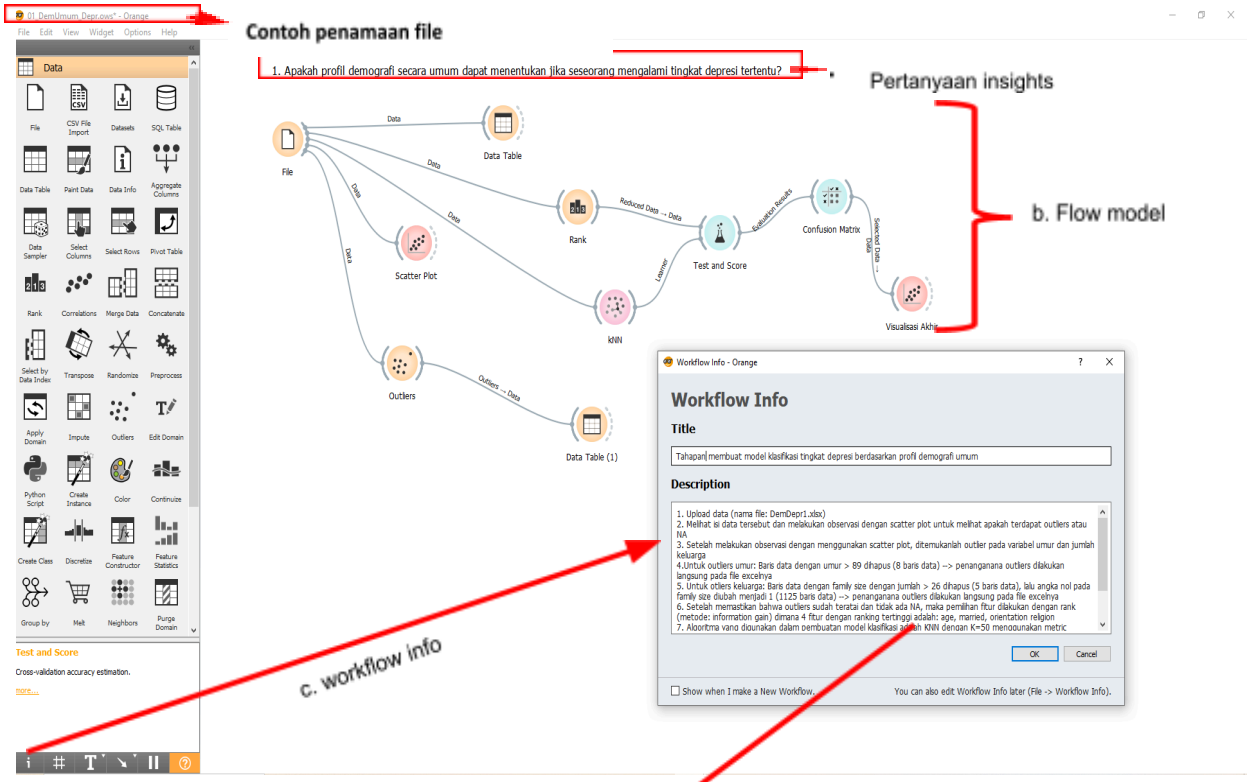
Tugas ke-4 Data Science pada Domain Spesifik
Membuat Pemodelan, Data Visualisasi dan Dashboard
dari Dataset hotel_booking.csv
Dosen Kelas: Lydia Mutiara Dewi, MA

Petunjuk pengerjaan:

1. Dengan menggunakan software Orange / Python / Excel, buatlah 6 buah pemodelan untuk **dataset hotel_booking.csv**
   Keterangan:
   - Model dapat berupa klasifikasi atau clustering. Minimal harus terdapat 3 model klasifikasi, entah menggunakan software Orange atau Python
   - Selain klasifikasi, anda bisa membuat pemodelan dengan menggunakan Pivot Table seperti yang anda buat di dataset ke-1 (Sebelum UTS)
2. Anda bisa mengulik dari segi fitur, algoritma (K-NN, Decision Tree, dll) .
3. Setiap pemodelan harus mengandung komponen-komponen sebagai berikut:
   a. Pertanyaan insights
   b. Flow model / coding / Pivot table pada Excel
   c. Workflow info - jika anda menggunakan Orange (mulai dari data di upload sampai dengan nilai precission dan recall atau kriteria pengukuran yang lain jika anda tidak menggunakan precission atau recall):



