LAPORAN
PEMBAHASAN HASIL STUDI KASUS
DATA COMPETITION ISFEST 2022

Kelompok Oh Data Euy (ODE) :

1. Felix Fernando

2. Gerend Christopher

3. Jeremy

INFORMATION SYSTEM FESTIVAL UMN

2022

BUSINESS UNDERSTANDING

Permasalahan masalah penelitian mengacu pada performansi studi mahasiswa suatu perguruan tinggi berdasarkan waktu studi. Tujuan dari penelitian ini adalah menentukan model klasifikasi yang tepat untuk dapat melakukan prediksi kelulusan yang baik dari beberapa model. Penerapan *data mining* pada penelitian ini berhubungan langsung dengan data nilai mahasiswa untuk menggali pengetahuan tentang suatu pola kelulusan mahasiswa yang tepat waktu, serta untuk menentukan model klasifikasi yang memberikan nilai akurasi yang baik berdasarkan pola kelulusan yang ada berkaitan dengan parameter-parameter nilai *input*.

DATA UNDERSTANDING

Pada fase ini dilakukan pengumpulan data awal sebagai syarat kelulusan dengan rincian sebagai berikut:

a) Sudah menyelesaikan 144 SKS
b) Tidak ada nilai D, E, dan F pada setiap mata kuliah
c) IPK di atas 2.5

Namun, beberapa penilaian subyektif dapat saja terjadi dengan anggapan bahwa sistem ajar dosen memengaruhi kemampuan mahasiswa menyerap materi di kelas. Dengan demikian, akan dilakukan juga analisis terhadap klasifikasi kualitas dosen berdasarkan kompetensi sesuai dengan evaluasi pengajaran yang terdiri dari nilai-nilai evaluasi yang diberikan mahasiswa per nomor pertanyaan yang diberikan.

Pemahaman data mengacu pada dokumen transkrip nilai mahasiswa dan evaluasi dosen per mata kuliah. Berikut atribut yang terdapat dalam dokumen transkrip nilai mahasiswa:

a) NIM/Nomor Induk Mahasiswa,
b) Angkatan Tahun Mahasiswa,
c) Periode Semester berupa empat digit kode dengan dua digit pertama menunjukkan tahun 20XX dan digit ketiga menunjukkan semester ganjil (1) atau semester genap (2). Misalnya, atribut 1211 menunjukkan semester ganjil tahun 2012.

d) Mata Kuliah yang terdiri dari kode dan nama mata kuliah yang disesuaikan dengan masing-masing periode tahun ajaran atau semester, misalnya IS100 Sistem Informasi Manajemen, IS341 Sistem Basis Data, IS201 Proses Bisnis Korporat, dsb.
e) Jumlah SKS,
f) Nilai,
g) Indeks (*Grade*)

NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE
10110310002	2010	1011	EM100	EM100 Dasar-dasar Bisnis	3	57	C
10110310002	2010	1011	EM180	EM180 Matematika Bisnis	3	70	B
10110310002	2010	1011	TI100	TI100 Algoritma dan Pemrograman	4	57	C
10110310002	2010	1011	TI101	TI101 Matematika Diskrit	3	59	C
10110310002	2010	1011	TI110	TI110 Pengantar Teknologi Multimedia	3	74	B
10110310002	2010	1011	UM121	UM121 Bahasa Inggris 1	2	59	C
10110310002	2010	1011	UM151	UM151 Agama	3	71	B
10110310002	2010	1021	EM201	EM201 Dasar-dasar Manajemen	3	56	C
10110310002	2010	1021	IK402	IK402 Komunikasi Interpersonal	2	57	C

Gambar 1. Cuplikan Dataset Transkrip Nilai Mahasiswa

Dataset transkrip nilai mahasiswa ini mencakup data dari 770 mahasiswa. Kemudian, dilakukan pengecekan data yang *non-null* pada setiap atribut dengan hasil sebagai berikut.

```
RangeIndex: 30870 entries, 0 to 30869
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   NIM       30870 non-null  object
 1   ANGKATAN  30870 non-null  int64
 2   SEMESTER  30870 non-null  int64
 3   KODE_MK   30870 non-null  object
 4   NAMA_MK   30847 non-null  string
 5   SKS       30870 non-null  int64
 6   NILAI     30858 non-null  float64
 7   GRADE     30318 non-null  object
dtypes: float64(1), int64(3), object(3), string(1)
```

Gambar 2. Hasil Pengecekan *Non-Null* Dataset Transkrip Nilai Mahasiswa

Dari hasil tersebut, diperoleh bahwa terdapat atribut nama mata kuliah, nilai, dan *grade* yang *null* karena jumlah entri yang diperoleh lebih kecil dari total entri data, yaitu sebanyak 30870 entri.

Sementara itu, dokumen evaluasi dosen per mata kuliah memiliki susunan atribut sebagai berikut:

a) Tahun ajaran,
b) Semester berupa empat digit kode dengan dua digit pertama menunjukkan tahun 20XX dan digit ketiga menunjukkan semester ganjil (1) atau semester genap (2). Misalnya, atribut 1211 menunjukkan semester ganjil tahun 2012.
c) Mata Kuliah, terdiri atas kode dan nama mata kuliah
d) Pertanyaan,
e) Keterangan berupa deskripsi dari pertanyaan yang bersangkutan,
f) Nilai yang diberikan mahasiswa per nomor pertanyaan yang bersangkutan.

TAHUN	SEMESTER	MATAKULIAH	PERTANYAAN	KETERANGAN	NILAI
2015	1511	IS100 Management Information Systems	1	Kesiapan memberikan perkuliahan/praktikum	3,28
2015	1511	IS100 Management Information Systems	2	Upaya menyampaikan materi perkuliahan/praktikum dengan jelas	3,25
2015	1511	IS100 Management Information Systems	3	Sistematis dalam menyampaikan materi perkuliahan/praktikum	3,24
2015	1511	IS100 Management Information Systems	4	Kemampuan memberikan contoh yang relevan dari materi yang diajarkan	3,3
2015	1511	IS100 Management Information Systems	5	Penyampaian materi perkuliahan sesuai dengan kontrak perkuliahan	3,27
2015	1511	IS100 Management Information Systems	6	Pemakaian buku teks sebagai buku utama perkuliahan	3,19
2015	1511	IS100 Management Information Systems	7	Memberi review materi perkuliahan sebelumnya	3,18
2015	1511	IS100 Management Information Systems	8	Pemberian kesempatan bertanya, berdiskusi serta berkonsultasi (baik di dalam maupun di luar kelas)	3,32
2015	1511	IS100 Management Information Systems	9	Kejelasan menjawab pertanyaan /diskusi di kelas	3,29
2015	1511	IS100 Management Information Systems	10	Pemberian tugas/kuis serta pembahasannya di kelas	3,24
2015	1511	IS100 Management Information Systems	11	Kemampuan memotivasi semangat belajar mahasiswa	3,31
2015	1511	IS100 Management Information Systems	12	Kemampuan menerima kritik, saran dan pendapat	3,27
2015	1511	IS100 Management Information Systems	13	Fairness dalam memberikan penilaian	3,26
2015	1511	IS100 Management Information Systems	14	Ketepatan waktu dalam memulai dan mengakhiri perkuliahan/praktikum	3,31

Gambar 3. Cuplikan Dataset Evaluasi Dosen per Mata Kuliah

Selanjutnya, dilakukan pengecekan data yang *non-null* pada setiap atribut dengan hasil sebagai berikut.

```
RangeIndex: 2114 entries, 0 to 2113
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   TAHUN       2114 non-null   int64
 1   SEMESTER    2114 non-null   int64
 2   MATAKULIAH  2114 non-null   object
 3   PERTANYAAN  2114 non-null   int64
 4   KETERANGAN  2114 non-null   object
 5   NILAI       2114 non-null   float64
dtypes: float64(1), int64(3), object(2)
```

Gambar 4. Hasil Pengecekan *Non-Null* Dataset Evaluasi Dosen per Mata Kuliah

Dari hasil tersebut, diperoleh bahwa tidak ada nilai *null* pada setiap atribut karena jumlah entri masing-masing atribut sama dengan total entri, yaitu sebanyak 2114 entri.

DATA PREPARATION

A. *Dataframe* Transkrip Nilai Mahasiswa (*Dataframe 1)*

Pada fase ini dilakukan pemilihan dan pembersihan data. Dataset transkrip nilai mahasiswa yang terdiri dari 30870 entri memiliki beberapa *missing value* pada atribut nama mata kuliah sebanyak 23 entri, atribut nilai sebanyak 12 entri, dan atribut *grade* sebanyak 552 entri.

