Final Project - Probability Course

Sekolah Data - Pacmann

OUTLINE

—

Background Project

Salah satu peran dari seorang Data Analyst adalah mampu menarik insight atau informasi sederhana pada data. Terkadang, melalui eksplorasi dataset saja sudah cukup membantu pengambilan keputusan bisnis daripada melakukan model yang kompleks. Melalui project ini, anda akan diminta untuk menganalisa variable-variabel pada data yang anda minati dengan menggunakan ilmu dasar probability. Anda juga boleh mencantumkan visualisasi data apabila diperlukan. Berikan hasil ulasan dari analisa yang anda lakukan sehingga bisa menjadi salah satu pembantu keputusan bisnis pada stakeholder terkait.

Catatan:

- Project **dapat** dikerjakan & **harus** dikumpulkan secara **individu.**
- Anda diperbolehkan untuk membentuk kelompok diskusi selama pengerjaan project.
- Dalam kelompok diskusi
    - **Anda dapat berdiskusi tentang**
        - Pemilihan topik & dataset
        - Alur pengerjaan project
        - Analisa hasil
    - **Anda tidak diperbolehkan untuk**
        - Memiliki easy report, presentation & rekaman, dan code (notebook) yang sama dengan sesama tim
- Khusus untuk **Career Program**, kelompok sudah ditentukan lewat small group discussion (SGD)
- Untuk **Skill Program**, Anda dapat mencari kelompok secara mandiri dengan sheet yang telah disiapkan oleh tim Pacmann
- Ingat, hasil project **dikumpulkan secara individu** karena akan menjadi **portofolio pribadi**.

Petunjuk Pengerjaan

Langkah #1 - Pilih Dataset dan Tentukan Objective

Expected time to start: Week 1
Expected time of completion: End of Week 1

1. **Dataset**
   Hal yang pertama Anda lakukan adalah memilih dataset. Anda dibebaskan untuk memilih dataset dari berbagai sumber baik itu dari website seperti Kaggle maupun dari tempat kerja Anda. Hal yang perlu diperhatikan adalah sifat confidential dari dataset-nya, jika Anda bertujuan untuk menjadikan project ini sebagai portofolio maka sebaiknya tidak menggunakan dataset yang bersifat confidential.

   Pastikan Anda cukup memahami dataset-nya sehingga Anda dapat melakukan analisa dari dataset yang didapat. Beberapa hal yang dapat membantu dalam pemilihan adalah sebagai Berikut:
   - Seberapa paham Anda terhadap fitur-fitur yang ada pada dataset?
   - Seberapa menarik insight yang bisa Anda dapat dan bahas dari dataset tersebut?

   Jika Anda bingung akan menggunakan dataset apa, berikut kami beri contoh dataset yang mungkin bisa Anda gunakan:
   - Data Tagihan Kesehatan.
     - Analisa variable-variabel yang memiliki hubungan dengan tagihan kesehatan
     - Membantu memberikan gambaran karakteristik pengguna yang seperti apa yang memiliki tagihan kesehatan yang tinggi untuk pertimbangan pihak asuransi untuk menentukan nilai premi.

2. **Objective**
   Selanjutnya, tentukan tujuan analisis yang ingin dilakukan. Objective akan membantu anda memberikan gambaran fokus analisis yang ingin dituju.

   Contoh, anda ingin melakukan analisa data tagihan kesehatan, objektif yang bisa digunakan:
   - Mengetahui statistik deskriptif dari pengguna (misal rata rata tagihan kesehatan dari variable-variabel yang berkaitan)
   - Mencari peluang kondisi tertentu terhadap tagihan kesehatan (misal mana yang lebih besar peluang nya, perokok yang memiliki tagihan tinggi atau orang yang obesitas)
   - Mencari variabel yang memiliki hubungan paling kuat dengan tagihan kesehatan
   - Menguji klaim tentang tagihan kesehatan

Langkah #2 - Lakukan Analisa

Expected time to start: Week 2
Expected time of completion: End of Week 8

Gunakan ilmu yang telah anda dapatkan di kelas probability, anda **diharapkan** dapat melakukan **analisa secara saintifik** untuk menarik insight pada variabel-variabel yang ada di data. Di bagian ini, diberikan petunjuk langkah-langkah analisa yang bisa dilakukan. Langkah-langkah ini bertujuan untuk membantu menjawab objektif pada data yang ingin dicapai.