Contoh isi workflow info:

1. Upload data (nama file: DemDepr1.xlsx) sebutkan nama kolomnya jika perlu
2. Melihat isi data tersebut dan melakukan observasi dengan scatter plot untuk melihat apakah terdapat outliers atau NA
3. Setelah melakukan observasi dengan menggunakan scatter plot, ditemukanlah outliers pada variabel umur dan jumlah keluarga
4.Untuk outliers umur: Baris data dengan umur > 89 dihapus (8 baris data) --> penanganana outliers dilakukan langsung pada file excelnya
5. Untuk otliers keluarga: Baris data dengan family size dengan jumlah > 26 dihapus (5 baris data), lalu angka nol pada family size diubah menjadi 1 (1125 baris data) --> penanganana outliers dilakukan langsung pada file excelnya
6. Setelah memastikan bahwa outliers sudah teratasi dan tidak ada NA, maka pemilihan fitur dilakukan dengan rank (metode: information gain) dimana 4 fitur dengan ranking tertinggi adalah: age, married, orientation religion
7. Algoritma yang digunakan dalam pembuatan model klasifikasi adalah KNN dengan K=50 menggunakan metric euclidean uniform
8. Model kemudian dijalankan (test and score) dengan menggunakan random sampling, repeat train set = 10, dan komposisi training set 80%

Keterangan: workflow di atas hanya sekedar ilustrasi yang menggunakan dataset lain

d. Jika anda menggunakan Python / Pivot Table, sebagai pengganti workflow, jelaskan langkah-langkah pengerjaannya.
e. Untuk setiap model, tampilkan **visualisasi yang anda anggap sesuai.**

4. Jika anda sudah mengerjakan poin 3, buatlah sebuah **folder** dengan nama : **T4M_DSDSA_NoKel**
5. Gantilah komponen NoKel pada folder tersebut dengan No Kelompok anda yang terdaftar di google classroom, contoh: **T4M_DSDSA_NoKel** menjadi **T4M_DSDSA_01**
6. Masukkan file-file model ke dalam folder yang sudah anda buat (file apapun, apakah orange, python atau excel), dengan penamaan sebagai berikut:

   MX_NamaModel_YY

   Keterangan:
   X : no model
   YY: 2 digit no kelompok

   Contoh:
   Model 1: M1_KlasifikasiDT_01.ows file berjenis orange
   Model 2: M2_KlasifikasiNBC_01.py file berjenis python
   Model 3: M3_KlasifikasiKNN_01.xlsx file berjenis excel
   Dan seterusnya sampai model ke-6

7. Dari 6 model yang sudah anda buat, pilihlah 4 **buah model terbaik** beserta alasannya (perbesar sendiri ukutan kotak pada table di bawah ini jika dirasa perlu):

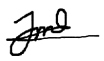
No	Nama File	Pengukuran (mis: Precision / Recall/ F1 / Accuracy)*	Informasi yang anda peroleh dari model ini contoh: dari model ini didapat informasi bahwa tamu dengan lead time …. Cenderung tidak membatalkan pesanan.
1	M3_KlasifikasiKNN_01.py	Accuracy: 0.7556884167060949 Precision: 0.6701030927835051 Recall: 0.6408124480342083	Informasi yang kami peroleh dari model ini : Pembatalan pemesanan tampaknya cukup umum, mengingat recall model sebesar 64.08%. Hal ini menunjukkan bahwa ada sejumlah pemesanan yang dibatalkan di hotel tersebut, dan model mampu mendeteksi sebagian besar dari pembatalan tersebut. Selain itu, model memiliki akurasi yang cukup tinggi sebesar 75.56%, presisi sebesar 67.01% mengindikasikan bahwa sebagian besar prediksi model benar, precision yang lebih tinggi mengurangi kesalahan ketika model memprediksi pembatalan.
2	M1_KlasifikasiDTree_01.ows	Accuracy: 0.585 F1-Score:0.582	Dari Hasil model yang ada, dapat dilihat bila dilihat dari pengukuran, model