Pertama, dilakukan pengecekan terhadap atribut mata kuliah pada atribut nama mata kuliah yang *null* dan ditemukan bahwa 23 entri nama mata kuliah yang *null* memiliki kode mata kuliah SI863 sehingga 23 entri tersebut ditambahkan nama mata kuliah yang berkorespondensi dengan kode tersebut, yaitu Tugas Akhir.

Berikutnya, untuk atribut *grade* yang *null* kita lakukan pemberian *grade* berdasarkan nilai yang diperoleh mahasiswa dengan ketentuan sebagai berikut.

Numerik	Grade	Bobot Nilai	Deskripsi
85-100	A	4	Sangat Baik
80 – 84.99	A-	3.7	Baik
75 – 79.99	B+	3.3	
70 – 74.99	B	3.0	
65 - 69.99	B-	2.7	Cukup
60 – 64.99	C+	2.3	
55 – 59.99	C	2.0	

45 – 54.99	D	1.0	Kurang
0 – 44.99	E	0	Sangat Kurang
	F	0	Pelanggaran Akademik

Gambar 5. Ketentuan *Grade* Nilai Mahasiswa

Lalu, untuk mempermudah perhitungan Indeks Prestasi nantinya, ditambahkan atribut bobot nilai sesuai *grade* yang diperoleh dan total bobot berupa hasil perkalian dari bobot nilai dan jumlah SKS.

NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE	BOBOT	TOTAL_BOBOT
10110310002	2010	1011	EM100	EM100 Dasar-dasar Bisnis	3	57.0	C	2.0	6.0
10110310002	2010	1011	EM180	EM180 Matematika Bisnis	3	70.0	B	3.0	9.0
10110310002	2010	1011	TI100	TI100 Algoritma dan Pemrograman	4	57.0	C	2.0	8.0
10110310002	2010	1011	TI101	TI101 Matematika Diskrit	3	59.0	C	2.0	6.0
10110310002	2010	1011	TI110	TI110 Pengantar Teknologi Multimedia	3	74.0	B	3.0	9.0

Gambar 6. Cuplikan Dataset Transkrip Nilai Mahasiswa yang Baru

Data pada kolom NAMA_MK, menunjukkan adanya ketidakkonsistenan dalam penamaannya, terdapat kasus dengan NAMA_MK yang tidak didahului dengan KODE_MK, sehingga untuk agar konsistenan, semua data pada kolom NAMA_MK ditambahkan dengan KODE_MK sebelum nama dari masing-masing kuliahnya.

Untuk kasus nilai yang *null* ditemukan bahwa 12 entri tersebut memiliki *grade* F sehingga akan diberikan nilai nol sesuai dengan ketentuan nilai mahasiswa. Kemudian, dilakukan pengecekan untuk mahasiswa yang mengambil mata kuliah lebih dari satu kali atau mengulang mata kuliah dengan mengecek apakah ada NIM yang memiliki mata kuliah duplikasi. Diperoleh bahwa ada 17 mahasiswa yang mengambil suatu mata kuliah secara berulang, yaitu mahasiswa dengan NIM '10110310011', '10110310022', '10110310036', '10110310037', '10110310044', '10110310045', '10110310047', '10110310083', '11110310003', '11110310004', '11110310005', '11110310009', '11110310020', '11110310026', '11110310035', '11110310080', '13110310069'. Dengan demikian, dilakukan modifikasi dataset dengan mengambil nilai terakhir dari mata kuliah yang

diambil secara berulang tersebut dan menghilangkan (*dropping*) mata kuliah duplikasi yang lainnya.

B. *Dataframe* Evaluasi Dosen (*Dataframe 2)*

Untuk data pada evaluasi Dosen, tidak terdapat data dengan entri *null*, data evaluasi dosen ini yang kemudian dilakukan modifikasi dengan adanya penambahan kolom nilai rata-rata dari 14 pertanyaan yang ada ke dalam *dataframe* baru.

C. *Dataframe* Tambahan

Dengan dua data sebelumnya, untuk membantu modelling dan proses EDA, dibuat beberapa dataframe baru yaitu, *Dataframe* transkrip nilai mahasiswa untuk masing-masing NIM (*Dataframe 3)*, *Dataframe* evaluasi dosen per semester dan mata kuliah (*Dataframe 4)*, dan *Dataframe* kelulusan mahasiswa per tahun (*Dataframe 5)*.

a. *Dataframe* Nilai Mahasiswa per NIM (*Dataframe 3)*

Untuk *dataframe* transkrip nilai mahasiswa untuk masing-masing NIM (*Dataframe 3)*, dibuat dengan cara memodifikasi *dataframe* nilai mahasiswa *(Dataframe 1)*, yaitu dengan melakukan penjumlahan terhadap total SKS dari masing-masing mahasiswa, total bobot dari masing-masing mahasiswa, IPK dari masing-masing mahasiswa, kumpulan nilai dari mata kuliah yang gagal, waktu kuliah mahasiswa (dalam tahun), banyaknya mata kuliah yang gagal, dan kolom hasil yang memiliki nilai“Lulus Tepat Waktu”, “Lulus Telat”, atau “Tidak Lulus”. Setiap data dibuat dalam bentuk kolom masing-masing,

	NIM	ANGKATAN	TOTAL_SKS	FAILED_GRADE	TOTAL_BOBOT	WAKTU_KULIAH	IPK	TOTAL_FAILED_GRADE	HASIL
0	00000008429	2015	145	D	474.5	3.5	3.272414	1.0	Tidak Lulus
1	00000008455	2015	136	DD	416.2	3.5	3.060294	2.0	Tidak Lulus
2	00000008481	2015	127	DDDEEED	313.6	3.5	2.469291	7.0	Tidak Lulus
3	00000008631	2015	145		519.3	3.0	3.581379	0.0	Lulus Tepat Waktu
4	00000008684	2015	145		512.1	3.0	3.531724	0.0	Lulus Tepat Waktu

Gambar 7. Cuplikan *Dataframe* Transkrip Nilai Mahasiswa per NIM

b. *Dataframe* Evaluasi Dosen per Semester dan Mata Kuliah (*Dataframe 4)*

Untuk *dataframe* evaluasi dosen per semester dan mata kuliah (*Dataframe 4)*, dibuat dengan cara menggabungkan dan memodifikasi dua *dataframe* awal yaitu *dataframe* transkrip nilai

mahasiswa dan *dataframe* evaluasi dosen. Dari *dataframe* transkrip nilai mahasiswa diperhitungkan jumlah indeks dari masing-masing mata kuliah untuk tiap semester yang berbeda, yang kemudian dibedakan mahasiswa yang lulus dan tidak lulus,

	KODE_MK	SEMESTER	GRADE	TOTAL
0	CE441	1421	A	2
1	CE441	1421	A-	7
2	CE441	1421	B	10
3	CE441	1421	B+	3
4	CE441	1421	B-	7

Gambar 8. Cuplikan *Dataframe* Jumlah Indeks Masing-masing KODE_MK Setiap Semester

Kemudian, *dataframe* evaluasi dosen *(Dataframe 2)*, modifikasi sehingga, untuk masing-masing dosen, 14 pertanyaan yang menjadi evaluasi dosen diubah menjadi kolom dan ditambahkan kolom nilai rata-rata dari penilaian dosen,

	TAHUN	NAMA_MK	SEMESTER	KODE_MK	NILAI_RATA-RATA	LULUS	TIDAK_LULUS	1	2	3	...	6	7	8	9	10	11	12	13	14
0	2015	IS100 Management Information Systems	1511	IS100	3.265000	99.0	4.0	3.28	3.25	3.24	...	3.19	3.18	3.32	3.29	3.24	3.31	3.27	3.26	3.31
1	2015	IS110 Business Mathematics	1511	IS110	3.087143	76.0	18.0	3.19	2.98	3.08	...	3.05	2.91	3.20	3.07	3.07	2.97	3.08	3.19	3.22
2	2015	IS201 Corporate Business Processes	1511	IS201	3.170000	2.0	1.0	3.27	3.18	3.16	...	3.20	3.02	3.20	3.14	3.14	2.98	3.09	3.20	3.33
3	2015	IS201 Corporate Business Processes	1521	IS201	2.958571	61.0	6.0	2.96	2.83	2.93	...	2.96	3.01	2.99	3.00	2.96	2.81	2.93	3.00	3.04
4	2015	IS220 Human and Computer Interaction	1521	IS220	3.295000	96.0	0.0	3.26	3.25	3.25	...	3.30	3.23	3.38	3.30	3.30	3.29	3.35	3.41	3.32

