

PREDIKSI PELUANG MAHASISWA UNTUK LOLOS SELEKSI PROGRAM S2

Oleh : Muhammad Rizal Nurcahyo, A11.2019.11740

A. DESKRIPSI JURNAL

a. Jurnal 1

Paper yang saya baca berjudul “PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN PERBANDINGAN ALGORITMA C5.0 DENGAN REGRESSION LINEAR”. Sesuai dengan judul paper tersebut, didalamnya membahas mengenai perbandingan akurasi dari training dan testing data antara algoritma C5 dengan algoritma Regresi Linier. Peneliti tertarik untuk membandingkan akurasi dari algoritma C5 dan Regresi Linier karena prediksi kelulusan mahasiswa yang menerapkan algoritma lain sudah banyak dilakukan.

Adapun tahapan penelitian yang dilakukan adalah sebagai berikut :

1) Mengumpulkan Data

Peneliti menggunakan data dari fakultas teknologi dan ilmu komputer. Data Mahasiswa/Alumni yang diambil memiliki jumlah 715 baris serta 9 atribut/kolom dari tahun 2017-2020.

2) Analisis Data

Langkah yang dilakukan dalam menganalisis data:

- a) Data Preprocessing : dilakukan data cleaning untuk menghilangkan atribut yang tidak bernilai atau missing data (NaN).
- b) Data Transformation : dilakukan normalisasi data dengan tujuan agar data yang bersifat numerik berada range yang tidak terlalu jauh dan data yang bersifat kategorikal akan dikonversi menjadi numerik.
- c) Data mining : proses mencari pola data menggunakan algoritma yang sudah dipilih.

3) Implementasi Python

Melakukan pengujian data menggunakan program python.

4) Evaluasi

Tahap pengambilan keputusan dari pemecahan masalah berkaitan dengan penggunaan hasil data mining.

Hasil :

Hasil perbandingan akurasi training dan testing pada algoritma C5 dengan Regresi Linier adalah sebagai berikut :

Model	Akurasi Training (%)	Akurasi Testing (%)
Algoritma C5	98.85	93.72
Regresi Linier	33.31	40.30

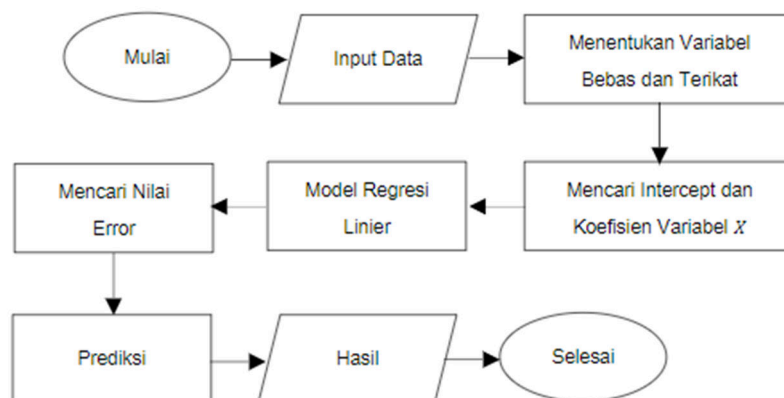
b. Jurnal 2

Judul Jurnal : PREDIKSI JUMLAH MAHASISWA BARU MENGGUNAKAN METODE REGRESI LINIER SEDERHANA

Pada jurnal ini peneliti membahas mengenai prediksi jumlah mahasiswa 5 tahun mendatang di Universitas Islam Negeri Sunan Ampel Surabaya. Tujuan peneliti dari dibuatnya prediksi tersebut adalah untuk mengetahui rasio dosen yang tersedia dengan jumlah mahasiswa baru, mempersiapkan ruang kuliah dan juga fasilitas lainnya.

