LAPORAN
PEMBAHASAN HASIL STUDI KASUS
DATA COMPETITION ISFEST 2022

KELOMPOK Fortuna101:

1. **Indrayoga Putra**
2. **Joshua Febrian Wiranata**

INFORMATION SYSTEM FESTIVAL UMN

2022

BUSINESS UNDERSTANDING

Prestasi belajar mahasiswa merupakan salah satu indikator dalam menentukan peringkat suatu universitas. Untuk mengukur prestasi belajar dapat ditinjau dari beberapa aspek, seperti indeks prestasi mahasiswa, kualitas tenaga pengajar, dan lingkungan di universitas. Pada kesempatan kali ini, kami akan melakukan sejumlah analisis berdasarkan data nilai mahasiswa dan data evaluasi mata kuliah untuk mencari faktor-faktor penting studi mahasiswa serta mencari solusi dalam upaya meningkatkan prestasi belajar mahasiswa di universitas X.

Beberapa permasalahan yang akan kami tinjau dari kedua kami peroleh adalah:

1. Bagaimana tren kelulusan tepat waktu untuk mahasiswa masih menempuh pendidikan di universitas X?
2. Bagaimana pengaruh mata kuliah terhadap hasil studi mahasiswa?
3. Apa faktor-faktor yang mempengaruhi waktu studi mahasiswa agar dapat lulus tepat waktu?

Harapannya, hasil akhir berupa kesimpulan dari model machine learning yang kami buat dapat membantu universitas X dalam mengambil keputusan untuk meningkatkan prestasi belajarnya.

Analytical Approach

Dalam menyelesaikan permasalahan yang sudah dirumuskan, kami menggunakan bantuan machine learning dalam membuat beberapa model yang dapat membantu menjawab permasalahan tersebut. Model yang kami buat dan gunakan menggunakan sejumlah parameter berdasarkan data nilai mahasiswa maupun data evaluasi mata kuliah. Beberapa model yang akan digunakan:

- Decision Tree Classifier,Logistic Regression,Random Forest Classifier ,KNN, Support Vector Machine untuk memprediksi tren kelulusan mahasiswa tepat waktu ,dengan parameter keberhasilan angka
- Kmeans untuk membentuk klaster mata kuliah antara nilai evaluasi mata kuliah dan nilai ujian mata kuliah.

DATA UNDERSTANDING

Sebelum menerapkan model-model yang sudah dipaparkan, akan diperhatikan data yang akan digunakan sehingga tidak mengakibatkan kesalahan dalam pengolahan data dan penarikan kesimpulan.

Data Nilai Mahasiswa (Dataset1_TranscriptMahasiswa)

1) Identifikasi Isi dan Variabel Data

Pada data nilai mahasiswa memuat laporan nilai mahasiswa untuk semua mata kuliah yang diikuti pada suatu semester. Berikut merupakan variabel dan keterangan dari data nilai mahasiswa

- NIM: Nomor Induk Mahasiswa yang merupakan kode unik untuk setiap mahasiswa
- ANGKATAN: Tahun mahasiswa saat masuk universitas
- SEMESTER: Waktu pengambilan mata kuliah, dengan semester ganjil diikuti dengan nilai XY11 dan semester genap diikuti dengan nilai XY21
- NAMA_MK: Nama mata kuliah yang mahasiswa ambil
- KODE_MK: Kode unik untuk mewakili suatu nama mata kuliah
- SKS: Beban kuliah dari mata kuliah yang diambil
- NILAI: Nilai akhir yang diperoleh mahasiswa
- GRADE: Indeks yang bersesuaian dengan nilai akhir mahasiswa

Berikut merupakan beberapa sampel data nilai mahasiswa

```
df.head()
```

	NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE
0	10110310002	2010	1011	EM100	EM100 Dasar-dasar Bisnis	3	57.0	C
1	10110310002	2010	1011	EM180	EM180 Matematika Bisnis	3	70.0	B
2	10110310002	2010	1011	TI100	TI100 Algoritma dan Pemrograman	4	57.0	C
3	10110310002	2010	1011	TI101	TI101 Matematika Diskrit	3	59.0	C
4	10110310002	2010	1011	TI110	TI110 Pengantar Teknologi Multimedia	3	74.0	B

+ Code + Text

2) Format dan Kelengkapan Data

Pertama akan ditinjau kelengkapan isi dan tipe variabel dari data nilai mahasiswa

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30870 entries, 0 to 30869
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   NIM       30870 non-null  int64
 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  object
 5   SKS       30870 non-null  int64
 6   NILAI     30858 non-null  float64
 7   GRADE     30318 non-null  object
dtypes: float64(1), int64(4), object(3)
memory usage: 1.9+ MB
```

Berdasarkan informasi diatas diperoleh informasi bahwa jumlah baris pada data adalah 30870, sehingga masih terdapat sejumlah data yang kosong pada variabel NAMA_MK, NILAI, GRADE. Selain itu perlu diperhatikan untuk variabel NIM, ANGKATAN, SEMESTER memiliki tipe data yang tidak dapat langsung diolah informasinya.

Kemudian pada variabel SEMESTER, terdapat data dengan kode xy22, di mana nilai tersebut tidak terdefinisi.

```
df["SEMESTER"].unique() # indikasi perlu di bersihkan untuk kode xx22
```

```
array([1011, 1021, 1111, 1121, 1211, 1221, 1311, 1321, 1322, 1122, 1222,
       1411, 1421, 1022, 1422, 1511, 1621, 1521, 1611, 1711, 1721, 1722,
       1522, 1821, 1811, 1622, 1822, 1911])
```

Pada variabel ANGKATAN sudah memuat nilai yang sesuai dan angkatan yang terdapat dalam data disajikan dalam informasi dibawah ini.

```
[2010 2011 2012 2013 2014 2015 2016 2017 2018]
```

Selain itu jumlah nama mata kuliah yang berbeda jauh lebih banyak dibandingkan jumlah kode mata kuliah yang berbeda. Hal ini mengindikasikan terdapat nama mata kuliah yang salah ketik, terdapat duplikat pada nama kuliah, atau kesalahan lainnya .

