**Probability Project Pacmann – Data Insurance Analyst**

1. **Pendahuluan**

Salah satu cara menarik insight atau informasi pada data adalah dengan mengeksplorasi dataset. Harapannya hal ini mampu membantu pengambilan keputusan bisnis tanpa membuat model yang kompleks. Pada projek ini saya akan menganalisa variable-variabel pada data dengan menggunakan ilmu dasar probability.

Dataset yang saya gunakan adalah Insurance dataset. Deskripsi setiap kolom dari dataset adalah sebagai berikut:

* age: usia customer
* sex: gender (male/female)
* bmi: body mass index, ukuran untuk menghitung berat badan ideal berdasarkan tinggi dan berat badan. Idealnya 18.5—24.9
* children: jumlah anak yang dicover dalam asuransi
* smoker: perokok atau non perokok (yes/no)
* region: residential area in the US (northeast, southeast, southwest, northwest)
* charges: jumlah tagihan

Projek ini akan membahas tentang Analisa Deskriptif analisis, Analisa Variabel Diskrit, Analisa Variabel Kontinu, Analisa Korelasi Variabel, serta Pengujian Hipotesis.

1. **Analisa Deskriptif analisis**

Deskriptif analisis merupakan langkah pertama yang dilakukan untuk mengeksplorasi data. Secara sederhana Analisa Deskriptif analisis bisa dilakukan dengan sintaks df.describe() untuk varibel numerik dan df.describe(include=['O']) untuk variabel kategorik.

Berikut adalah beberapa pertanyaan yang bisa dijawab dengan

1. Berapa rata-rata umur pada data? 39 tahun
2. Berapa rata-rata nilai BMI pada data? 30.66
3. Berapakah nilai median dari tagihan? 9386.16
4. Lebih banyak mana data perempuan atau laki-laki? Laki-laki
5. Lebih banyak perokok atau non perokok? Non perokok
6. **Analisa Variabel Diskrit**
7. Gender mana yang memiliki tagihan paling tinggi?

Gender dengan tagihan tertinggi adalah perempuan, namun perbedaannya tidak terlalu signifikan. Hal ini bisa dimaklumi karena secara alami perempuan lebih banyak peluang melakukan pemeriksaan atau pengobatan. Misalnya karena hamil dan melahirkan.

1. Bagaimana Distribusi tagihan di tiap-tiap region?

Secara garis besar distribusi tagihan di setiap region hampir sama. Dari visualisasi kita tahu bahwa southwest memiliki jumlah tertinggi dan northwest dengan jumlah terendah. Apakah ini dipengaruhi oleh proporsi data di tiap-tiap region?

1. Apakah setiap region memiliki proporsi data banyak orang yang sama?

Tidak sama. Southeast memiliki proporsi pling banyak dengan 27.2%. Berarti jumlah tagihan pada pertanyaan sebelumnya tidak terpengaruh dengan proporsi data di tiap region.

1. Berapa peluang seseorang tersebut adalah perempuan jika diketahui dia adalah perokok?

Jumlah perokok perempuan: 115

Jumlah total perokok: 274

Peluang seseorang adalah perempuan jika diketahui dia adalah perokok: 0.42

1. Berapa peluang seseorang tersebut adalah laki-laki jika diketahui dia adalah perokok?

Jumlah perokok laki-laki: 159

Jumlah total perokok: 274

Peluang seseorang adalah perempuan jika diketahui dia adalah perokok: 0.58

1. **Analisa Variabel Kontinu**
2. Mana yang lebih mungkin terjadi: (a) Seseorang dengan BMI diatas 25 (overweight) mendapatkan tagihan kesehatan diatas 16,7k, atau (b) Seseorang dengan BMI dibawah 25 (normal weight) mendapatkan tagihan kesehatan diatas 16,7k

\*\*proporsi\_overweight lebih tinggi, maka kondisi "seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16,7k" lebih mungkin terjadi.\*\*

1. Mana yang lebih mungkin terjadi: (a) Seseorang perokok mendapatkan tagihan kesehatan diatas 16,7k, atau (b) Seseorang non perokok mendapatkan tagihan kesehatan diatas 16,7k

\*\*proporsi pertama lebih tinggi, maka kondisi "seseorang perokok mendapatkan tagihan kesehatan diatas 16,7k" lebih mungkin terjadi.\*\*

1. **Analisa Korelasi Variabel**
2. Korelasi antara Tagihan dengan variabel usia (age), gender (sex), BMI (bmi), dan jumlah anak (children).

Semua berkorelasi positif.

1. **Pengujian Hipotesis**
2. Tagihan kesehatan perokok lebih tinggi daripada tagihan kesehatan non perokok

H0 = Tagihan kesehatan perokok dan non perokok sama.

H1 = Tagihan kesehatan perokok lebih tinggi daripada non perokok.

Dengan melakukan uji t-test didapatkan nilai:

P-value: 1.4067220949376498e-282

alpha: 0.05

------------------------------------

Menolak hipotesis nol.

Tagihan kesehatan perokok lebih tinggi daripada non perokok.

1. Proporsi perokok laki-laki lebih besar daripada perempuan

H0 = Proporsi perokok laki-laki sama dengan perokok perempuan.

H1 = Proporsi perokok laki-laki lebih besar daripada perokok perempuan.

Dengan melakukan uji chi-square didapatkan nilai:

P-value: 0.0062765550120107375

alpha: 0.05

------------------------------------

Menolak H0.

Proporsi perokok laki-laki lebih besar daripada perokok perempuan.

1. Variasi tagihan kesehatan perokok dan non perokok sama

H0 = Variasi tagihan kesehatan perokok dan non perokok sama.

H1 = Variasi tagihan kesehatan perokok dan non perokok berbeda.

Dengan melakukan uji lavane untuk mencasi variasi didapatkan nilai:

P-value: 1.670117565125241e-66

alpha: 0.05

------------------------------------

Menolak H0. Variasi tagihan kesehatan perokok dan non perokok berbeda.

1. Tagihan Kesehatan dengan BMI diatas 25 (overweight) lebih tinggi daripada tagihan kesehatan dengan BMI dibawah 25 (normal weight)

H0 = Tagihan kesehatan dengan BMI di atas 25 sama dengan BMI di bawah 25.

H1 = Tagihan kesehatan dengan BMI di atas 25 lebih tinggi daripada BMI di bawah 25.

Dengan melakukan uji t-test didapatkan nilai:

P-value: 1.6909664264439038e-05

alpha: 0.05

------------------------------------

Menolak H0.

Tagihan kesehatan dengan BMI di atas 25 lebih tinggi daripada BMI di bawah 25.

1. Tagihan kesehatan laki-laki lebih besar dari perempuan

H0 = Tagihan kesehatan laki-laki sama dengan perempuan.

H1 = Tagihan kesehatan laki-laki lebih besar daripada perempuan.

Dengan melakukan uji t-test didapatkan nilai:

P-value: 0.0338207919951195

alpha: 0.05

------------------------------------

Menolak H0.

Tagihan kesehatan laki-laki lebih besar daripada perempuan.

1. **Penutup**

Demikianlah short report Probability Project Pacmann – Data Insurance Analyst. Informasi lengkap tentang projek ini bisa dilihat pada link berikut:

Link GitHub

Link Video YouTube

Link Medium