Petunjuk Analisa

Untuk mempermudah dan memperdalam analisa, berikut adalah hal-hal komprehensif yang dapat Anda lakukan.

#1 - Analisa Descriptive Statistic

Expected time to start: Week 4
Expected time of completion: End of Week 4

Kita awali proses analisa ini dengan hal yang paling dasar, yakni merangkum karakter-karakter berdasarkan data seperti mencari rata-rata & persebaran data.

Materi pertemuan : 7 & 8

#2 - Analisa Peluang pada Variabel Diskrit

Expected time to start: Week 2
Expected time of completion: End of Week 5
Selanjutnya, untuk memperdalam analisa, Anda dapat mengidentifikasi peluang kondisi tertentu yang berpotensi memiliki besaran variabel tertentu.

Materi pertemuan : 3 - 10

#3 - Analisa Peluang pada Variabel Kontinu

Expected time to start: Week 6
Expected time of completion: End of Week 6
Variabel dalam data yang kita punya bisa jadi tidak semuanya berbentuk diskrit, untuk memahami kemungkinan kondisi variabel bernilai kontinu pada data, kita bisa melakukan analisa pada data tersebut.

> *Note: Anda dibebaskan memilih cara untuk menghitung peluang baik menggunakan asumsi distribusi atau menggunakan pendekatan diskrit (binning interval pada variabel kontinu).*

Materi pertemuan : 11 & 12

#4 - Analisa Korelasi Variabel

Expected time to start: Week 7
Expected time of completion: End of Week 7
Setelah menjawab kondisi-kondisi yang lebih mungkin dari langkah sebelumnya. Kita juga dapat mencari keterhubungan antara kondisi-kondisi tersebut dengan variabel yang kita inginkan. Analisa korelasi akan diperlukan disini.

Materi pertemuan : 13 & 14

#5 - Pengujian Hipotesis

Expected time to start: Week 8
Expected time of completion: End of Week 8
Langkah terakhir, kita cari apakah ada bukti statistik yang cukup terhadap klaim atau hipotesis pada data kita

Materi pertemuan : 15 & 16

Berikut adalah **contoh** detail langkah pada dataset asuransi kesehatan sebagai referensi yang bisa dilakukan di setiap langkah-langkahnya.

#1 - Analisa Descriptive Statistic

Contoh, anda bisa memilih 5 pertanyaan dibawah ini untuk melakukan eksplorasi awal pada data tagihan kesehatan.

1. Berapa rata rata umur pada data tersebut?
2. Berapa rata rata nilai BMI dari yang merokok?
3. Apakah variansi dari tagihan kesehatan perokok dan non perokok sama?
4. Apakah rata rata umur perempuan dan laki-laki yang merokok sama?
5. Mana yang lebih tinggi, rata rata tagihan kesehatan perokok atau non merokok?
6. Mana yang lebih tinggi, rata rata tagihan kesehatan perokok yang BMI nya diatas 25 atau non perokok yang BMI nya diatas 25 (overweight)
7. BMI mana yang lebih tinggi, seseorang perokok atau non perokok?

#2 - Analisa Variabel Diskrit

Anda bisa memilih 5 pertanyaan dibawah ini untuk pengecekan kondisi pada data.

1. Gender mana yang memiliki tagihan paling tinggi?
2. Distribusi peluang tagihan di tiap-tiap region
3. Apakah setiap region memiliki proporsi data banyak orang yang sama?
4. Mana yang lebih tinggi proporsi perokok atau non perokok?
5. Berapa peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok?
6. Berapa peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok?
7. Bagaimana bentuk distribusi peluang besar tagihan dari tiap-tiap region?