```
print('Banyak kategori variabel KODE_MK adalah ',df['KODE_MK'].nunique())
print('Banyak kategori variabel NAMA_MK adalah ',df['NAMA_MK'].nunique())
```

```
Banyak kategori variabel KODE_MK adalah  142
Banyak kategori variabel NAMA_MK adalah  175
```

Contoh data yang memiliki kode mata kuliah yang sama, tetapi nama mata kuliahnya berbeda:

1. Kode mata kuliah tertulis ulang

	NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE
0	10110310002	2010	1011	EM100	EM100 Dasar-dasar Bisnis	3	57.0	C
87	10110310004	2010	1311	EM100	Dasar-dasar Bisnis	3	73.0	B

2. Kode mata kuliah berbeda yang berisi nama mata kuliah yang sama

UM222	Bahasa Inggris 2	2	182
	UM222 Bahasa Inggris 2	2	53
UM223	Bahasa Inggris 2	2	535

Di samping itu, untuk variabel NILAI dan SKS, terlihat bahaw tidak ada indikasi data yang *error* atau nilainya terdefinisi sesuai dengan rentang yang seharusnya.

	NILAI	SKS
count	30858.000000	30870.000000
mean	74.357930	2.838063
std	15.650969	0.564980
min	0.000000	2.000000
25%	66.000000	3.000000
50%	74.000000	3.000000
75%	85.000000	3.000000
max	100.000000	6.000000

Selain itu masih terdapat sejumlah data yang diindikasikan merupakan duplikat karena terdapat mahasiswa yang mengambil lebih dari 1 mata kuliah yang sama pada satu semester.

```
df[df.duplicated(subset=['NIM','KODE_MK','GRADE','SEMESTER'],keep=False)]
```

	NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE
345	10110310011	2010	1011	EM100	Dasar-dasar Bisnis	3	71.0	B
346	10110310011	2010	1011	EM100	EM100 Dasar-dasar Bisnis	3	74.0	B
347	10110310011	2010	1011	EM180	EM180 Matematika Bisnis	3	73.0	B
348	10110310011	2010	1011	EM180	Matematika Bisnis	3	74.0	B
349	10110310011	2010	1011	TI100	Algoritma dan Pemrograman	4	72.0	B
...	...	...	...	...	...	...	...	...
6876	11110310080	2011	1411	SI725	SI725 Knowledge Management	3	79.0	B+
6877	11110310080	2011	1411	SI729	SI729 Sistem dan Aplikasi Perusahaan 3	3	83.0	A-
6878	11110310080	2011	1411	SI729	Sistem dan Aplikasi Perusahaan 3	3	84.0	A-
6881	11110310080	2011	1411	SI863	SI863 Tugas Akhir	6	78.0	B+
6882	11110310080	2011	1411	SI863	NaN	6	78.0	B+

3) Menilai dan Mengevaluasi Kualitas Data

Berdasarkan informasi yang sudah digali dari data nilai mahasiswa, dapat dinilai bahwa data ini masih tergolong kurang baik, hal ini dikarenakan variabel yang tipe datanya tidak sesuai dengan definisinya, terdapat sejumlah data kosong, adanya ketidaksesuaian data pada variabel SEMESTER, terdapat kasus ketidaksesuaian antara kode mata kuliah dengan nama mata kuliah, serta terdapat indikasi data yang duplikat.

Data Evaluasi Mata Kuliah (Dataset2_EvaluasiDosenPerMK)

1) Identifikasi Isi dan Variabel Data

Data Evaluasi Mata Kuliah merupakan data yang berisi laporan penilaian evaluasi pengajar mata kuliah pada suatu semester. Berikut merupakan variabel dan keterangan dari setiap variabel untuk Data Evaluasi Mata kuliah:

- TAHUN: Tahun ajaran saat performa pengajar dievaluasi
- SEMESTER: Semester pada saat performa pengajar dievaluasi dengan ketentuan sama seperti data nilai mahasiswa
- MATAKULIAH: Mata kuliah yang diambil pengajar
- PERTANYAAN: Nomor berdasarkan kategori pertanyaan evaluasi
- KETERANGAN: Keterangan aspek penilaian yang dievaluasi dosen
- NILAI: Hasil penilaian evaluasi pengajar berdasarkan kategori keterangan

Berikut merupakan beberapa sampel data nilai mahasiswa

```
dfs.head()
```

	TAHUN	SEMESTER	MATAKULIAH	PERTANYAAN	KETERANGAN	NILAI
0	2015	1511	IS100 Management Information Systems	1	Kesiapan memberikan perkuliahan/praktikum	3.28
1	2015	1511	IS100 Management Information Systems	2	Upaya menyampaikan materi perkuliahan/praktik...	3.25
2	2015	1511	IS100 Management Information Systems	3	Sistematis dalam menyampaikan materi perkuliah...	3.24
3	2015	1511	IS100 Management Information Systems	4	Kemampuan memberikan contoh yang relevan dari ...	3.30
4	2015	1511	IS100 Management Information Systems	5	Penyampaian materi perkuliahan sesuai dengan k...	3.27

2) Format dan Kelengkapan Data

Pertama akan ditinjau kelengkapan isi dan tipe variabel dari data evaluasi mata kuliah

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 99.2+ KB
```

Diperoleh bahwa terdapat 2114 baris pada data yang semuanya memiliki nilai. Perlu diperhatikan juga bahwa terdapat beberapa variabel yang tipe datanya tidak sesuai dengan definisi yang seharusnya, sehingga harus dilakukan pengolahan lanjut. Selain itu, tidak terindikasi bahwa terdapat nilai error pada data evaluasi mata kuliah.

```
Kategori pada variabel SEMESTER adalah [1511 1521 1611 1621 1711 1721 1811 1821]
Kategori pada variabel TAHUN adalah [2015 2016 2017 2018]
Kategori pada variabel PERTANYAAN  [ 1  2  3  4  5  6  7  8  9 10 11 12 13 14]
```

Terakhir, tidak terdapat adanya ada duplikat pada data ini

```
df_d[df_d.duplicated(subset=['TAHUN','SEMESTER','MATAKULIAH','KETERANGAN'],keep=False) ] #tidak ada data duplikat
```

TAHUN	SEMESTER	MATAKULIAH	PERTANYAAN	KETERANGAN	NILAI

3) Menilai dan Mengevaluasi Kualitas Data

Kualitas data evaluasi mata kuliah sudah baik karena data tersebut dapat dikategorikan sudah bersih.

DATA PREPARATION

Data Nilai Mahasiswa (Dataset1_TranscriptMahasiswa)

1) Variabel

Variabel yang kami gunakan berasal dari data nilai mahasiswa atau variabel yang merupakan transformasi dari variabel yang sudah ada. Selain itu untuk mengolah data, kami mengubah tipe variabel untuk variabel SEMESTER, ANGKATAN, NIM.

```
df['SEMESTER']=df['SEMESTER'].astype('str')
df['ANGKATAN']=df['ANGKATAN'].astype('str')
df['NIM']=df['NIM'].astype('str')
```

2) Data Cleaning

Data Filling

1. Variabel Nilai

   Dalam mengisi nilai yang kosong pada variabel NILAI, perhatikan bahwa semua data yang kosong memiliki indeks GRADE bernilai ‘F’. Sehingga kita tidak perlu mengisi data yang kosong pada variabel NILAI.

```
df[df.NILAI.isna()]
```

	NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE
18866	9931	2015	1811	IS341	Sistem Basis Data	3	NaN	F
21965	12407	2015	1621	IS432	Pengantar E-Business	3	NaN	F
22633	13019	2016	1821	IS670	Audit Sistem Informasi	3	NaN	F
25920	19365	2017	1811	IS341	Sistem Basis Data	3	NaN	F
26030	19601	2017	1721	IS201	Proses Bisnis Korporat	3	NaN	F
26760	21116	2017	1711	IF110	Pengantar Teknologi Multimedia	3	NaN	F
26901	21159	2017	1821	UM321	Bahasa Inggris 3	2	NaN	F
27520	22742	2017	1721	IS230	Algoritma dan Pemrograman	3	NaN	F
27774	23564	2017	1721	IS230	Algoritma dan Pemrograman	3	NaN	F
29423	28446	2018	1821	IS220	Interaksi Manusia dan Komputer	3	NaN	F
29723	29747	2018	1811	IF100	Dasar-Dasar Pemrograman	3	NaN	F
30066	31318	2018	1821	IS201	Proses Bisnis Korporat	3	NaN	F

2. Variabel Grade

Untuk mengisi data kosong pada variabel GRADE, akan dilakukan transformasi nilai akhir yang diperoleh mahasiswa berdasarkan rentang indeks yang berlaku di universitas tersebut.

3. Variabel NAMA_MK
   Data yang kosong pada NAMA_MK dapat disesuaikan pasangan KODE_MK yang bersesuaian dengan nama mata kuliah tersebut.

Data Transformation

1. Nilai semester yang tidak memiliki definisi
   Karena untuk data dengan entri XY22 tidak terdefinisi, maka kami akan menganalisis lebih lanjut terkait data nilai tersebut.

SEMESTER	SKS
1022	50
1122	160
1222	84
1322	127
1422	71
1522	28
1622	127
1722	300
1822	358

Jika meninjau dari persebaran data, dapat dilihat persebaran jumlah SKS untuk semester XY22 tidaklah banyak, sehingga kami mengasumsikan hal ini merupakan kesalahan input pada variabel SEMESTER. Lebih lanjut jika melihat jumlah SKS yang diambil setiap semester, dapat dilihat jumlah SKS pada semester ganjil sedikit lebih banyak daripada semester genap. Dengan ini kami mengasumsikan bahwa entri XY22 merupakan entri semester genap, dengan alasan bahwa dengan penambahan entri ini maka perbedaan jumlah SKS pada semester ganjil dan genap tidak jauh berbeda.

2. Data duplikat

   Perhatikan contoh sampel data di bawah

	NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE
345	10110310011	2010	1011	EM100	Dasar-dasar Bisnis	3	71.0	B
346	10110310011	2010	1011	EM100	EM100 Dasar-dasar Bisnis	3	74.0	B
347	10110310011	2010	1011	EM180	EM180 Matematika Bisnis	3	73.0	B
348	10110310011	2010	1011	EM180	Matematika Bisnis	3	74.0	B

Terdapat data duplikat dengan NIM serupa yang mengambil mata kuliah dengan kode yang sama di satu semester. Dalam hal ini kami menggunakan asumsi bahwa nilai yang kami ambil sebagai nilai akhir merupakan nilai baris terakhir dari data duplikat. Asumsi ini dikarenakan kami beranggapan bahwa terjadi revisi pada nilai akhir mahasiswa, sehingga sangat memungkinkan adanya input berulang.

Validasi nilai data

Setelah melakukan data filling dan data transformation, diperoleh jumlah data yang kosong untuk setiap variabelnya kecuali variabel NILAI sudah tidak ada. Nilai kosong pada variabel nilai diperoleh akibat indikasi kecurangan oleh mahasiswa sehingga GRADE menjadi ‘F’.

```
(NIM          0
 ANGKATAN     0
 SEMESTER     0
 KODE_MK      0
 NAMA_MK      0
 SKS          0
 NILAI       12
 GRADE        0
```

Analisis lanjut dalam mencari duplikat data

Untuk mencari indikasi data duplikat lanjut, kami ingin mencari fenomena mahasiswa yang mengambil mata kuliah yang sama di semester yang berbeda.

Hasilnya, hanya dijumpai 1 mahasiswa yang mengambil suatu mata kuliah lebih dari 1 kali. Maka dari itu kami akan melakukan pembersihan data dengan cara membuang data dengan variabel NILAI terkecil.

```
df[df.duplicated(subset=['NIM','KODE_MK'],keep=False)]
```

	NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	SKS	NILAI	GRADE
12407	13110310069	2013	1711	IS571	Tata Kelola Teknologi Informasi 1	3	53.0	D
12415	13110310069	2013	1811	IS571	Tata Kelola Teknologi Informasi 1	3	7.0	E

3) Feature Engineering

Mencari Indeks Prestasi Semester

Untuk mencari nilai Indeks Prestasi Semester (IPS) mahasiswa setiap semesternya, kami membuat variabel baru yang merupakan kombinasi dari variabel NILAI, SKS, dan jumlah SKS pada semester yang bersangkutan.

			SKS/semester	BOBOTxSKS	IPS
NIM	ANGKATAN	SEMESTER			
10064	**2015**	**1511**	21	77.8	3.704762
		1521	22	83.0	3.772727
		1611	21	73.3	3.490476
		1621	20	74.3	3.715000
		1711	20	74.3	3.715000
		1721	22	88.0	4.000000
		1811	15	58.2	3.880000
		1821	4	14.8	3.700000
10082	**2015**	**1511**	21	56.6	2.695238
		1521	19	64.5	3.394737

Mencari Indeks Prestasi Kumulatif

Indeks Prestasi Kumulatif (IPK) dapat dicari dengan mencari rataan dari seluruh semester yang diikuti oleh mahasiswa.

	TOTAL_IPS	TOTAL_SEMESTER	IPK
NIM			
10064	29.977965	8	3.747246
10082	22.931968	7	3.275995
10110310002	23.296627	8	2.912078
10110310004	29.790209	10	2.979021
10110310005	29.824176	8	3.728022

Mencari semester terakhir yang diikuti mahasiswa

Untuk keperluan model dan menentukan status waktu kelulusan mahasiswa, akan dicari reka jejak semester terakhir yang ditempuh oleh mahasiswa.

Mencari syarat kelulusan

Dalam menentukan status keterangan kelulusan mahasiswa, perlu diuji apakah mahasiswa masih mempunyai indeks ‘D’, ‘E’, ‘F’ pada variabel GRADE.