		Precision: 0.783 Recall: 0.673	cukup dapat mendeteksi atau memisahkan pemesanan yang dibatalkan berdasarkan Previous_Cancellation dan Previous_booking_not_canceled. Bila divisualisasikan juga bahwa semakin tinggi nilai previous_booking_not_canceled, maka kemungkinan sebuah pemesanan kamar tidak dibatalkan semakin tinggi, tapi bila nilai Previous_Cancellation tinggi, maka kemungkinan pemesanan kamar dibatalkan cukup tinggi juga. Ada kemungkinan kecil dimana jika jumlah previous_cancellations nya lebih dari 2, maka booking akan dibatalkan dan juga jika perbandingan antara previous_cancellations dan previous_booking_not_canceled nya cukup jauh berbeda. Sehingga dapat disimpulkan bahwa Previous_Cancellation dan Previous_booking_not_canceled cukup berpengaruh untuk melihat pemesanan dibatalkan atau tidak
3	M4_ClusteringKmeans_01.py	-	Informasi yang kami peroleh dari model ini : **Cluster 1 (Short Stay Short Lead Time)** cenderung memiliki risiko pembatalan yang moderat. Hal ini mungkin terjadi karena tamu dengan lead time yang singkat (pemesan kamar dalam waktu dekat dengan tanggal kedatangan) cenderung membuat keputusan pemesanan yang lebih cepat dan impulsif yang mana berpotensi lebih mudah membatalkan jika ada perubahan rencana mendadak dan juga mungkin mereka memiliki jadwal yang lebih fleksibel, sehingga lebih mungkin untuk mencari alternatif lain jika ada kesempatan yang lebih baik atau lebih murah. **Cluster 1 (Long Stay Long Lead Time)** cenderung memiliki risiko pembatalan yang lebih rendah. Tamu dalam kelompok ini mungkin lebih yakin dengan rencana perjalanan mereka karena mereka memesan jauh-jauh hari dan menginap lebih lama.

4	M6_ClusteringK-Means_01.py	Silhouette Score untuk k=4 setelah t-SNE: 0.5432828664779663	Cluster 0 berisi customer yang cenderung melakukan booking dari jauh hari yang diindikasikan dengan median dan Inter Quartil Range yang lebih lebar. Cluster 1 dan 2 menunjukkan lead time yang sangat pendek, berarti booking yang terburu-buru (Last Minute). Segmen ini dapat ditarget dengan promo-promo last minute maupun policy yang lebih fleksibel. Cluster 3 memiliki lead time yang umum, mengindikasikan kestabilan waktu membooking (tidak terlalu cepat maupun terlalu lama).

**Jika model yang anda pilih tidak mengandung pengukuran beri tanda: -*

8. Jika sudah, isikan tabel kontribusi kelompok di bawah ini (terurut **secara ascending** berdasarkan NPM):

NPM	Nama	Kelas	Kontribusi (%)	Tanda Tangan
6182001001	Jenson Mark Lowell	A	100%	
6182101040	Samuel Edward Winoto		100%	
6182101054	Steffi Widjaya		100%	

9. Ubahlah **file ini ke dalam bentuk pdf dan masukkan** juga ke dalam folder **T4M_DSDSA_NoKel** yang tadi sudah anda buat
10. Lihat kembali poin no 3e. Di no 3e, anda sudah membuat 6 buah model lengkap dengan visualisasinya. **Dari 6 visualisasi yang sudah anda buat, pilihlah 4 visualisasi, screenshot visualisasi tersebut dan buatlah dashboard pada file Excel bernama: T4_DB_NoKel.xlsx. Lihat petunjuk contoh layout dashboard yang berada di dalam file ini.**
11. Jika sudah, unggahlah pekerjaan ke dalam google classroom. Lihat no 12 untuk mengetahui file-file apa saja yang harus anda unggah.
12. Jadi, di dalam folder **T4M_DSDSA_NoKel**, nantinya akan terdapat:
    - 6 buah file pemodelan yang sudah anda buat
    - **File ini** yang sudah anda ubah ke dalam bentuk pdf (**T4_DSDSA_Nokelompok.pdf**)
    - File Excel yang berisi dashboard yang berisi screenshot visualisasi, informasi, insights, dan actionable insights bernama **T4_DB_NoKel.xlsx.**
    - Dataset asli yang anda gunakan untuk membuat pemodelan pada Orange / Python / Excel.