Tugas 2 Mata Kuliah PADK

Nama: Rheyhan Fahry

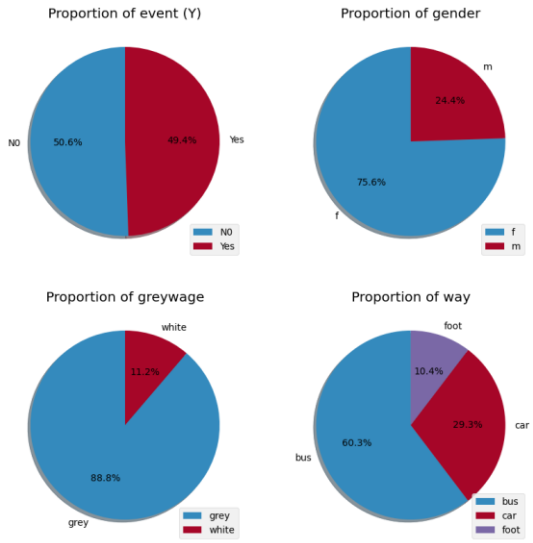
Nim: G1401211030

Menggunakan bahasa pemrograman python, data dan metadata dibaca menggunakan *modules* pandas. Lalu data turnover atau target diubah dari [1, 0] menjadi [“yes”, “no’]. Selanjutnya dilakukan pengecekan dan penghapusan terhadap data bermissing value. Berikut adalah tipe data dan peubah yang digunakan:

Tabel 1 Peubah dan tipe data yang dipakai

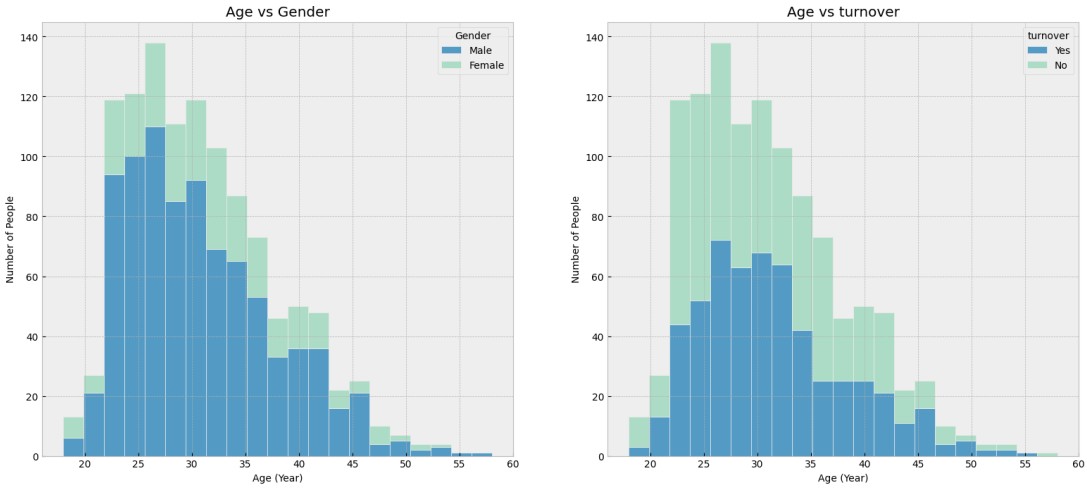
Peubah	Keterangan	Tipe Data
Event (Y)	Kejadian Turnover	*Object*
Gender	Jenis Kelamin	*Object*
Age	Usia	*Float64*
Industry	Bidang Pekerjaan	*Object*
Profession	Profesi	*Object*
Greywage	Membayar Pajak / Tidak	*Object*
Way	Transportasi Ke Kantor	*Object*

Dilanjutkan pada tahapan eksplorasi, proporsi setiap peubah yang memiliki value *unique* kurang dari 10 diplot menggunakan *module* seaborn dan matplotlib. Pada proporsi turnover telihat jika data bervalue "ya" dan "tidak" memiliki ukuran yang sama. Sehingga data yang digunakan tidak mengalami keadaan yang dinamakan *data imbalanced*. Pada proporsi gender, data dipenuhi oleh kaum Perempuan dengan rasio hamper ¾. Terlihat pula di peubah greywage jika proporsi dipenuhi dengan value "grey" sebanyak 88%. Terakhir pada proporsi kendaraan, bis menjadi alat transportasi yang digemari oleh kaum pekerja.



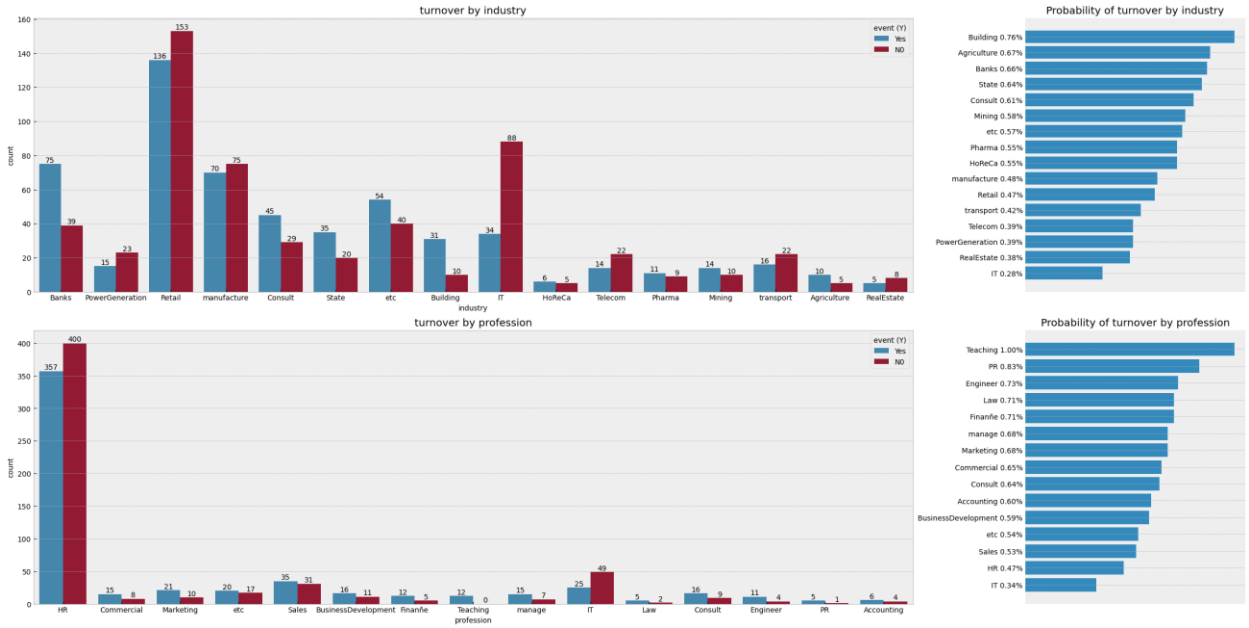
Gambar 1 Proporsi pie chart dengan value unique kurang dari 10

Lalu, eksplorasi dilanjutkan untuk mengetahui distribusi peubah umur terhadap gender dengan turnover menggunakan distribusi histogram. Pada distribusi umur terhadap gender, terlihat bahwa distribusi tersebut adalah *right skewed* dan tidak ada korelasi antar gender terhadap umur. Tempat bekerja didominasi dengan orang yang berada pada di rentang umur 22-30. Lalu pada distribusi umur terhadap turnover, terlihat apabila semakin bertambahnya umur maka semakin besar kemungkinan seorang pekerja untuk melakukan aksi turnover.



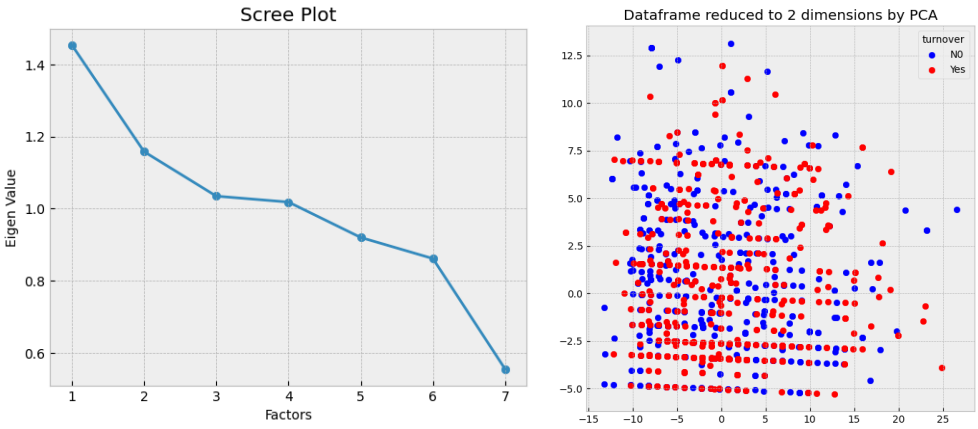
Gambar 2 Distribusi peubah umur terhadap gender dan turnover

