Final Project Bootcamp ML & AI

Data: admission

1. Preprocessing Data
2. Cek data

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 505 entries, 0 to 504

Data columns (total 8 columns):

# Column Non-Null Count Dtype

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

0 gre\_score 496 non-null float64

1 toefl\_score 497 non-null float64

2 motiv\_letter\_strength 503 non-null float64

3 recommendation\_strength 499 non-null float64

4 gpa 501 non-null float64

5 research\_exp 505 non-null object

6 admit\_status 505 non-null object

7 univ\_tier 505 non-null object

dtypes: float64(5), object(3)

memory usage: 31.7+ KB

1. Cek duplikasi



Ada 5 data yang terduplikasi, selanjutnya data duplicate didrop, sehingga tersisa 500 data.

1. Cek missing data

gre\_score 9

toefl\_score 8

motiv\_letter\_strength 2

recommendation\_strength 6

gpa 4

research\_exp 0

admit\_status 0

univ\_tier 0

dtype: int64

Atau jika dilihat prosentasenya:

gre\_score 0.018

toefl\_score 0.016

motiv\_letter\_strength 0.004

recommendation\_strength 0.012

gpa 0.008

research\_exp 0.000

admit\_status 0.000

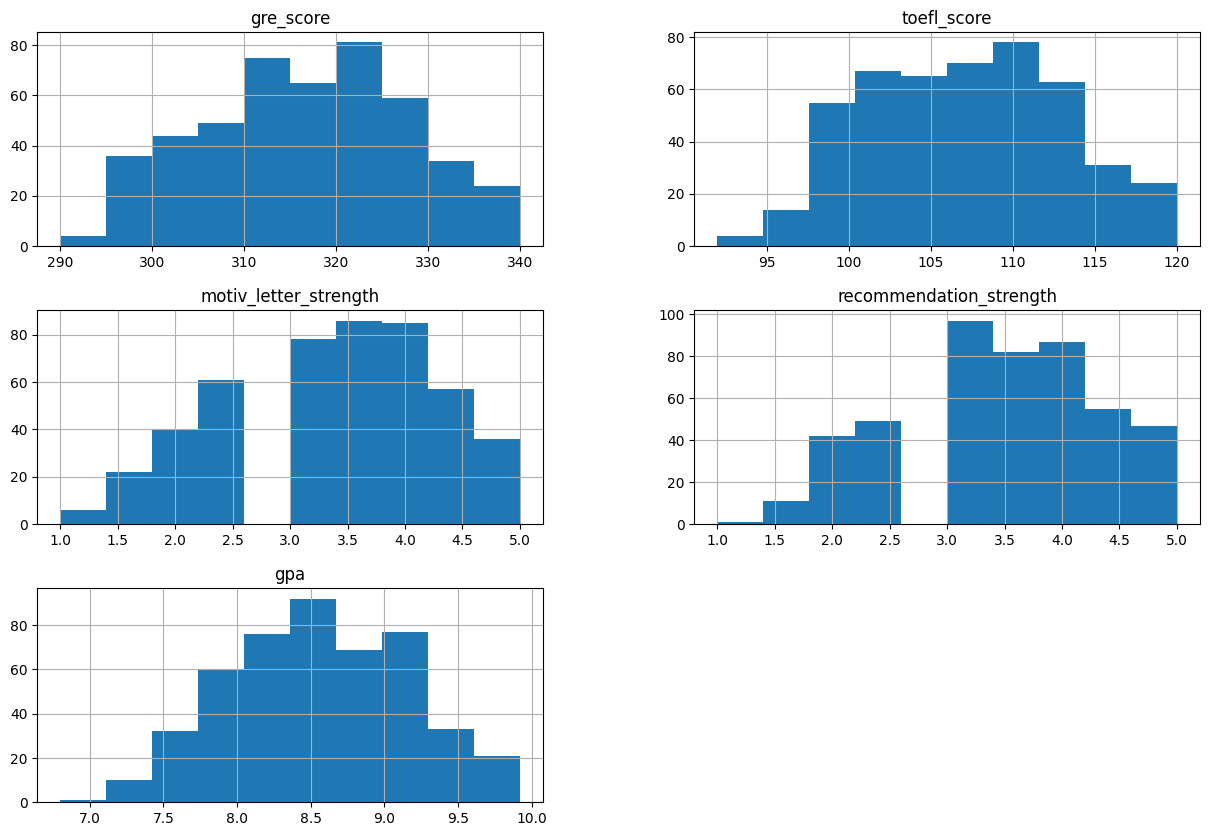
univ\_tier 0.000

dtype: float64

Semua missing < 5 % sehingga didrop saja.