Gambar 9. Cuplikan *Dataframe* Evaluasi Dosen Termodifikasi

Lalu, kedua *dataframe* digabung menjadi satu buah *dataframe* dengan penambahan kolom proporsi kelulusan mahasiswa, proporsi kelulusan dihitung dengan,

$$proporsi\ lulus = \frac{Lulus}{Lulus + Tidak\ Lulus}$$

hasil terakhir dari *dataframe* evaluasi dosen per semester dan mata kuliah (*Dataframe 4) yaitu,*

	TAHUN	NAMA_MK	SEMESTER	KODE_MK	NILAI_RATA-RATA	LULUS	TIDAK_LULUS	1	2	3	...	6	7	8	9	10	11	12	13	14	PROPORSI_LULUS
0	2015	IS100 Management Information Systems	1511	IS100	3.265000	99.0	4.0	3.28	3.25	3.24	...	3.19	3.18	3.32	3.29	3.24	3.31	3.27	3.26	3.31	0.961165
1	2015	IS110 Business Mathematics	1511	IS110	3.087143	76.0	18.0	3.19	2.98	3.08	...	3.05	2.91	3.20	3.07	3.07	2.97	3.08	3.19	3.22	0.808511
2	2015	IS201 Corporate Business Processes	1511	IS201	3.170000	2.0	1.0	3.27	3.18	3.16	...	3.20	3.02	3.20	3.14	3.14	2.98	3.09	3.20	3.33	0.666667
3	2015	IS201 Corporate Business Processes	1521	IS201	2.958571	61.0	6.0	2.96	2.83	2.93	...	2.96	3.01	2.99	3.00	2.96	2.81	2.93	3.00	3.04	0.910448
4	2015	IS220 Human and Computer Interaction	1521	IS220	3.295000	96.0	0.0	3.26	3.25	3.25	...	3.30	3.23	3.38	3.30	3.30	3.29	3.35	3.41	3.32	1.000000

Gambar 10. Cuplikan *Dataframe* Evaluasi Dosen per Semester dan Mata Kuliah

c. *Dataframe* Kelulusan Mahasiswa per Tahun (*Dataframe 5)*

Untuk *dataframe* kelulusan mahasiswa per tahun (*Dataframe 5)*, *dataframe* dilakukan perhitungan terhadap jumlah mahasiswa yang lulus tiap angkatannya, *dataframe* ini merupakan modifikasi dari *dataframe* Transkrip Nilai Mahasiswa per NIM,

	ANGKATAN	JUMLAH KELULUSAN	JUMLAH MAHASISWA	PROPORSI KELULUSAN
0	2010	34	3157	0.010770
1	2011	33	3395	0.009720
2	2012	22	2002	0.010989
3	2013	41	3880	0.010567
4	2014	27	4464	0.006048
5	2015	24	4643	0.005169

Gambar 11. Cuplikan *Dataframe* Kelulusan Mahasiswa per Tahun

PREDICTION MODEL AND EVALUATION

A. Exploratory Data Analysis (EDA)

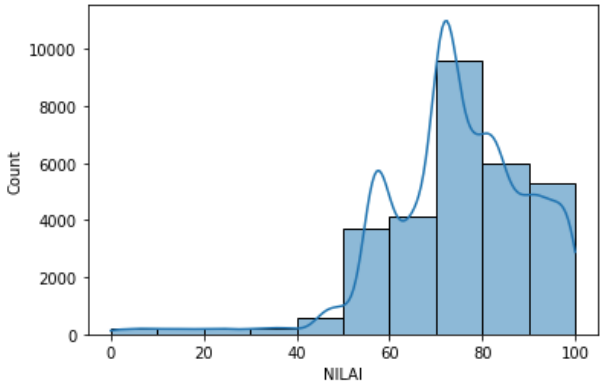
a. *Dataframe* Pertama: Nilai Mahasiswa

Dataframe nilai mahasiswa dan memiliki statistika deskriptif sebagai berikut.

	NILAI
count	30085.000000
mean	74.263321
std	15.751127
min	0.000000
25%	66.000000
50%	74.000000
75%	85.000000
max	100.000000

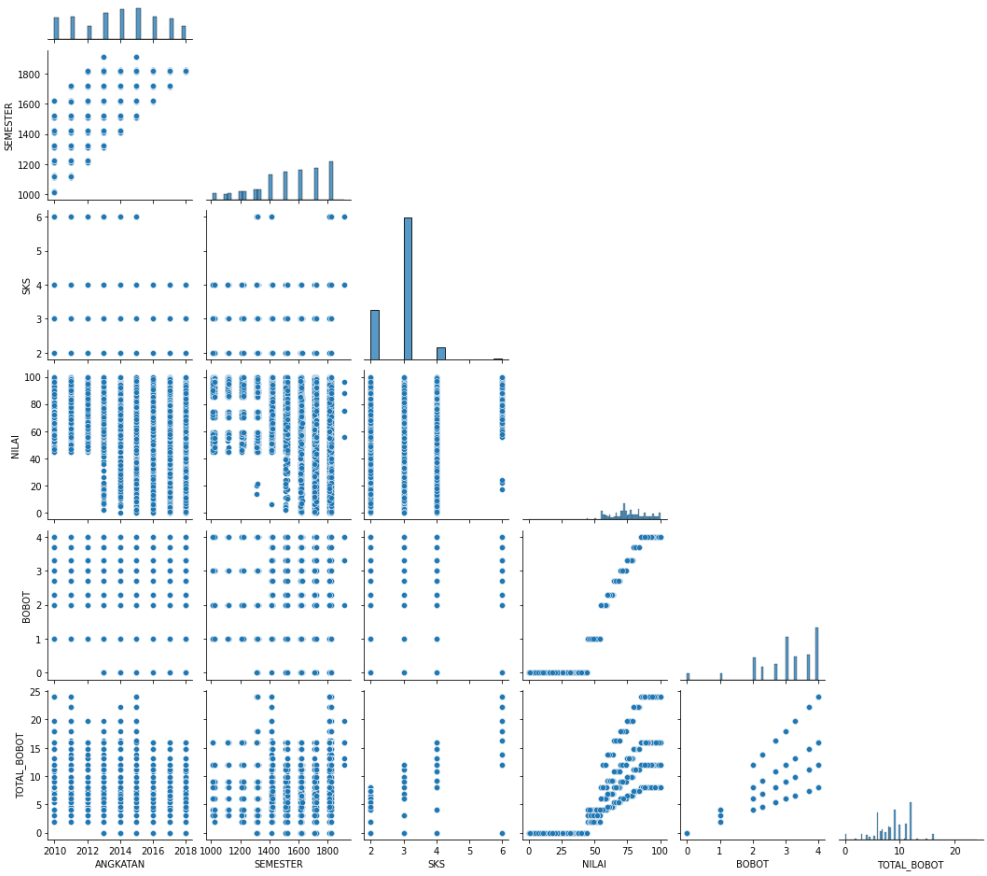
Gambar 12. Stastistika Deskriptif *Dataframe 1*

Lalu, diperoleh juga distribusi nilai mahasiswa dalam bentuk histogram sebagai berikut.



Gambar 13. Histogram Distribusi Nilai Mahasiswa

Terlihat bahwa distribusi nilai mahasiswa menceng ke kiri (*negatively skewed*).