Pada penelitian tersebut digunakan data jumlah mahasiswa baru salah satu universitas di Surabaya fakultas sains dan teknologi 5 tahun terakhir yaitu mulai dari tahun 2016 hingga tahun 2020

Adapun tahapan penelitian yang dilakukan adalah sebagai berikut :



B. PROGRES CAPAIAN DATASET

- 1) Jenis dataset yang digunakan : Public (sumber : Kaggle.com)
- 2) Dataset yang digunakan : Graduate Admission Prediction
- 3) Jumlah dataset : 1
- 4) Format dataset : CSV
- 5) Atribut dataset :

No.	Nama Atribut	Format Dataset
1	Serial No.	Numerik
2	GRE Score	Numerik
3	TOEFL Score	Numerik
4	University Rating	Numerik
5	SOP	Numerik
6	LOR	Numerik
7	CGPA	Numerik
8	Research	Numerik
9	Chance of Admit	Numerik

C. LATAR BELAKANG MASALAH

Saat ini banyak sekali mahasiswa yang telah lulus S1 dan ingin melanjutkan kembali pendidikannya ke jenjang S2. Namun terkadang timbul masalah apabila mahasiswa lulusan S1 tersebut ingin mengetahui peluangnya lolos ujian S2. Karena ujian S2 menjadi suatu bagian penting untuk menguji kompetensi calon mahasiswa pascasarjana.

Untuk itu dibutuhkan suatu sistem untuk memprediksi peluang kelulusan masuk pascasarjana. Sehingga dapat membantu calon mahasiswa pascasarjana dalam memberikan gambaran apakah ada peluang untuk lulus seleksi masuk pascasarjana atau tidak.

D. METODE PENYELESAIAN

Saya menggunakan metode Regresi Linear untuk menyelesaikan masalah diatas. Regresi linear merupakan pendekatan untuk memantapkan hubungan antara

satu atau lebih variabel dependen (regresi linear sederhana) dan juga variabel independen (regresi linear banyak).

Pada penelitian ini data yang digunakan berupa data sekunder, yaitu data yang didapatkan secara tidak langsung atau data yang didapat dari dokumen dan penyampaian orang lain. Regresi linier merupakan salah satu perhitungan time series metode kuantitatif dimana waktu digunakan sebagai dasar prediksi.

Data yang didapatkan merupakan data yang sudah matang yang didownload dari website Kaggle.com. data ini bersifat supervised yang artinya data sudah diberi label dan dikelompokkan berdasarkan labelnya.

E. TAHAPAN PENYELESAIAN

Tahapan yang dilakukan dalam pelaksanaan penelitian adalah sebagai berikut :

a. Pengumpulan Data

Tahap pertama dalam pelaksanaan penelitian ini adalah melakukan pengumpulan data. Data yang dikumpulkan merupakan data prediksi penerimaan lulusan yang didownload dari situs kaggle.com. terdapat 500 baris data yang terdiri dari 9 kolom dengan atribut seperti tabel diatas.

Tabel 1 : Data Variabel Prediksi Kelulusan

Serial_No.	GRE_Score	TOEFL_Score	University_Rating	SOP	LOR	CGPA	Research	Chance_of_Admit	
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65
...
495	496	332	108	5	4.5	4.0	9.02	1	0.87
496	497	337	117	5	5.0	5.0	9.87	1	0.96
497	498	330	120	5	4.5	5.0	9.56	1	0.93
498	499	312	103	4	4.0	5.0	8.43	0	0.73
499	500	327	113	4	4.5	4.5	9.04	0	0.84

500 rows x 9 columns

Tabel yang disajikan diatas merupakan data yang akan digunakan untuk menentukan prediksi peluang masuk program pascasarjana.

b. Preprocessing

- Cleaning Data

Tahap ini digunakan untuk mengisi nilai yang hilang, menghilangkan data yang bersifat noise, dan mengatasi ketidakkonsistenan data.

```
# Handling missing value
raw_data.isnull().sum()
```

```
Serial_No.      0
GRE_Score       0
TOEFL_Score     0
University_Rating 0
SOP             0
LOR             0
CGPA            0
Research        0
Chance_of_Admit 0
dtype: int64
```