#3 - Analisa Variabel Kontinu

Anda bisa menggunakan 2 pertanyaan dibawah ini untuk pengecekan kondisi pada data tagihan kesehatan.

1. Mana yang lebih mungkin terjadi
    a. Seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k, atau
    b. Seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k
2. Mana yang lebih mungkin terjadi
    a. Seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k, atau
    b. Seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k

#4 - Analisa Korelasi Variabel
Anda bisa memilih mengecek korelasi tagihan kesehatan minimal dengan 2 variabel lainnya, misalnya dengan bmi dan variable tanggungan anak.

#5 - Pengujian Hipotesis
Anda bisa mengecek 3 hipotesis tentang karakter populasi dari data. Hipotesis bisa anda pilih adalah

1. Tagihan kesehatan perokok lebih tinggi daripada tagihan kesehatan non perokok
2. Proporsi perokok laki laki lebih besar dari perempuan
3. Variansi tagihan kesehatan perokok dan non perokok sama
4. Tagihan kesehatan dengan BMI diatas 25 lebih tinggi daripada tagihan kesehatan dengan BMI dibawah 25
5. Tagihan kesehatan laki-laki lebih besar dari perempuan

—

Setelah melalui 5 langkah ini, Anda akan mendapatkan bahan dalam menjawab kondisi atau faktor dari pengguna asuransi kesehatan yang memiliki hubungan dengan besar tagihan.

Langkah #3 - Buat Report dan Video Presentasi

Expected time to start: Week 2
Expected time of completion: End of Week 8
Setelah Anda mengerjakan per langkahnya, kami ingin Anda dapat merangkum analisa & menuliskan hasilnya dalam sebuah **short report.** Anda juga diminta melakukan **presentasi penjelasan teori**. Buatlah short report serta cantumkan file pengerjaan didalamnya dan rekam presentasi penjelasan salah satu teori yang anda gunakan (misalnya conditional probability) melalui youtube. Berikan link medium (short report) dan link youtube (presentasi penjelasan teori) ke dalam form submission akhir.

Silahkan mencicil membuat report dimulai dari **Week 2,** Berikut gambaran timeline pengerjaan yang bisa anda ikuti

Short Report

- *Introduction (Week 2 - End of Week 2)*
- *Research question (Week 3 - End of Week 8)*
- *Conclusion, Further Research, Link Pengerjaan (Week 8 - End of Week 8)*

Presentasi Teori

- Bisa dimulai kapan saja sesuai tergantung apakah teori yang anda pilih sudah diajarkan di kelas

Note - Cicil Progres!

Anda bisa mulai mencicil mengerjakan progres project dimulai dari minggu pertama!.

- Cicil langkah pada petunjuk pengerjaan dan cicil outcome project sesuai dengan ekspektasi pengerjaan dapat dimulai (perhatikan Expected time to start dan Expected time of completion)
- Anda dapat mengupload progres pengerjaan project dalam submission yang diberikan

Outcome Project

1. Short report Di upload ke **Medium**
    a. Buatlah short report dalam bentuk artikel pendek
    b. Contoh short report bisa dilihat disini (bagian Spring 2020/2021/2022 -> Stats Graphics & Visualization) atau disini.
    c. Outline short report bisa berbentuk sebagai berikut:
        i. Introduction
        Isi dengan tujuan project atau fokus analisa yang ingin anda eksplor, dan ceritakan dataset yang anda akan pakai untuk analisis.
        ii. Research Question
        Isi dengan pertanyaan yang ingin anda jawab, uraian jawabannya, beserta insight atau hasil analisis yang anda dapatkan.
        iii. Conclusion
        Isi dengan temuan menarik dari keseluruhan pertanyaan yang telah dijawab.
        iv. Further Research
        Sampaikan saran perbaikan (jika ada) untuk pengerjaan yang telah dilakukan.
        v. Reference
        Cantumkan referensi yang anda pakai untuk membantu pengerjaan
        vi. Link pengerjaan
            1. Buatlah sebuah repository di github anda (anda bisa memakai google drive jika belum familiar dengan github).
            2. Simpan hasil pengerjaan anda ke dalam repository tersebut berupa File **code python, file excel, atau dokumen pendukung** apapun yang digunakan untuk analisa.