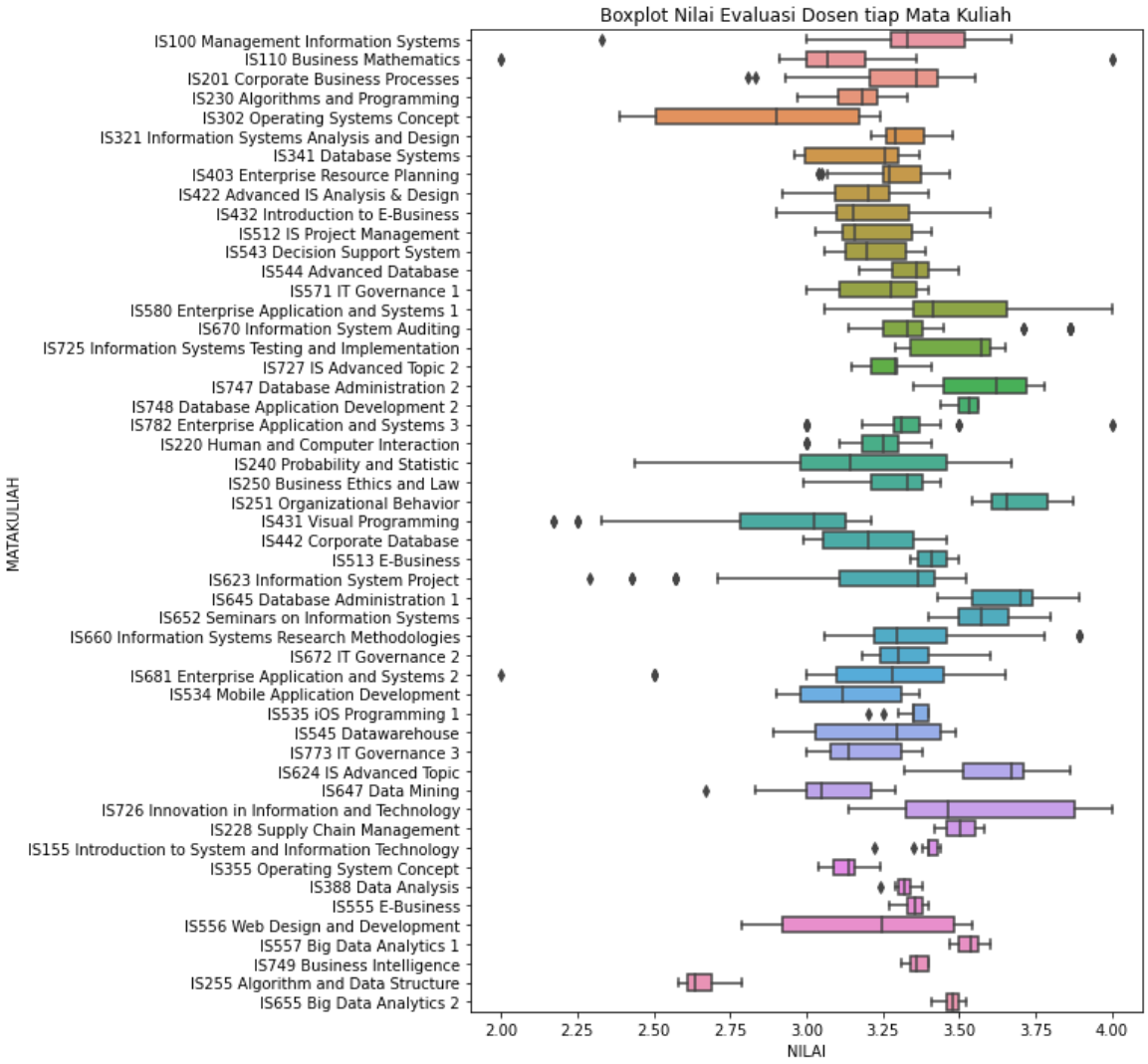


Gambar 14. *Scatter Matrix* Kolom pada *Dataframe 1*

Melalui *scatter matrix*, dapat dilihat juga hubungan linier setiap atribut data. Terlihat bahwa antar atribut pada data mahasiswa tidak memiliki hubungan linier kecuali untuk atribut semester dan angkatan namun kedua hal tersebut tidak menjadi ketertarikan dalam tujuan penelitian saat ini.

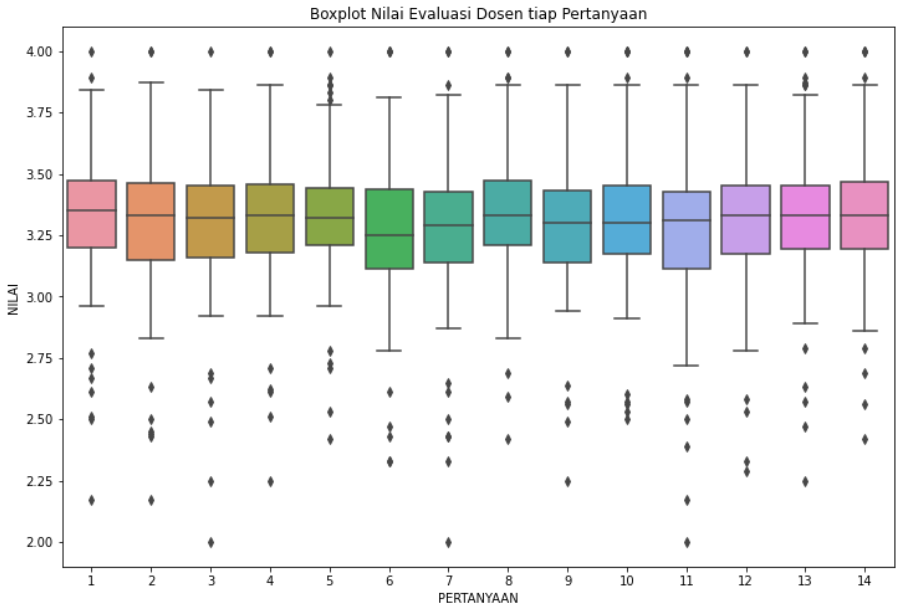
b. *Dataframe* Kedua: Evaluasi Dosen per Mata Kuliah

Berikutnya, kita akan menganalisis dataframe kedua yang berisi mengenai evaluasi dosen per mata kuliah. Pertama, dilakukan *plotting box plot* nilai evaluasi dosen per mata kuliah.



Gambar 15. Boxplot Nilai Evaluasi Dosen tiap Mata Kuliah pada Dataframe 2

Berdasarkan plot tersebut, dapat dilihat bahwa dosen mata kuliah IS225 *Algorithm and Data Structure* memiliki nilai yang relatif kecil dibandingkan mata kuliah lainnya. Selanjutnya, dilakukan *plotting boxplot* untuk nilai evaluasi dosen setiap pertanyaan.



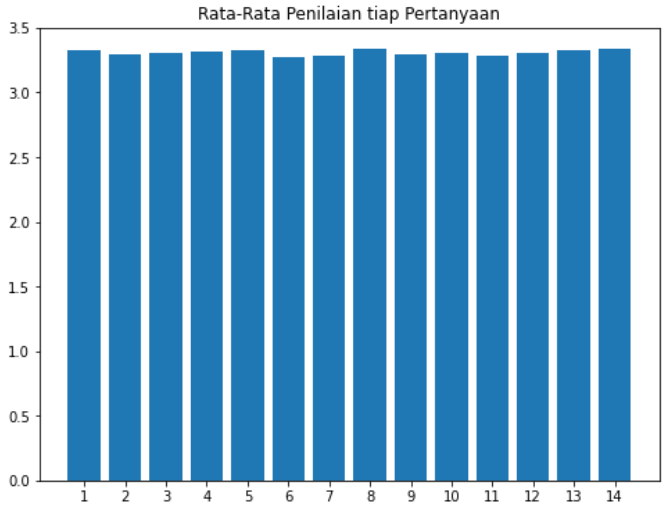
Gambar 16. Boxplot Nilai Evaluasi Dosen tiap Pertanyaan pada Dataframe 2

Berdasarkan plot di atas, setiap pertanyaan memiliki nilai rata-rata yang cenderung sama. Kemudian, diperoleh bahwa nilai rata-rata dosen per tahunnya tidak berbeda jauh.

TAHUN	NILAI
2015	3.269231
2016	3.267432
2017	3.363956
2018	3.351299

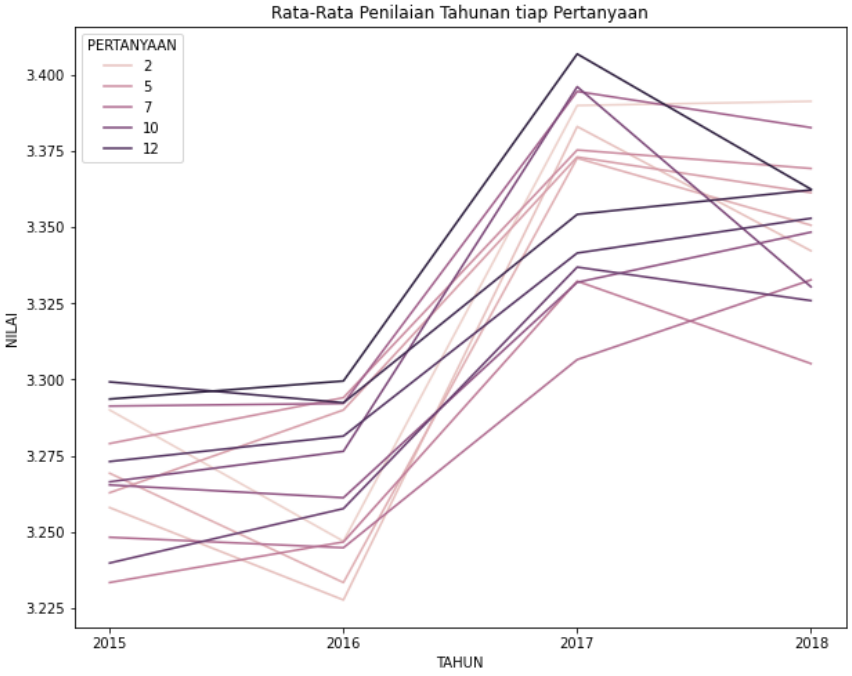
Gambar 17. Rata-Rata Nilai Evaluasi Dosen tiap Tahun

Hal ini juga berlaku untuk rata-rata penilaian setiap pertanyaan yang diberikan kepada mahasiswa.