Terlihat bahwa data sudah siap untuk dilanjutkan ke proses selanjutnya

	Datatype	Null Values	Null %	No: Of Unique Values	count	mean	std	min	25%	50%	75%	max
Serial_No.	int64	0	0.0	500	500.0	250.50000	144.481833	1.00	125.7500	250.50	375.25	500.00
GRE_Score	int64	0	0.0	49	500.0	316.47200	11.295148	290.00	308.0000	317.00	325.00	340.00
TOEFL_Score	int64	0	0.0	29	500.0	107.19200	6.081868	92.00	103.0000	107.00	112.00	120.00
University_Rating	int64	0	0.0	5	500.0	3.11400	1.143512	1.00	2.0000	3.00	4.00	5.00
SOP	float64	0	0.0	9	500.0	3.37400	0.991004	1.00	2.5000	3.50	4.00	5.00
LOR	float64	0	0.0	9	500.0	3.48400	0.925450	1.00	3.0000	3.50	4.00	5.00
CGPA	float64	0	0.0	184	500.0	8.57644	0.604813	6.80	8.1275	8.56	9.04	9.92
Research	int64	0	0.0	2	500.0	0.56000	0.496884	0.00	0.0000	1.00	1.00	1.00
Chance_of_Admit	float64	0	0.0	61	500.0	0.72174	0.141140	0.34	0.6300	0.72	0.82	0.97

Gambar diatas adalah hasil cek dari meta data

- Menghapus Atribut yang tidak digunakan

Saya melakukan drop atribut Serial.No dan melakukan penghapusan space dari kolom Chance_of_Admit

```
data_col_chng = raw_data.copy()
data_col_chng.drop(columns = ['Serial_No.'], axis = 1, inplace = True)
data_col_chng.columns = data_col_chng.columns.str.strip()
```

Dan berikut adalah hasilnya

	GRE_Score	TOEFL_Score	University_Rating	SOP	LOR	CGPA	Research	Chance_of_Admit
0	337	118	4	4.5	4.5	9.65	1	0.92
1	324	107	4	4.0	4.5	8.87	1	0.76
2	316	104	3	3.0	3.5	8.00	1	0.72
3	322	110	3	3.5	2.5	8.67	1	0.80
4	314	103	2	2.0	3.0	8.21	0	0.65
...
495	332	108	5	4.5	4.0	9.02	1	0.87
496	337	117	5	5.0	5.0	9.87	1	0.96
497	330	120	5	4.5	5.0	9.56	1	0.93
498	312	103	4	4.0	5.0	8.43	0	0.73
499	327	113	4	4.5	4.5	9.04	0	0.84

500 rows x 8 columns

c. Processing

- Splitting Data ke dalam variabel dependen dan independen

```
X = data_col_chng.iloc[:, :-1].values
y = data_col_chng.iloc[:, -1].values
print('Shape of X: ', X.shape)
print('Shape of y: ', y.shape)
```

```
Shape of X: (500, 7)
Shape of y: (500,)
```

- Splitting Data ke dalam data train dan data test

```
Shape of X_train: (400, 7)
Shape of X_test: (100, 7)
Shape of y_train: (400,)
Shape of y_test: (100,)
```

- Modeling dengan MinMaxScaler

Membuat model multi linear regression dengan minmaxscaler menghasilkan output :

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
l_regressor = LinearRegression()
l_regressor.fit(X_train_mms_scaled, y_train)
y_pred_l_reg = l_regressor.predict(X_test_mms_scaled)
print(r2_score(y_test, y_pred_l_reg))
```