Tahapan eksplorasi dilanjutkan dengan melihat proporsi turnover berdasarkan industri dan profesi. Pada plot bar turnover berdasarkan industri. Terlihat apabila probabilitas seseorang yang bekerja pada industri pembangunan memiliki peluang yang paling besar untuk melakukan turnover diikuti dengan industry agrikultur. Lalu, pada proporsi turnover berdasarkan profiesi dapat terlihat apabila profesi dalam mengajar memiliki probabilitas aksi turnover terbesar.



Gambar 3 Barplot turnover berdsarkan peubah industri dan profesi

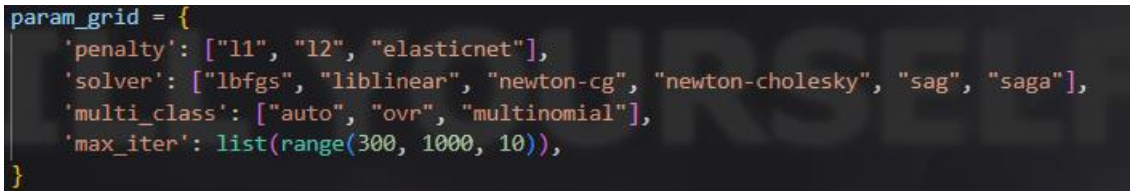
Dilakukan faktor analisis untuk melihat seberapa besar faktor yang dapat diinterpretasikan apabila dimensi peubah dirubah. Lalu dicobakan pemplottingan scatter apabila dimensi direduksi menjadi 2 dimensi.



Gambar 4 Scree plot dan scatterplot dimensi terreduksi

Setelah tahap eksplorasi dilakukan, dilanjutkan dengan tahapan pre modelling. Pada tahapan ini data dibagi menjadi feature dan target. Pada kolum data feature yang berbentuk kategorik (*object*) diubah menggunakan *dummy variable* dan data target diubah menjadi biner [0, 1]. Seusai itu, data dibagi menjadi data latih dan data uji dengan *size* data latih sebesar 0.8.

Model awal didapatkan menggunakan model parameter default yang disediakan oleh *module* sklearn. Pada model tersebut model difit menggunakan data latih. Akurasi data latih dan data uji diterima sebesar 0.6157 dan 0.6416. Oleh karena itu, diperlukan *hyperparameter tuning* untuk mencari model terbaik. Berikut merupakan parameter yang diuji cobakan:



Gambar 5 Paramater yang diuji cobakan pada *hyperparameter tuning*

Setelah parameter yang ingin diuji cobakan telah dideklarisasikan. Tuning dimulai dengan 150 iterasi. Diambil 5 terbaik dari 150 iterasi yang telah dicoba sebagai berikut:

Tabel 2 Paramater terbaik

penalty	solver	multi_class	max_iter	train_score	test_score
l1	saga	auto	600	0.6035	0.6592
l1	saga	multinomial	510	0.6035	0.6592
l1	saga	multinomial	990	0.6035	0.6592
l1	saga	ovr	940	0.6035	0.6592
l1	saga	ovr	870	0.6035	0.6592

Tabel 3 Perbandingan model awal dan model terbaik

Jenis Model	penalty	solver	multi_class	max_iter	train_score	test_score
Model Terbaik	l1	saga	auto	600	0.6035	0.6592
Model Awak	l2	lbfgs	auto	100	0.6157	0.6416

Saran

Jangan menggunakan regresi logistik karena kemungkinan besar banyak asumsi yang tidak terpenuhi dan data tidak linear scara umumnya. Banyak alternatif lainnya yang menghiraukan penggunaan asumsi yang digunakan pada regresi logistik. seperti Neural Network, random forest classifier, classifiertree, dll.

Github: https://github.com/Silly-little-fellas-11/Tugas-2-Sesi-Uas-PADK