Gambar 18. Rata-Rata Nilai Evaluasi Dosen tiap Pertanyaan

Terakhir, dilakukan plot *line chart* terhadap rata-rata penilaian dosen setiap pertanyaan per tahun dan diperoleh bahwa setiap pertanyaan memiliki rataan penilaian yang memiliki tren naik. Dengan demikian, dosen memiliki perkembangan kinerja secara keseluruhan.



Gambar 19. Rata-Rata Nilai Evaluasi Dosen Tahunan tiap Pertanyaan

c. *Dataframe* Ketiga: Transkrip Nilai Mahasiswa per NIM

Dataframe ketiga memiliki statistika deskriptif sebagai berikut.

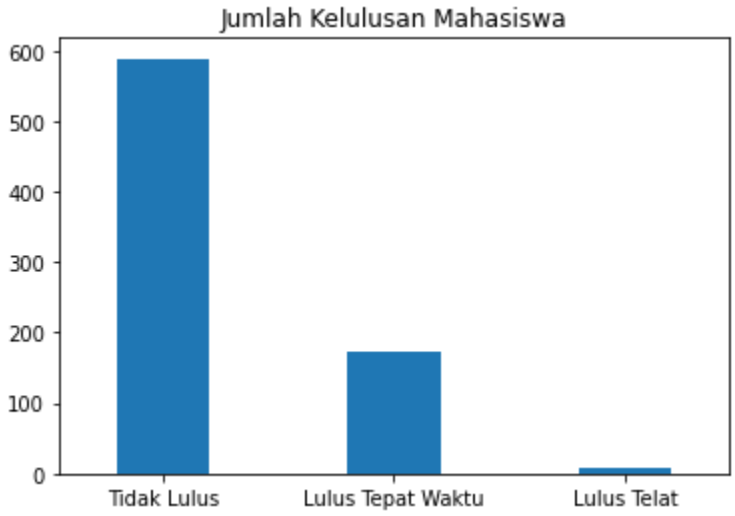
	TOTAL_SKS	TOTAL_BOBOT	WAKTU_KULIAH	IPK
count	770.000000	770.000000	770.000000	770.000000
mean	110.763636	337.295714	2.716234	2.943970
std	41.083773	153.354122	1.437582	0.608181
min	40.000000	0.000000	0.500000	0.000000
25%	82.000000	193.625000	1.500000	2.673277
50%	134.000000	395.850000	3.500000	3.034729
75%	145.000000	463.875000	3.500000	3.340000
max	148.000000	580.000000	6.500000	4.000000

Gambar 20. Statistika Deskriptif Dataframe 3

Diperoleh juga hasil dari setiap mahasiswa sebagai berikut.

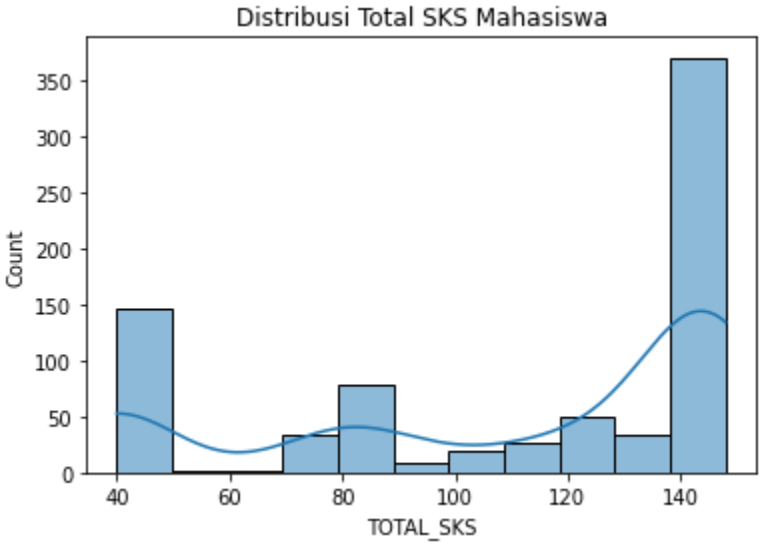
```
Tidak Lulus          589
Lulus Tepat Waktu    174
Lulus Telat            7
Name: HASIL, dtype: int64
```

Gambar 21. Jumlah Kelulusan Mahasiswa



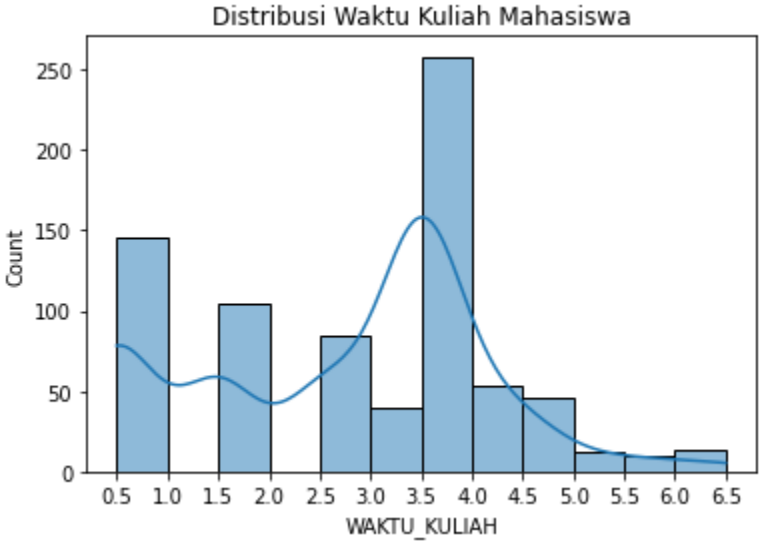
Gambar 22. Grafik Batang Jumlah Kelulusan Mahasiswa

Berdasarkan plot, kita dapat melihat bahwa masih banyak mahasiswa yang belum lulus karena belum memenuhi syarat kelulusan.



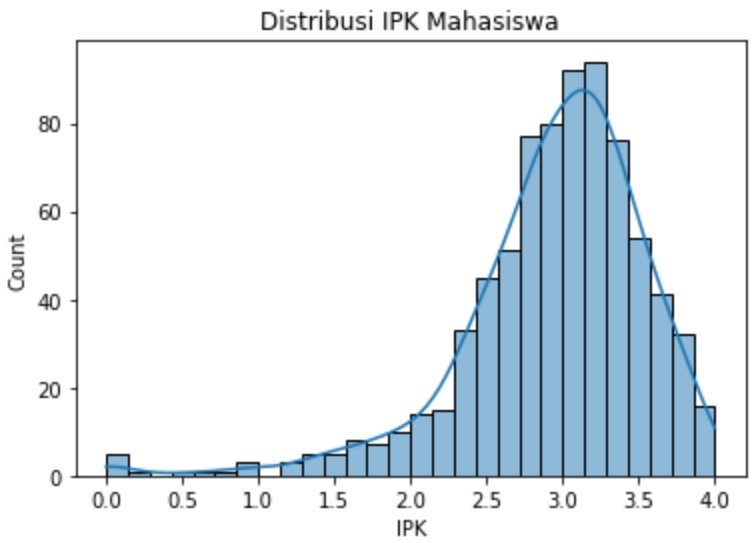
Gambar 23. Distribusi Total SKS Mahasiswa

Dari histogram tersebut, kita juga melihat bahwa sebagian besar mahasiswa telah mengambil SKS dengan total lebih dari 140SKS.