```
0.8188432567829629
```

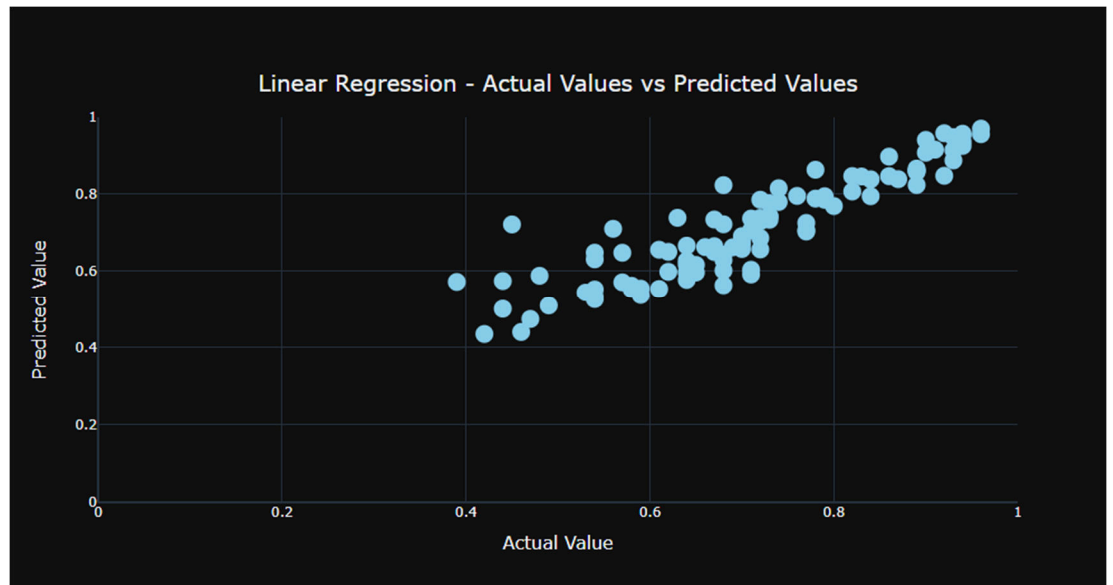
- Modeling dengan StandardScaler

Membuat model multi linear regression dengan standardscaler menghasilkan output :

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
l_regressor_ss = LinearRegression()
l_regressor_ss.fit(X_train_ss_scaled, y_train)
y_pred_l_reg_ss = l_regressor_ss.predict(X_test_ss_scaled)
print(r2_score(y_test, y_pred_l_reg_ss))
```

0.8188432567829629

- Menampilkan hasil regresi linier actual value dengan predicted value



d. Uji MSE, RMSE, dan MAPE

Hasil prediksi yang telah didapatkan kemudian selanjutnya memasuki tahapan pengujian untuk memastikan keakurasiannya. Pengujian keakurasian ini dilakukan menggunakan tiga pengujian yang terdiri dari uji MSE, RMSE, dan MAPE. Uji akurasi dengan ketiga metode tersebut menghasilkan output :

```
MAE: 0.042722654277053664
MSE: 0.00370465539878841
RMSE: 0.060865880415783113
```

Hasil dari uji akurasi diatas menunjukkan bahwa nilai mendekati nol sehingga pengujian dengan model linear regression sangat cocok untuk digunakan sebagai prediksi.

Tools yang digunakan :

- Visual Studio Code

Digunakan untuk melakukan pembuatan fungsi-fungsi sebagai bentuk implementasi program

- Jupyter Notebook
Digunakan untuk melakukan processing data
- Xampp
Digunakan untuk menjalankan perintah eksekusi implementasi program python ke dalam bentuk web
- CMD
Digunakan untuk membuka jupyter notebook maupun membuka code

Bahasa Pemrograman yang digunakan :

- Python

F. DAFTAR PUSTAKA

- [1] Almumtazah, N., Azizah, N., Putri, Y, L., Novitasari, C, R, (2021), *“Prediksi Jumlah Mahasiswa Baru Menggunakan Metode Regresi Linier Sederhana”*, Jurnal Ilmiah Matematika dan Terapan, Vol. 18, No. 1, 31-40.
- [2] Zebua, Falda, Junisman., Manalu, Ribka, Permatasari., Nababan, Marlince, Novita, Karoseri, 2021, *“Prediksi Kelulusan Mahasiswa Menggunakan Perbandingan Algoritma C5.0 Dengan Regression Linier”*, Jurnal TEKINKOM, Vol. 4, No. 2, 230-238.