Contoh hasil pengolahan dataframe

	NIM	TOTAL_IPS	TOTAL_SEMESTER	IPK	SKS	indi	SEMESTER_LAST
0	10064	29.977965	8	3.747246	145	0	8
1	10082	22.931968	7	3.275995	136	1	7
2	10110310002	23.296627	8	2.912078	145	1	8
3	10110310004	29.790209	10	2.979021	145	1	10
4	10110310005	29.824176	8	3.728022	145	0	8

Mencari Keterangan Kelulusan Mahasiswa

Akan ditentukan keterangan kelulusan dari mahasiswa berdasarkan syaratsyarat kelulusan mahasiswa, seperti lulus minimal 144 SKS dan tidak memiliki indeks D/E/F. Kami menglasifikasikan keterangan kelulusan menjadi empat kategori yaitu :

- lulus_lebih_cepat: mahasiswa yang lulus dengan waktu kurang dari 4 tahun
- lulus_tepat_waktu: mahasiswa lulus dengan waktu 4 tahun
- lulus_telat: mahasiswa yng lulus lebih dengan waktu lebih dari 4 tahun
- belum_lulus: Mahasiswa yang belum lulus hingga saat ini

	NIM	TOTAL_IPS	TOTAL_SEMESTER	IPK	SKS	indi	SEMESTER_LAST	keterangan_lulus
0	10064	29.977965	8	3.747246	145	0	8	lulus_tepat_waktu
1	10082	22.931968	7	3.275995	136	1	7	belum_lulus
2	10110310002	23.296627	8	2.912078	145	1	8	belum_lulus
3	10110310004	29.790209	10	2.979021	145	1	10	belum_lulus
4	10110310005	29.824176	8	3.728022	145	0	8	lulus_tepat_waktu

Data Evaluasi Mata Kuliah (Dataset2_EvaluasiDosenPerMK)

1) Variabel yang digunakan

Variabel yang kami gunakan berasal dari data nilai evaluasi mata kuliah atau variabel yang merupakan transformasi dari variabel yang sudah ada.

2) Data Cleaning

Ubah tipe data

```
df_d['TAHUN']=df_d['TAHUN'].astype('str')
df_d['SEMESTER']=df_d['SEMESTER'].astype('str')
df_d['PERTANYAAN']=df_d['PERTANYAAN'].astype('int')
```

Validasi kebersihan data

Dapat dilihat pada gambar di samping bahwa sudah tidak terdapat data yang kosong. Selain itu tipe data sudah sesuai sehingga data siap untuk diolah.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2114 entries, 0 to 2113
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   TAHUN        2114 non-null   object
 1   SEMESTER     2114 non-null   object
 2   MATAKULIAH   2114 non-null   object
 3   KODE_MK      2114 non-null   object
 4   PERTANYAAN   2114 non-null   int32
 5   KETERANGAN   2114 non-null   object
 6   NILAI        2114 non-null   float64
dtypes: float64(1), int32(1), object(5)
memory usage: 107.5+ KB
```