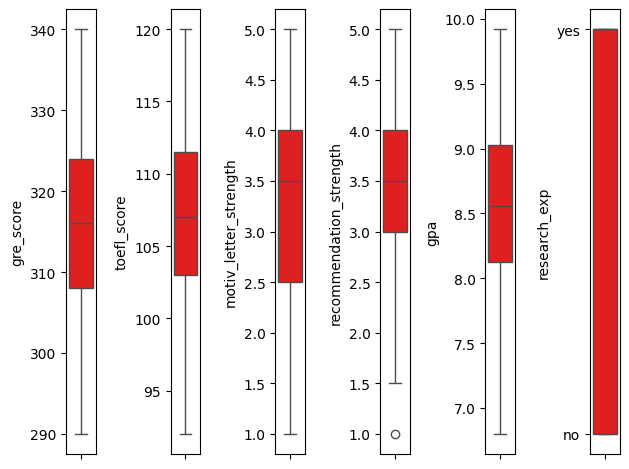
Data yang tersisa tinggal 471 observasi.

1. Cek Distribusi



Seluruh data yang float sudah berdistribusi normal sehingga tidak perlu transformasi.

1. Cek Outlier



Pada kolom recommendation\_strength, outlier di drop, sehingga yang tersisa 470 observasi.

1. Feature Encoding

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Value Counts of research\_exp

research\_exp

yes 256

no 214

Name: count, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Value Counts of admit\_status

admit\_status

yes 268

no 202

Name: count, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Value Counts of univ\_tier

univ\_tier

high 301

low 169

Name: count, dtype: int64

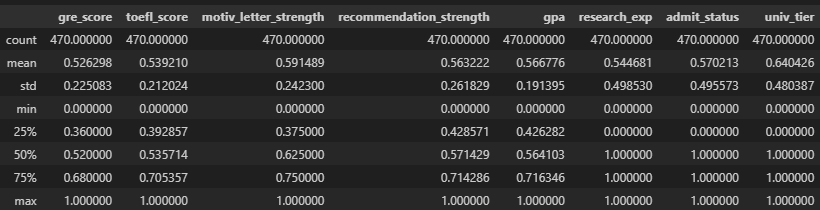
Semua variabel kategorik memiliki dua nilai sehingga data diubah ke 0 atau 1.

1. Scaling

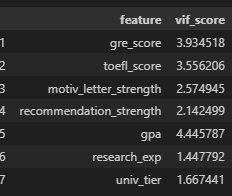
Dilakukan scaling dengan minmax scaler dan diperoleh semua variabel pada dataset memiliki nilai minimal 0 dan maksimal 1.

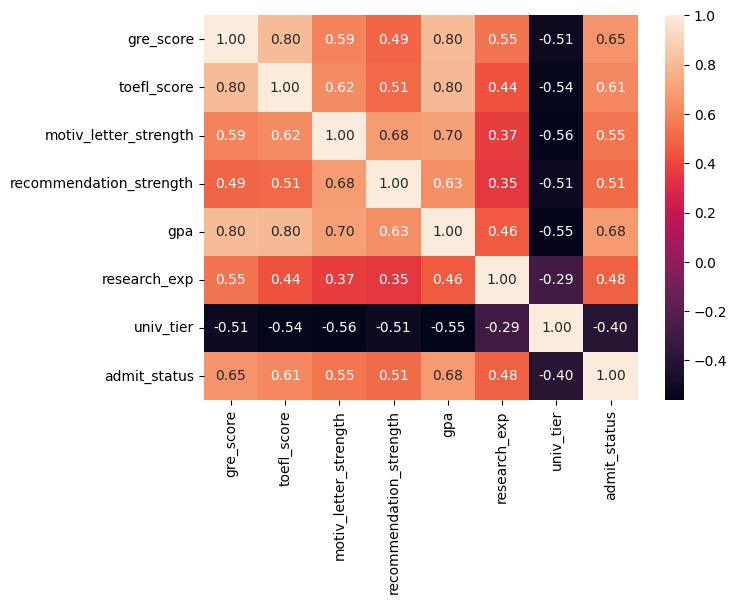
Admission\_clean





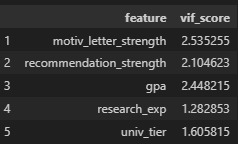
1. Pemodelan regresi Lasso
2. Membagi dataset menjadi training dan testing (80:20)
3. Handling Multicolinearity



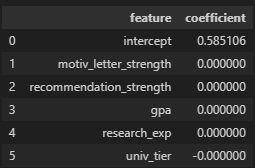


Keputusan: toef dan gre didrop. Sehingga ada 5 fitur yang berpengaruh terhadap admit\_status.

Setelah dicek ulang diperoleh vif:



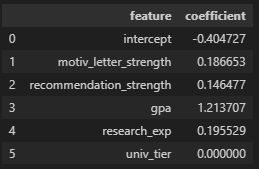
1. Transformasi ke numpy
2. Pemodelan regresi lasso, diperoleh hasil:



Diperoleh semua koefieisen nol. Perlu dilakukan hyperparameter tuning.

1. Tuning dilakukan dengan LassoCV, dan diperoleh alpha terbaik 0.001677

dengan koefisien-koefisien:



Artinya univ\_tier tidak berpengaruh terhadap status admission seseorang.

1. Model evaluasion

Alpha optimal: 0.001677

RMSE for training data is 0.34600583378291166

MAE for training data is 0.28442963082154105

MAPE for training data is 26.2768439563897

RMSE for testing data is 0.31511706450172633

MAE for testing data is 0.2557227540373843

MAPE for testing data is 23.532819075063113

Keasimpulan: