Final Project - Probability Course

Sekolah Data - Pacmann

OUTLINE

—

Background Project

Asuransi kesehatan adalah salah satu hal yang patut diperhatikan karena bersangkutan dengan kebutuhan perencanaan masa depan. Pengguna asuransi kesehatan diwajibkan untuk membayar besaran uang secara rutin (premi) kepada pihak perusahaan asuransi. Premi tersebut diolah oleh perusahaan asuransi untuk membayarkan tagihan kesehatan pengguna yang tertanggung. Penentuan nilai premi menjadi tantangan tersendiri bagi pihak asuransi mengingat ada banyak faktor yang dapat mempengaruhi & meningkatkan profil resiko pengguna.

Melalui project ini, Anda akan diminta untuk membantu menganalisa variable-variabel yang memiliki hubungan dengan tagihan kesehatan yang diterima oleh setiap pengguna. Anda akan diberikan data yang berisi data personal pengguna seperti umur, gender, tempat tinggal pengguna, banyak anak tertanggung asuransi, nilai bmi, keadaan merokok atau tidaknya pengguna.

Petunjuk Analisa

Dengan menggunakan dasar ilmu probability, Anda **diharapkan** dapat melakukan **analisa secara saintifik** untuk mencari variabel-variabel pengguna yang berhubungan dengan tagihan kesehatan.

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

Langkah #1 - Analisa Descriptive Statistic

Kita awali proses analisa ini dengan hal yang paling dasar, yakni merangkum karakter-karakter berdasarkan data seperti mencari rata-rata & persebaran data. Anda bisa memilih 5 pertanyaan dibawah ini untuk melakukan eksplorasi data. Beberapa hal yang dapat Anda jawab adalah

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

Materi pertemuan : 7 - 12

Langkah #2 - Analisa Variabel Kategorik (PMF)

Selanjutnya, untuk memperdalam analisa, Anda dapat mengidentifikasi peluang kondisi tertentu yang berpotensi memiliki besaran tagihan kesehatan tertentu. Anda bisa memilih 5 pertanyaan dibawah ini untuk pengecekan kondisi pada data. Beberapa hal yang dapat Anda jawab adalah

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 tagihan dari tiap-tiap region?

Materi pertemuan : 1 - 8

Langkah #3 - Analisa Variabel Kontinu (CDF)

Variabel dalam data yang kita punya tidak semuanya berbentuk kategorik, untuk memahami kemungkinan kondisi variabel bernilai kontinu terhadap tagihan kesehatan, kita bisa melakukan analisa CDF pada data. Beberapa hal yang dapat Anda jawab adalah

1. Mencari peluang besar tagihan berdasarkan BMI
2. Mencari kemungkin terjadi, seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.700.
3. Berapa peluang seseorang acak tagihan kesehatannya diatas 16.7k diketahui dia adalah perokok
4. 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
5. 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

Materi pertemuan : 9 - 12

Langkah #4 - Analisa Korelasi Variabel

Setelah menjawab kondisi-kondisi yang lebih mungkin memiliki tagihan kesehatan yang tinggi dari langkah sebelumnya. Kita juga dapat mencari keterhubungan antara kondisi-kondisi tersebut dengan tagihan kesehatan. Analisa korelasi akan diperlukan disini.

Materi pertemuan : 13 & 14

Langkah #5 - Pengujian Hipotesis

Langkah terakhir, kita cari apakah ada bukti statistik yang cukup terhadap klaim atau hipotesis tentang tagihan kesehatan. Anda wajib mengecek 3 hipotesis tentang karakter populasi dari data. Hipotesis yang **wajib** uji adalah

1. Tagihan kesehatan perokok lebih tinggi daripada tagihan kesehatan non perokok
2. Tagihan kesehatan dengan BMI diatas 25 lebih tinggi daripada tagihan kesehatan dengan BMI dibawah 25

Satu hipotesis lain, anda bisa ***pilih salah satu*** hipotesis dibawah ini, atau ***anda dapat membuat hipotesis lainnya***

1. BMI laki-laki dan perempuan sama
2. Tagihan kesehatan laki-laki lebih besar dari perempuan
3. Proporsi perokok berbeda di tiap region

Materi pertemuan : 15 & 16

—

Setelah melalui 5 langkah ini, Anda akan mendapatkan bahan untuk melakukan analisa mendalam serta dapat menjawab kondisi atau faktor dari pengguna asuransi kesehatan yang mempengaruhi besar tagihan.

Outcome Project

Setelah Anda mengerjakan itu semua, kami ingin Anda dapat melakukan analisa & merangkum hasilnya dalam sebuah **short report** & **presentasi**. Simpan short report beserta file pengerjaan dalam tautan repository github dan rekam presentasi anda melalui youtube. Berikan link repository project dan link youtube presentasi ke dalam form submission.

1. Repository Project
    a. Buatlah sebuah repository di github anda.
    b. Simpan hasil pengerjaan anda ke dalam repository tersebut berupa:
        i. Short report dalam bentuk slide presentasi.
        ii. File code python, file excel, atau dokumen pendukung apapun yang digunakan untuk analisa.
    c. Anda bisa menggunakan format slide presentasi yang dapat diunduh disini. Tercantum juga cara pengisian untuk outline presentasinya. Anda boleh menggunakan template slide lainnya, asalkan mengikuti format outline sebagai berikut:
        i. Introduction
        ii. Question and Answer
        iii. Hasil Analisa
        iv. Kesimpulan
        v. Saran Perbaikan
        vi. Referensi
2. Link youtube Presentasi
    Record penjelasan anda tentang hasil project pada slide presentasi yang sudah anda kerjakan dalam durasi 10 menit.

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 (CDF) - *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	**25 poin**
- Komunikasikan pengerjaan project dengan hasil analisa secara efektif	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

Dataset yang disediakan adalah [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