```
: (TAHUN          0
   SEMESTER       0
   MATAKULIAH     0
   KODE_MK        0
   PERTANYAAN     0
   KETERANGAN     0
   NILAI          0
   dtype: int64,
   None)
```

3) Feature engineering

Variabel kode mata kuliah

Kami membentuk variabel untul mengidentifikasi kode mata kuliah (KODE_MK). Variabel ini nantinya akan digunakan untuk mencocokan kode matkul dengan data nilai mahasiswa.

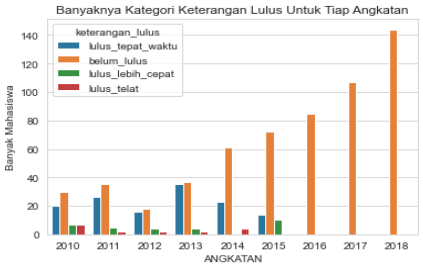
```
df_d['KODE_MK'] =df_d['MATAKULIAH'].str.slice(0, 5)
df_d=df_d[['TAHUN','SEMESTER','MATAKULIAH','KODE_MK','PERTANYAAN','KETERANGAN','NILAI']]
df_d.head()
```

	TAHUN	SEMESTER	MATAKULIAH	KODE_MK	PERTANYAAN	KETERANGAN	NILAI
0	2015	1511	IS100 Management Information Systems	IS100	1	Kesiapan memberikan perkuliahan/praktikum	3.28
1	2015	1511	IS100 Management Information Systems	IS100	2	Upaya menyampaikan materi perkuliahan/praktik...	3.25
2	2015	1511	IS100 Management Information Systems	IS100	3	Sistematis dalam menyampaikan materi perkuliah...	3.24
3	2015	1511	IS100 Management Information Systems	IS100	4	Kemampuan memberikan contoh yang relevan dari ...	3.30
4	2015	1511	IS100 Management Information Systems	IS100	5	Penyampaian materi perkuliahan sesuai dengan k...	3.27

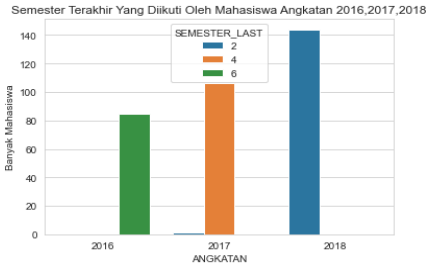
PREDICTION MODEL AND EVALUATION

A. Exploratory Data Analysis (EDA)

1. Analisa Angkatan Mahasiswa



Gambar A. 1 Histogram Keterangan Kelulusan untuk setiap angkatan



Gambar A. 2 Histogram semester terakhir yang ditempuh oleh 2016-2018

Pada Gambar A.1, terlihat bahwa masih banyak mahasiswa yang belum lulus, bahkan semua mahasiswa angkatan 2016-2018 belum lulus. Setelah ditelusuri lebih lanjut, terlihat pada Gambar A.2 belum ada mahasiswa yang menempuh semester delapan. Hal ini mengakibatkan jumlah SKS yang baru ditempuh masih kurang dari 144. Oleh karena itu kami menggunakan asumsi dan penyederhanaan masalah berdasarkan data sebagai berikut:

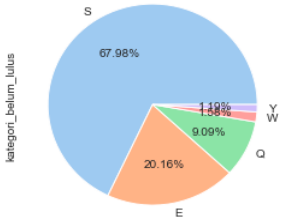
- Mahasiswa angkatan 2018 adalah mahasiswa tingkat 1 karena semester terakhir yang ditempuh oleh semua mahasiswanya adalah semester dua.
- Mahasiswa angkatan 2017 adalah mahasiswa tingkat 2 karena semester terakhir yang ditempuh oleh semua mahasiswanya adalah semester empat.
- Mahasiswa angkatan 2016 adalah mahasiswa tingkat 3 karena semester terakhir yang ditempuh oleh semua mahasiswanya adalah semester enam.

Selanjutnya kami akan menganalisa terkait penyebab masih ada mahasiswa angkatan 2010-2015 yang belum lulus.

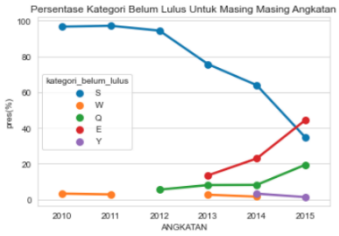
2. Analisa Faktor yang Menghambat Kelulusan Mahasiswa Angkatan 2010-2015

Kami ingin menganalisa faktor yang menyebabkan kebanyakan mahasiswa angkatan 2010- 2015 belum lulus. Kami membagi penyebab ketidaklulusan mahasiswa menjadi sejumlah kategori berdasarkan syarat kelulusan yang berlaku.

- Kategori Q: Masih memiliki indeks D/E/F, IPK < 2.5, dan jumlah SKS yang lulus <144
- Kategori W: Masih memiliki indeks D/E/F dan IPK < 2.5
- Kategori E: Masih memiliki indeks D/E/F dan jumlah SKS yang lulus <144
- Kategori R: IPK <2.5 dan jumlah SKS yang lulus <144
- Kategori S: Masih memiliki indeks D/E/F
- Kategori T: IPK < 2.5
- Kategori Y: Jumlah SKS yang lulus <144



Gambar A. 3 Diagram kategori untuk mahasiswa yang belum lulus



Gambar A. 4 Proporsi persentase untuk kategori setiap tahunnya

Berdasarkan gambar A.3, diperoleh kategori yang paling didominasi adalah kategori S, E, Q, W, dan Y dengan faktor utama penyebab mahasiswa belum lulus adalah masih memiliki indeks D/E/F. Hal ini diperkuat gambar A. 4, terkait kategori S yang menjadi faktor utama penyebab mahasiswa belum lulus di hampir setiap angkatan. Masalah ini mendorong kami untuk melihat hubungan antara pengaruh dosen di mata kuliah dengan nilai ujian yang diperoleh mahasiswa.

Hubungan Nilai Evaluasi Mata Kuliah dan Nilai Ujian

KODE_MK	NILAI	1	2	3	4	5	6	7
IS100	78.287281	3.462500	3.375000	3.362500	3.470000	3.450000	3.045000	3.240000
IS110	61.979452	3.110000	2.993333	3.073333	3.056667	3.130000	3.350000	2.706667
IS155	77.165517	3.440000	3.410000	3.400000	3.440000	3.420000	3.220000	3.440000
IS201	73.330396	3.362857	3.301429	3.282857	3.311429	3.328571	3.262857	3.260000
IS220	72.348519	3.230000	3.172500	3.215000	3.175000	3.250000	3.180000	3.242500