2. Link **Youtube** Presentasi (Theory Explanation)
    a. Record penjelasan anda tentang **salah satu teori probabilitas** yang sudah diajarkan dan anda gunakan dalam analisa pada slide presentasi dalam durasi maksimal 5 menit. Anda bisa memilih satu teori saja dari pilihan berikut:
        - Bayes Theorem
        - Covariance & Correlation
        - Hypothesis Testing (Pilih 1 Uji Statistik yang digunakan di Langkah #5)
    b. Berisi:
        - Pengenalan diri
        - Penjelasan Teori
    c. Di upload ke Youtube.
        - Judul: Probability - [Judul teori yang anda pilih]
        - Permission: Set access nya menjadi publik agar dapat tim Pacmann periksa.

Evaluasi

Kami akan mengevaluasi beberapa komponen berikut. Dengan fokus memeriksa ketepatan pengerjaan & analisa yang dihasilkan.

Komponen/*Grading Criteria*	**Poin maksimum**
Short Report	**75 poin**
Langkah #1: Analisa Descriptive Statistic - *Ketepatan cara pengerjaan* - *Analisa yang didapatkan dari jawaban pertanyaan-pertanyaan*	15 poin
Langkah #2: Analisa Variabel Kategorik (PMF) - *Ketepatan cara pengerjaan* - *Analisa yang didapatkan dari jawaban pertanyaan-pertanyaan*	15 poin
Langkah #1: Analisa Variabel Kontinu - *Ketepatan cara pengerjaan* - *Analisa yang didapatkan dari jawaban pertanyaan-pertanyaan*	15 poin
Langkah #1: Analisa Korelasi Variabel - *Ketepatan cara pengerjaan* - *Analisa yang didapatkan dari jawaban pertanyaan-pertanyaan*	15 poin
Langkah #5: Pengujian Hipotesis - *Ketepatan cara pengerjaan* - *Pengambilan kesimpulan pada setiap uji klaim*	15 poin
Presentation (Theory Explanation)	**25 poin**
- Komunikasikan salah satu teori yang digunakan pada project dengan intuitif dan ringkas	25 poin

Need Assistance?

Tentu project ini menantang!
Jika anda memiliki pertanyaan atau kesulitan dalam mengerjakan project ini, anda bisa memanfaatkan fasilitas Asistensi Via discord tag asisten.

Dataset & Tools

Dataset

1. Berikut adalah beberapa sumber dataset yang dapat Anda gunakan untuk membantu pemilihan dataset
   Open Data Jakarta
   Open Data Jawa Barat
   Badan Pusat Statistik Indonesia
   Satu Data Indonesia
   UCI Machine Learning Repository
   Kaggle
   Data World

2. Jika anda bingung anda bisa menggunakan dataset yang disediakan, data tagihan kesehatan personal. Data ini memiliki 7 variable dengan variable **charges** menunjukkan besaran tagihan kesehatan. Deskripsi setiap kolom dari dataset adalah sebagai berikut:
   - **age**
     Age of primary beneficiary
   - **sex**
     Insurance contractor gender, female, male
   - **bmi**
     Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9
   - **children**
     Number of children covered by health insurance / Number of dependents
   - **smoker**
     Smoking
   - **region**
     The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
   - **charges**
     Individual medical costs billed by health insurance

Tools

Anda dibebaskan untuk menggunakan tools apa saja untuk melakukan perhitungan, analisa, dan plotting data.

- Python
- Excel
- Atau lainnya