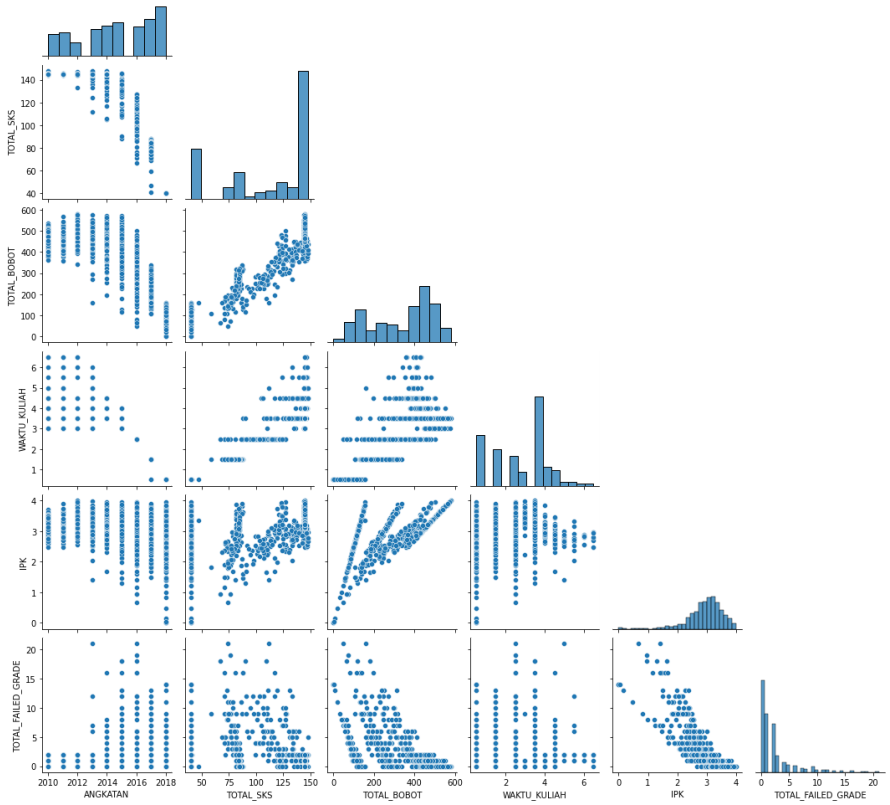
Gambar 24. Distribusi Waktu Kuliah Mahasiswa

Dari plot distribusi waktu kuliah mahasiswa (dalam tahun), kita melihat bahwa mayoritas mahasiswa berkuliah dengan durasi 3,5 hingga 4 tahun.



Gambar 25. Distribusi IPK Mahasiswa

Terlihat juga bahwa distribusi IPK mahasiswa menceng ke kiri (*negatively skewed*).

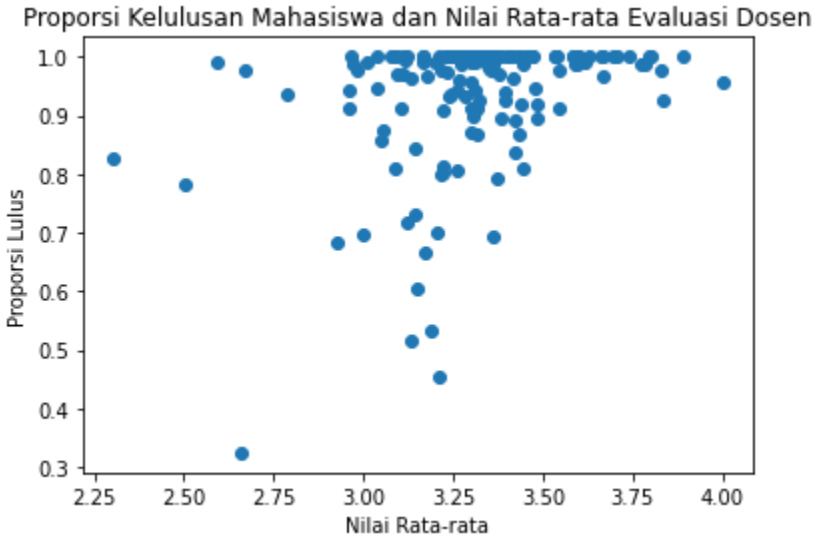


Gambar 26. *Scatter Matrix* Kolom pada *Dataframe 3*

Dari *scatter matrix*, diperoleh juga bahwa antar atribut pada *dataframe* ketiga ini tidak memiliki hubungan linier.

d. *Dataframe* Keempat: Evaluasi Dosen per Semester dan Mata Kuliah

Dataframe keempat berisi data evaluasi dosen per semester dan mata kuliah.

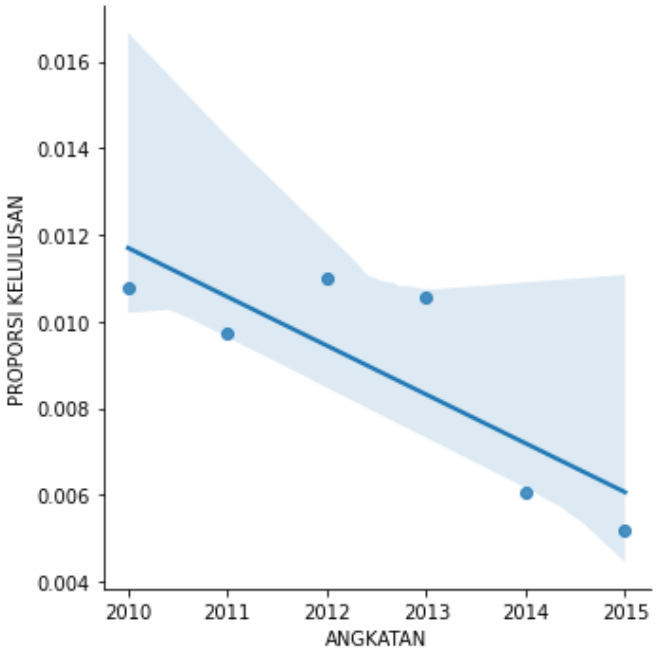


Gambar 27. *Scatter Plot* Proporsi Kelulusan Mahasiswa dan Nilai Rata-Rata Evaluasi Dosen

Dari *scatter plot* tersebut, terlihat bahwa nilai rata-rata penilaian dosen tidak menunjukkan adanya hubungan terhadap banyaknya mahasiswa yang lulus pada kelas yang bersangkutan. Dengan demikian, dapat disimpulkan bahwa nilai evaluasi dosen bukan faktor yang memengaruhi kelulusan mahasiswa.

e. Dataframe Kelima: Jumlah Kelulusan Mahasiswa Setiap Angkatan

Dataframe kelima berisi data kelulusan mahasiswa setiap angkatan.

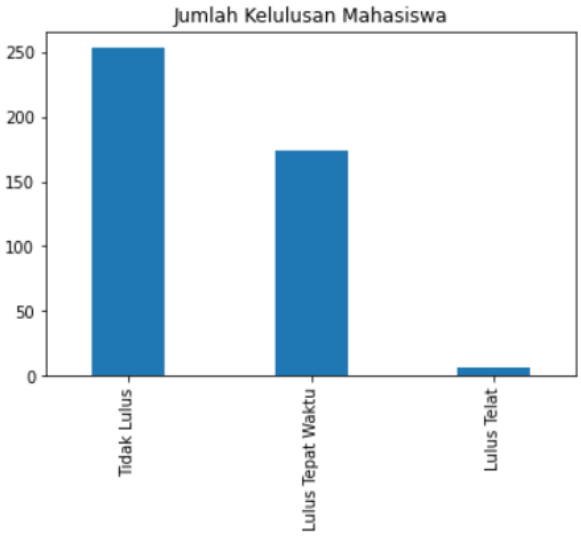


Gambar 28. Grafik Proporsi Kelulusan Mahasiswa tiap Angkatan

Dari *plot* tersebut, terlihat bahwa adanya tren menurun untuk jumlah mahasiswa yang lulus dari tahun 2010 hingga 2015. Plot dilakukan hingga tahun 2015 karena mahasiswa angkatan 2016 ke 2018 belum lulus karena belum memenuhi syarat SKS mengingat waktu studi yang telah ditempuh masih sedikit.