IS228	75.000000	3.530000	3.515000	3.540000	3.490000	3.505000	3.530000	3.505000
KODE_MK	NILAI	8	9	10	11	12	13	14
IS100	78.287281	3.392500	3.287500	3.337500	3.380000	3.362500	3.372500	3.465000
IS110	61.979452	3.136667	3.046667	3.403333	2.680000	3.096667	3.133333	3.193333
IS155	77.165517	3.400000	3.400000	3.380000	3.400000	3.400000	3.430000	3.350000
IS201	73.330396	3.265714	3.247143	3.272857	3.261429	3.271429	3.315714	3.357143
IS220	72.348519	3.300000	3.210000	3.212500	3.215000	3.280000	3.285000	3.232500

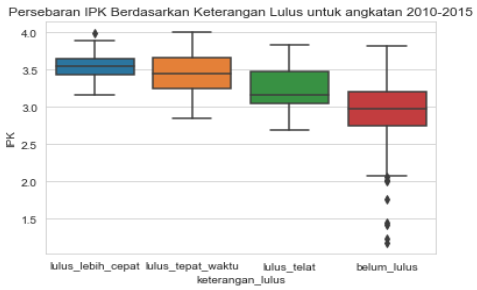
Gambar A. 5 Rataan nilai semua kode mata kuliah untuk setiap pertanyaan

	NILAI
NILAI	1.000000
1	0.622218
2	0.681035
3	0.661668
4	0.717488
5	0.670418
6	0.632448
7	0.687628
8	0.709802
9	0.699419
10	0.691562
11	0.724284
12	0.721236
13	0.750475
14	0.659519

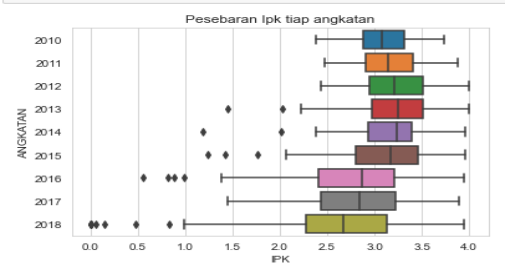
Gambar A. 6 Korelasi pengaruh setiap pertanyaan terhadap nilai mahasiswa

Hubungan antara nilai evaluasi mata kuliah dan nilai ujian mahasiswa dicari dengan menghitung nilai korelasi antar variabel pada data pada gambar A.5 yang merupakan rata rata nilai ujian dan rata-rata tiap pertanyaan evaluasi untuk setiap matkul. Pada Gambar A.6, terlihat hubungan antara rataan nilai evaluasi untuk tiap pertanyaan dengan rataan nilai ujian memiliki nilai korelasi yang cukup besar yaitu lebih besar dari 0,6 maka dapat disimpulkan nilai evaluasi memiliki pengaruh pada nilai ujian Dengan demikian karena nilai evaluasi mata kuliah mewakili kinerja dosen pada mata kuliah tersebut maka kualitas dosen memiliki hubungan korelasi positif dengan nilai ujian yang mempengaruhi kelulusan tepat waktu mahasiswa.

4) Analisa IPK



Gambar A. 7 Plot Persebaran IPK berdasarkan ketergan lulus untuk angkatan 2010-2015

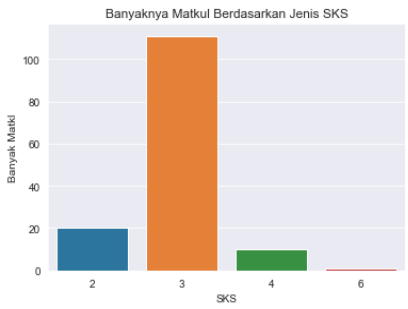


Gambar A. 8 Plot Persebaran IPK untuk setiap angkatan

Dengan meninjau persebaran IPK untuk setiap kategori kelulusan pada gambar A. 7 terlihat bahwa mahasiswa yang lulus tepat waktu memiliki nilai rataan IPK yang lebih tinggi dengan variansinya lebih kecil dibandingkan dengan kategori lainnya. Selanjutnya dengan meninjau persebaran IPK untuk setiap angkatan (gambar A.8) dapat dilihat bahwa angkatan 2018 dan 2016 memiliki persebaran yang cukup variatif. Mengingat kedua angkatan tersebut masing-masing masih berada di jenjang semester 2 dan semester 4, maka kita belum dapat menentukan kesimpulan terkait fenomena tersebut.

5) Analisa Mata Kuliah

Berdasarkan informasi data terdapat sejumlah mata kuliah dengan bobot SKS yang berbeda. Dengan melakukan distribusi persebaran beban SKS pada setiap mata kuliah (gambar A.9), diperoleh informasi tambahan bahwa terdapat banyak mata kuliah dengan beban 3 SKS. Selain itu terdapat juga mata kuliah yang hanya diikuti oleh 1 peserta mata kuliah, yakni dengan dengan kode IF501 (gambar A.10). Hal ini menjadi faktor penting dalam menganalisa mata kuliah yang memberikan pengaruh baik pada nilai mahasiswa.



Gambar A. 9 Histogram beban SKS mata kuliah yang terdapat pada data

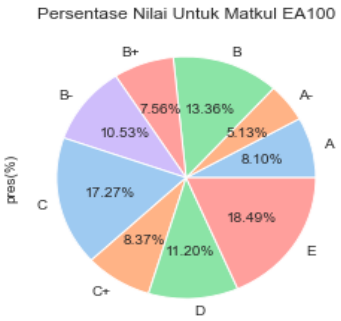
KODE_MK	Banyaknya Diambil
IF501	1
IF502	4
IF551	8
IF561	2

Gambar A. 10 Kode mata kuliah yang diikuti oleh sedikit mahasiswa

Selain itu juga akan dicari mata kuliah yang banyak diambil oleh mahasiswa, dan mata kuliah yang paling banyak diambil adalah mata kuliah dengan kode EA100. Kemudian akan dicari informasi juga terkait mata kuliah tersebut, apakah mata kuliah tersebut menjadi salah satu mata favorit mahasiswa dengan rata-rata kelulusan yang cukup baik. Berdasarkan diagram proporsi indeks mata kuliah kode EA100 pada gambar A.11, diperoleh banyak persebaran indeksnya cukup merata. Selain itu juga indeks nilai E menjadi yang paling dominan serta masih terdapat banyak mahasiswa yang belum lulus mata kuliah ini.

KODE_MK	Banyak diambil
EA100	741
IK402	625
EM604	603
IS220	600
IS240	542

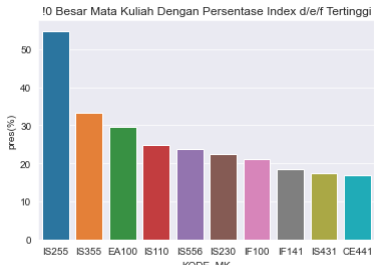
Gambar A. 11 Mata kuliah yang paling banyak diambil mahasiswa



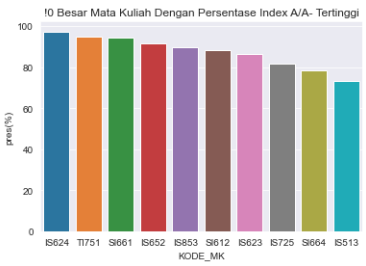
Gambar A. 12 Proporsi indeks yang diperoleh pada mata kuliah yang diambil banyak mahasiswa

6) Analisa Mata Kuliah Penentu