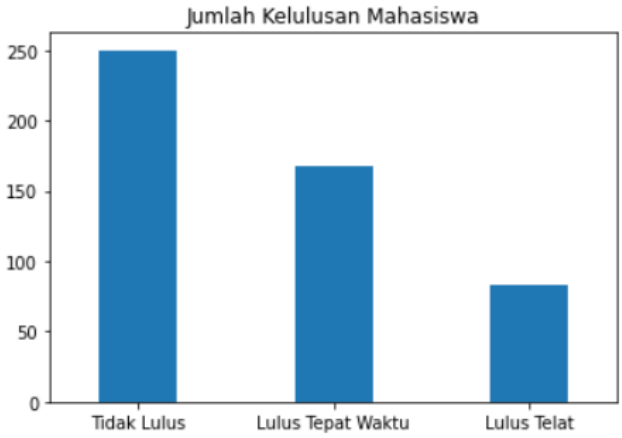
B. Modeling & Evaluation

Untuk memprediksi kelulusan mahasiswa, dibangun suatu model. Model dibangun dengan menggunakan *Dataframe* transkrip nilai mahasiswa untuk masing-masing NIM yang telah diberi label klasifikasi mahasiswa menjadi 3 kelas, yakni ‘Tidak Lulus’, ‘Lulus Telat’, dan ‘Lulus Tepat Waktu’. Dengan menggunakan *dataframe* yang telah diberi label, permasalahan model menjadi permasalahan klasifikasi *multiclass* sehingga digunakan model *machine learning* dengan *supervised learning styles*. Algoritma model *supervised learning* dibangun dengan memberikan input data atau *training data* klasifikasi mahasiswa sebagai proses pembelajaran.



Gambar 29. Data Jumlah Kelulusan Mahasiswa

Data yang akan digunakan untuk melakukan *training* pada model yang akan digunakan terlebih dahulu dilakukan proses sampling, karena persebaran data menunjukkan *imbalanced classification*, yang terlihat dari gambar 29, data dengan label 'Tidak Lulus' memiliki jumlah yang jauh lebih tinggi dibandingkan data dengan label 'Lulus Tepat Waktu' dan 'Lulus Telat', sehingga dilakukan proses *oversampling* dan *undersampling*. Namun, sebelum itu ditentukan terlebih dahulu variabel fitur atau atribut sebagai prediktor untuk memprediksi kelulusan mahasiswa. Fitur yang akan digunakan adalah 'TOTAL_FAILED_GRADED' dan 'IPK' karena banyaknya mata kuliah yang tidak lulus dan IPK memengaruhi kelulusan mahasiswa. Selanjutnya, data terlebih dahulu dilakukan *oversampling,* dan kemudian dilakukan *undersampling* dengan perbandingan 1:2:3 untuk 'Lulus Telat', 'Lulus Tepat Waktu', dan 'Tidak Lulus' secara berurutan dengan data 'Tidak Lulus' diambil sebanyak 295 data (setengah dari jumlah data aslinya).

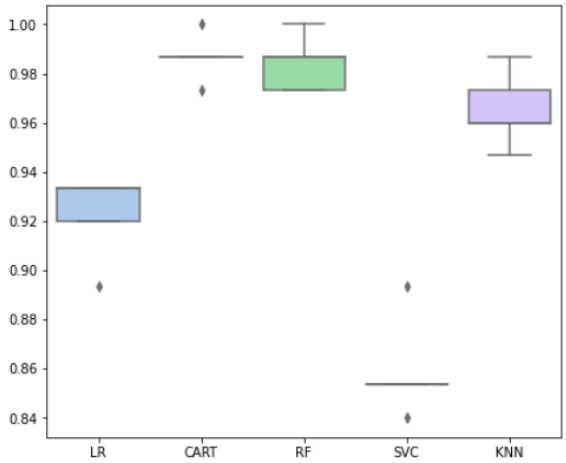


Gambar 30. Data Jumlah Kelulusan Mahasiswa Setelah Sampling

Setelah melakukan sampling data, data tersebut dibagi menjadi 75% *training set* dan 25% *validation set*. Lalu, untuk menentukan model yang akan digunakan, diperiksa akurasi 5 jenis model, yakni Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, dan K-Nearest Neighbours dengan menggunakan *cross validation* 5-*stratified fold*. Berikut rata-rata dan simpangan baku akurasi kelima model.

```
LR      | cv_score_mean: 0.9226666666666669     | cv_score_std: 0.015549205052920814
CART    | cv_score_mean: 0.9866666666666667     | cv_score_std: 0.008432740427115663
RF      | cv_score_mean: 0.984          | cv_score_std: 0.009977753031397158
SVC     | cv_score_mean: 0.8586666666666666     | cv_score_std: 0.018086213288334037
KNN     | cv_score_mean: 0.9653333333333333     | cv_score_std: 0.013597385369580781
```

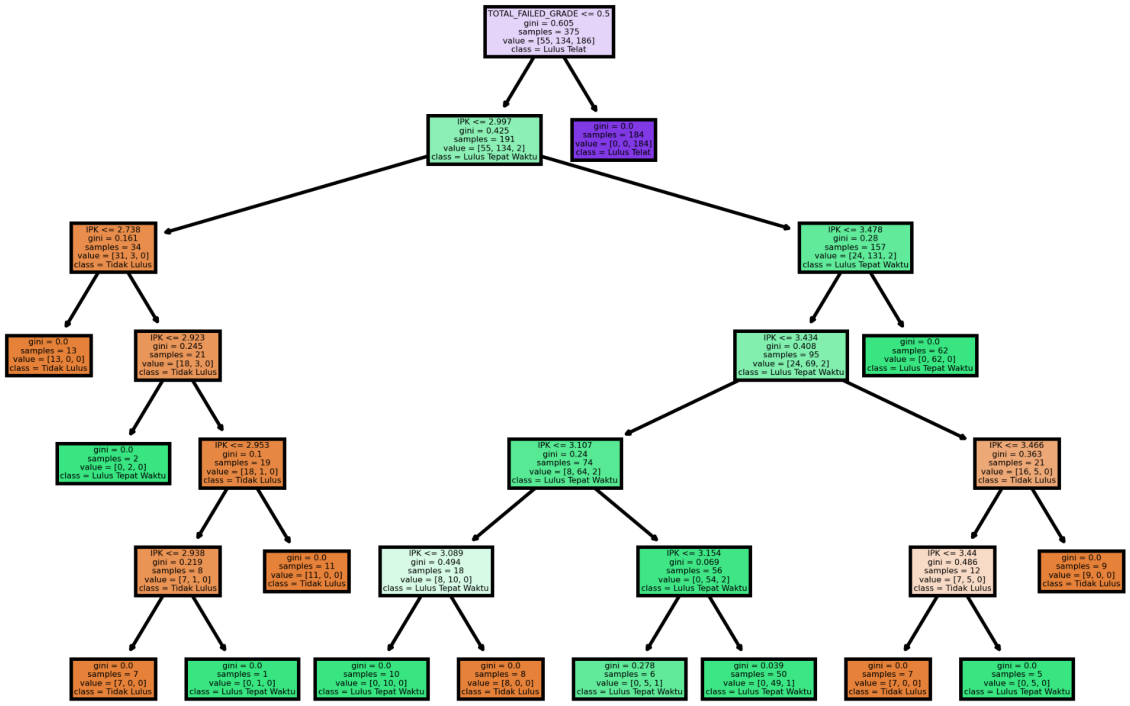
Gambar 31. Nilai Rata-Rata dan Simpangan Baku *Cross validation* dari Akurasi Berbagai Model



Gambar 32. Boxplot Akurasi Berbagai Model

Berdasarkan hasil akurasi model yang telah diuji dan standar deviasi, dapat disimpulkan model yang akan digunakan adalah Decision Tree Classifier.

Dalam membangun model Decision Tree, parameter Decision Tree dikalibrasi atau disesuaikan. Parameter Decision Tree yang dikalibrasi adalah maksimum tinggi atau kedalaman pohon. Maksimum tinggi pohon yang memiliki akurasi prediksi model yang terbaik adalah pohon dengan tinggi 6. Berikut gambar model



Gambar 33. Model Decision Tree Classifier

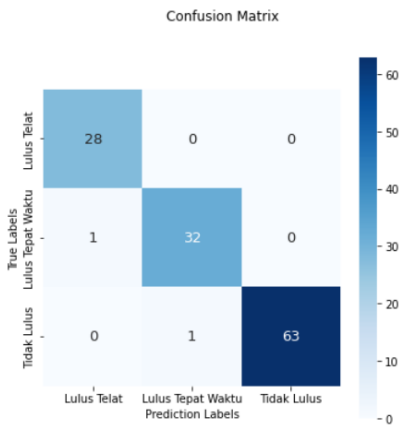
Setelah model Decision Tree dibangun, dilakukan evaluasi model menggunakan *validation set* yang telah dibagi sebelumnya. Berikut hasil evaluasi model.

```
DecisionTreeClassifier(max_depth=6) Evaluation
--------------------------------------------------------------------------------
accuracy score: 0.984

classification report:
                    precision    recall  f1-score   support

      Lulus Telat       0.97      1.00      0.98        28
Lulus Tepat Waktu       0.97      0.97      0.97        33
      Tidak Lulus       1.00      0.98      0.99        64

         accuracy                           0.98       125
        macro avg       0.98      0.98      0.98       125
     weighted avg       0.98      0.98      0.98       125
```



Gambar 34. Evaluasi Model Decision Tree Classifier

Keterangan:

True Positive (TP) = Hasil prediksi bernilai benar dan data asli bernilai benar

True Negative (TN) = Hasil prediksi bernilai salah dan data asli bernilai salah

False Positive (FP) = Hasil prediksi bernilai benar, tetapi data asli bernilai salah

False Negative (FN) = Hasil prediksi bernilai salah, tetapi data asli bernilai benar

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{28 + 32 + 63}{28 + 32 + 63 + 1 + 1} = \frac{123}{125} = 0.984$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F-1 Score} = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

$$\textit{Precision} \text{ Lulus Telat} = \frac{TP}{TP + FP} = \frac{28}{28 + 1} = \frac{28}{29} = 0.97$$

$$\textit{Precision} \text{ Lulus Tepat Waktu} = \frac{TP}{TP + FP} = \frac{32}{32 + 1} = \frac{32}{33} = 0.97$$

$$\textit{Precision} \text{ Tidak Lulus} = \frac{TP}{TP + FP} = \frac{63}{63} = 1$$

$$\textit{Recall} \text{ Lulus Telat} = \frac{TP}{TP + FN} = \frac{28}{28} = 1$$

$$\textit{Recall} \text{ Lulus Tepat Waktu} = \frac{TP}{TP + FN} = \frac{32}{32+1} = \frac{32}{33} = 0.97$$

Recall Tidak Lulus = $\frac{TP}{TP+FN} = \frac{63}{63+1} = \frac{63}{64} = 0.98$

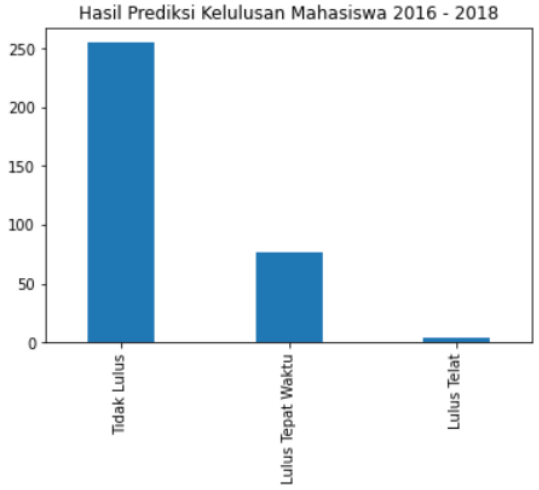
F-1 *Score* Lulus Telat = $\frac{2 * Precision * Recall}{(Precision + Recall)} = \frac{2*0.97*1}{0.97+1} = \frac{1.94}{1.97} = 0.98$

F-1 *Score* Lulus Tepat Waktu = $\frac{2 * Precision * Recall}{(Precision + Recall)} = \frac{2*0.97*0.97}{0.97+0.97} = 0.97$

F-1 *Score* Tidak Lulus = $\frac{2 * Precision * Recall}{(Precision + Recall)} = \frac{2*1*0.98}{1+0.98} = \frac{1.96}{1.98} = 0.99$

Berdasarkan nilai akurasi dan F-1 *score* yang cukup tinggi sehingga memiliki tingkat kesalahan yang rendah, model Decision Tree dengan maksimum kedalaman 4 dapat digunakan sebagai model untuk memprediksi kelulusan mahasiswa.

Model yang telah dilakukan proses *training* kemudian digunakan untuk memprediksi kelulusan dari mahasiswa pada tahun 2016 hingga 2018, didapatkan hasil sebagai berikut,



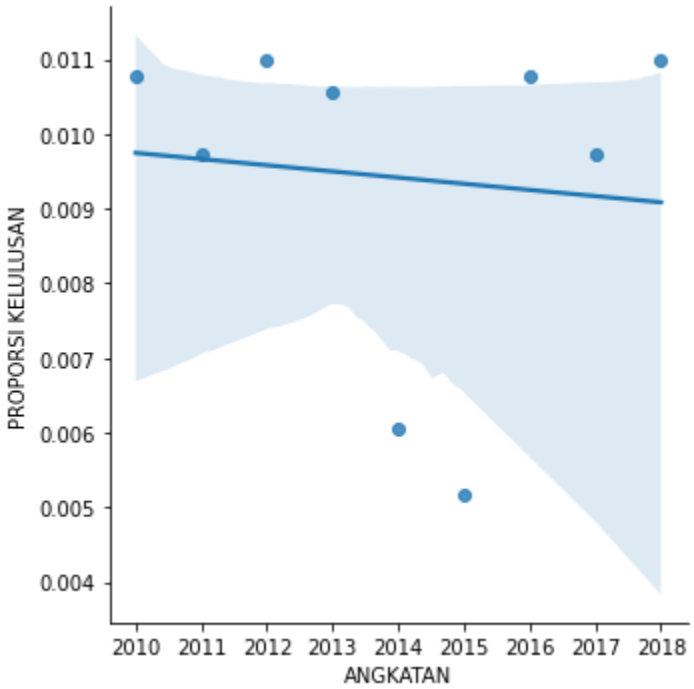
Gambar 35. Hasil Prediksi Kelulusan Mahasiswa 2016 - 2018

Hasil prediksi menunjukkan bahwa, data tertinggi adalah pada mahasiswa yang ‘Tidak Lulus’, yang apabila dilihat dari proporsi kelulusannya untuk tiap angkatan dari 2016 - 2018, didapatkan hasil sebagai berikut,

	ANGKATAN	JUMLAH KELULUSAN	JUMLAH MAHASISWA	PROPORSI KELULUSAN
0	2016	15	3395	0.010770
1	2017	25	3133	0.009720
2	2018	41	2016	0.010989

Gambar 36. Hasil Prediksi Kelulusan Mahasiswa 2016 - 2018

Data dari kelulusan mahasiswa pada tahun 2016 - 2018 apabila diplotkan dengan data dari 2010 - 2015, data menunjukkan bahwa tren 2016 - 2018 selaras dengan tren pada tahun 2010 - 2013.



Gambar 37. Tren Kelulusan Mahasiswa 2010 - 2018

CONCLUSION AND SUGGESTION

Berdasarkan hasil studi kasus dari dataset tersedia yaitu dataset nilai mahasiswa dari tahun 2010-2018 dan dataset evaluasi dosen per mata kuliah, didapatkan beberapa insight akhir yaitu,

1. Didapatkan bahwa, kualitas dari dosen atau hasil dari evaluasi dosen tidak mempengaruhi performa dari mahasiswa yang mengikuti kelasnya, hal ini ditunjukkan oleh tidak adanya hubungan linear / korelasi yang sangat rendah antara nilai rata-rata atau bahkan nilai per pertanyaan pada evaluasi dosen dengan proporsi kelulusan mahasiswa pada kelas yang bersangkutan.
2. Dengan menggunakan model Decision Tree Classifier, terdapat 2 buah data yang memiliki pengaruh besar terhadap waktu kelulusan dari mahasiswa, yaitu data IPK mahasiswa dan banyaknya mata kuliah yang tidak lulus (Mata kuliah dengan nilai dibawah C)
3. Berdasarkan tren kelulusan yang telah diplot dari angkatan 2010 hingga 2015, terlihat tren menurun pada proporsi kelulusan mahasiswanya. Akan tetapi, berdasarkan prediksi, terlihat bahwa, tren dari kelulusan mahasiswa sama dengan tren dari tahun 2010 hingga 2013. Dari analisis yang telah dilakukan, hal ini terjadi karena untuk mahasiswa angkatan 2014 dan 2015, terdapat banyak mahasiswa yang masih “belum lulus”.

Model Decision Tree yang telah dibentuk dapat digunakan untuk memprediksi kelulusan mahasiswa pada tahun 2016 hingga 2018, dengan menggunakan data pada kolom ‘TOTAL_FAILED_GRADE’ dan ‘IPK’. Dengan data tersebut, model ini dapat memprediksi apakah seorang mahasiswa “Lulus Tepat Waktu”, “Lulus Telat”, atau “Tidak Lulus”