Tentunya dalam memilih mata kuliah, banyak diantara mahasiswa yang meninjau persebaran nilai dari periode-periode semester sebelumnya. Maka dari itu, akan dicari mata kuliah penentu, di mana mata kuliah ini memiliki suatu gambaran *output* indeks berdasarkan data tahun-tahun sebelumnya. Dalam hal ini akan dicari mata kuliah yang memiliki indeks bagus seperti A/A+ tertinggi dan juga mata indeks yang masih kurang seperti D/E/F. Tentunya dalam mencari persentase ini, akan digunakan batasan bahwa mata kuliah tersebut sudah diikuti oleh minimal 100 peserta kuliah. Hal ini bertujuan agar mencegah terjadinya bias dalam menarik kesimpulan.



Gambar A. 13 Mata kuliah dengan persentase indeks mahasiswa D/E/F terbanyak



Gambar A. 14 Mata kuliah dengan persentase indeks A/A- terbanyak

Gambar A.14 menunjukan 10 mata kuliah dengan persentase indeks D/E/F tertinggi. Jelas terlihat bahwa mata kuliah dengan kode IS255 memiliki persentase tertinggi, yakni 54,8% (108 data). Sedangkan gambar A.15 menunjukan 10 mata kuliah dengan indeks A/A- tertinggi. Mata kuliah dengan kode IS624 memiliki persentase tertinggi, yaitu 97,3% (254 data), akan tetapi perbedaan dengan mata kuliah tertinggi lainnya tidak jauh berbeda.

7) **Analisa hubungan antara mata kuliah prasyarat**

Mata kuliah prasyarat kami definisikan sebagai mata kuliah yang memiliki topik lanjutan pada mata kuliah lainnya. Dalam menentukan mata kuliah yang kategori ini, kami melakukan penyederhanaan bahwa mata kuliah prasyarat memiliki kode seri yang berurutan pada akhir nama mata kuliah. Diperoleh data bersih mata kuliah prasyarat beserta nilai-ratanya sebagai berikut.

KODE_MK	NAMA_MK	SKS	
IS645	Administrasi Database 1	3	89.571429
SI642	Administrasi Database 1	3	84.097561
IS747	Administrasi Database 2	3	92.339286
SI744	Administrasi Database 2	3	89.285714
UM121	Bahasa Inggris 1	2	69.896104
UM222	Bahasa Inggris 2	2	68.610390
UM321	Bahasa Inggris 3	2	72.831967
IS557	Big Data Analytics 1	3	71.153846
IS655	Big Data Analytics 2	3	65.200000
CE551	Jaringan Komputer Terapan 1	3	71.666667
SK533	Jaringan Komputer Terapan 1	3	88.666667
CE651	Jaringan Komputer Terapan 2	3	66.000000
SK632	Jaringan Komputer Terapan 2	3	96.000000
CE751	Jaringan Komputer Terapan 3	3	75.733333
SK733	Jaringan Komputer Terapan 3	3	87.000000
IF551	Keamanan Jaringan Lanjutan 1	3	64.125000
IF652	Keamanan Jaringan Lanjutan 2	3	71.625000
IF753	Keamanan Jaringan Lanjutan 3	3	62.833333
IF501	Keamanan Siber 1	3	81.000000
IF601	Keamanan Siber 2	3	75.000000
IF701	Keamanan Siber 3	3	75.000000
IF561	Pemrograman & Pengembangan Game 1	3	90.500000
IF662	Pemrograman & Pengembangan Game 2	3	84.000000
EA100	Pengantar Akuntansi 1	3	59.022942
EA201	Pengantar Akuntansi 2	3	67.008000
SI643	Pengembangan Aplikasi Database 1	3	84.666667
SI745	Pengembangan Aplikasi Database 2	3	81.111111
IF502	Perancangan dan Pengembangan Game 1	3	85.000000
IF602	Perancangan dan Pengembangan Game 2	3	90.250000
IF702	Perancangan dan Pengembangan Game 3	3	85.000000
IS580	Sistem & Aplikasi Perusahaan 1	3	79.318182
IS681	Sistem & Aplikasi Perusahaan 2	3	76.324786
IS782	Sistem & Aplikasi Perusahaan 3	3	75.277778
SI520	Sistem dan Aplikasi Perusahaan 1	3	82.960784
SI624	Sistem dan Aplikasi Perusahaan 2	3	79.383562
SI729	Sistem dan Aplikasi Perusahaan 3	3	72.164384
IS761	Skripsi 1	3	79.577586
IS862	Skripsi 2	3	81.084071
IS571	Tata Kelola Teknologi Informasi 1	3	77.747126
IS672	Tata Kelola Teknologi Informasi 2	3	72.905882
IS773	Tata Kelola Teknologi Informasi 3	3	81.673077
IS624	Topik Lanjut Sistem Informasi 1	3	90.298851
IS727	Topik Lanjut Sistem Informasi 2	3	80.744681
IS535	iOS Programming 1	3	75.600000
IS636	iOS Programming 2	3	79.400000
IS737	iOS Programming 3	3	70.600000

Gambar A. 15 Data bersih mata kuliah prasyarat

Kemudian, akan ditinjau nilai rata-rata yang diperoleh untuk setiap tingkatannya secara keseluruhan. Hipotesis kami, nilai rata-rata mahasiswa akan cenderung menurun seiring bertingkatnya tingkat prasyarat. Akanetapi berdasarkan hasil diperoleh bahwa nilai mahasiswa justru meningkat untuk setiap tingkatannya (gambar A. 16). Tentunya untuk menganalisa lebih lanjut mengenai fenomena ini, perlu meninjau beberapa informasi tambahan seperti jumlah mahasiswa pada setiap tingkat prasyarat serta perubahan rataan nilai mahasiswa untuk setiap mata kuliah.

```
Nilai rata-rata untuk mata kuliah prasyarat 1 adalah  71.73660714285714
Nilai rata-rata untuk mata kuliah prasyarat 2 adalah  73.1072625698324
Nilai rata-rata untuk mata kuliah prasyarat 3 adalah  74.24117647058823
[<matplotlib.lines.Line2D at 0x7f5dccb0c310>]
```

Gambar A. 16 Plot rataan nilai mahasiswa terhadap tingkat mata kulliah prasyarat

B. Modeling & Evaluation

Teknik dan Algoritma

Beberapa algoritma yang akan digunakan beserta penggunaannya:

- ❖ K-means untuk membentuk klaster matkul antara nilai evaluasi mata kuliah dan nilai ujian mata kuliah
- ❖ Decision Tree Classifier, Logistic Regression, Random Forest Classifier, KNN, Support Vector Machine untuk memprediksi tren kelulusan mahasiswa tepat waktu, dengan parameter keberhasilan akurasi model yang cukup tinggi.

1) K-means

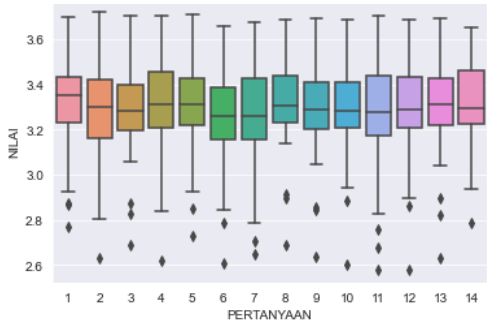
Data Train

Untuk mencari klaster diperlukan pasangan data dari rataan nilai evaluasi dosen untuk setiap kategori pertanyaan dan nilai mahasiswa pada semester yang sesuai. Jumlah data yang memenuhi kriteria ini hanya 50 data. Selanjutnya akan dilihat klasteringnya

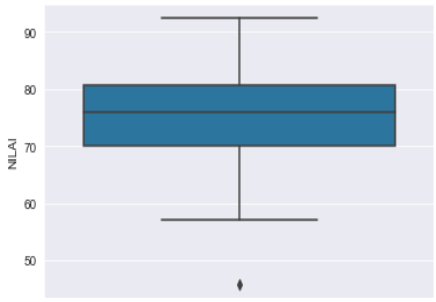
```
NILAI    50 non-null     float64
1        50 non-null     float64
2        50 non-null     float64
3        50 non-null     float64
4        50 non-null     float64
5        50 non-null     float64
6        50 non-null     float64
7        50 non-null     float64
8        50 non-null     float64
9        50 non-null     float64
10       50 non-null     float64
11       50 non-null     float64
12       50 non-null     float64
13       50 non-null     float64
14       50 non-null     float64
```

Gambar B. 1 Validasi data train

Scalling dan Pemilihan Parameter

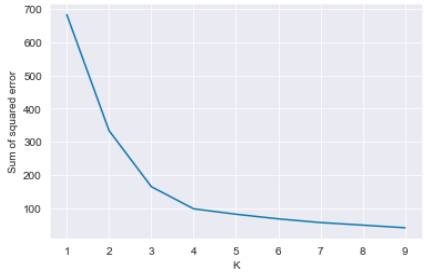


Gambar B. 2 Persebaran nilai evaluasi dosen untuk setiap pertanyaan



Gambar B. 3 Persebaran nilai mahasiswa

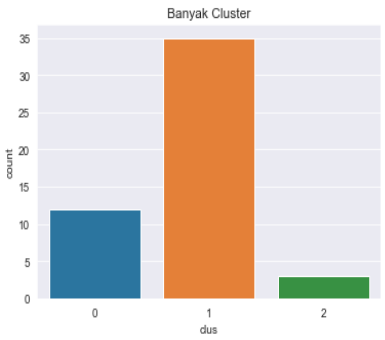
Berdasarkan gambar B.2 dan B.3 terlihat bahwa data masih memiliki banyak data pencilan/ *outlier*. Selain itu skala antara rataan nilai mahasiswa dan rataan nilai pertanyaan evaluasi mata kuliah memiliki rentang skala penilaian yang berbeda. Maka dari itu, kami menggunakan metode *robust scaler* untuk mentransformasikan data. Dapat diidentifikasi dengan *elbow method* pada gambar B.4 kami memilih pembagian 3 klaster (k = 3).



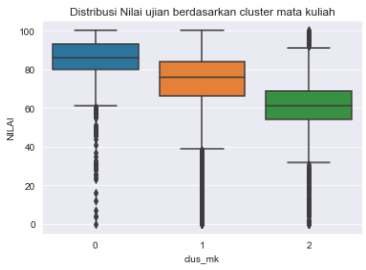
Gambar B. 4 Metode elbow pada k-means

Analisa Sifat klaster Mata Kuliah

Gambar B.5 di bawah menunjukan mata kuliah klaster 1 merupakan klaster yang paling banyak, sedangkan matkul klaster 2 adalah jenis klaster matkul yang paling sedikit. Beralih ke gambar B.6 mengenai persebaran nilai mahasiswa untuk setiap klaster, dapat dilihat bahwa rataan nilai klaster 0 paling tinggi dan klaster 2 memiliki rataan nilai ujian paling buruk. Kemudian jika meninjau persebaran nilai untuk setiap klaster (gambar B.7) terlihat bahwa klaster 1 memiliki persebaran nilai yang paling merata. Berbeda dengan klaster 0 cenderung memiliki persebaran nilai terkonsentrasi di atas dan matkul klaster 2 cenderung memiliki persebaran nilai terkonsentrasi di bawah.



Gambar B. 5 Jumlah data pada setiap klaster

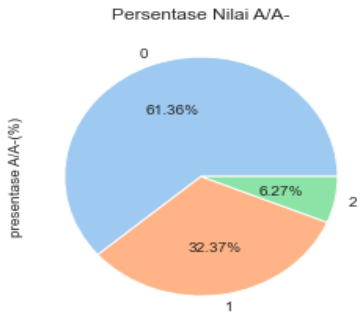


Gambar B. 6 Persebaran nilai mahasiswa per klaster

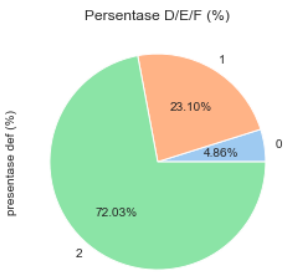
Setelah dilakukan analisis lanjut dengan meninjau persentase nilai penentu pada masing-masing kluster, terlihat pada gambar B.8 dan B.9 bahwa klaster 0 adalah matkul *output* yang cukup baik dengan persentase indeks bagus yang tinggi serta persentase indeks buruk yang sangat rendah. Dengan demikian dapat disimpulkan bahwa:

Gambar B. 7 Persebaran nilai untuk semua klaster

- Klaster 0 merupakan kategori mata kuliah yang memiliki peluang tinggi untuk mendapat indeks bagus. Hal ini dapat disebabkan oleh materi yang cukup mudah serta menyenangkan atau dosen baik dan menyenangkan
- Klaster 1 merupakan kategori mata kuliah yang memberikan kesan standar dalam perkuliahan
- Klaster 2 merupakan kategori mata kuliah yang sulit untuk mendapat indeks bagus. Hal ini dapat disebabkan oleh materi cenderung sulit serta menyenangkan atau dosen dan lingkungan belajar yang kurang menyenangkan. Argumen ini pun dapat dijelaskan oleh rataan nilai evaluasi mata kuliah yang rendah untuk setiap kategori pertanyaan (gambar B.10).



Gambar B. 8 Persentase indeks A/A- setiap klaster



Gambar B. 9 Persentase indeks D/E/F setiap klaster

	KODE_MK	NILAI	1	2	3	4	5	6	7	8	9	10	11	12	13	14
10	IS255	45.817259	2.770	2.630	2.690	2.620	2.730	2.610	2.650	2.6900	2.640	2.600	2.58	2.5800	2.6300	2.7900
11	IS302	62.344697	2.865	2.805	2.830	2.840	2.850	2.790	2.785	2.8950	2.845	2.885	2.76	2.8600	2.8200	2.9650
18	IS431	63.504178	2.875	2.830	2.875	2.865	2.925	2.845	2.880	2.9125	2.855	2.940	2.83	2.8975	2.8975	2.9375

Gambar B. 10 Sampel data nilai evaluasi untuk data di klaster 2

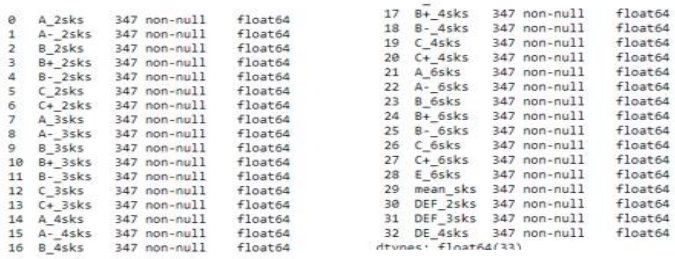
2) Decision Tree Classifier, Logistic Regression, Random Forest Classifier, KNN, Support Vector Machine

Data

Dalam memprediksi kelulusan tepat waktu mahasiswa, kami menggunakan sampel data nilai mahasiswa yang terdiri atas data mahasiswa angkatan 2010-2015 sebagai *data train.* Ini dikarenakan pada angkatan 2010-2015 sudah memiliki lulusan dan juga angkatan 2016-2018 sudah diasumsikan masih menempuh jenjang pendidikan. Dengan ini data mahasiswa yang akan diprediksi adalah data mahasiswa angkatan 2016-2018. Kami klasifikasikan mahasiswa yang lulus lebih cepat atau setara dengan 4 tahun adalah mahasiswa yang tepat waktu, sedangkan sisanya merupakan mahasiswa yang belum lulus.

Fitur

Faktor-faktor kelulusan kami pilih berdasarkan proporsi setiap indeks terhadap setiap beban SKS, kami namakan ini sebagai fitur (*features*). Selain itu juga kami melihat menduga bahwa jumlah SKS yang diambil setiap semester berpengaruh pada waktu kelulusan.



0	A_2sks	347 non-null	float64	17	B+_4sks	347 non-null	float64
1	A-_2sks	347 non-null	float64	18	B-_4sks	347 non-null	float64
2	B_2sks	347 non-null	float64	19	C_4sks	347 non-null	float64
3	B+_2sks	347 non-null	float64	20	C+_4sks	347 non-null	float64
4	B-_2sks	347 non-null	float64	21	A_6sks	347 non-null	float64
5	C_2sks	347 non-null	float64	22	A-_6sks	347 non-null	float64
6	C+_2sks	347 non-null	float64	23	B_6sks	347 non-null	float64
7	A_3sks	347 non-null	float64	24	B+_6sks	347 non-null	float64
8	A-_3sks	347 non-null	float64	25	B-_6sks	347 non-null	float64
9	B_3sks	347 non-null	float64	26	C_6sks	347 non-null	float64
10	B+_3sks	347 non-null	float64	27	C+_6sks	347 non-null	float64
11	B-_3sks	347 non-null	float64	28	E_6sks	347 non-null	float64
12	C_3sks	347 non-null	float64	29	mean_sks	347 non-null	float64
13	C+_3sks	347 non-null	float64	30	DEF_2sks	347 non-null	float64
14	A_4sks	347 non-null	float64	31	DEF_3sks	347 non-null	float64
15	A-_4sks	347 non-null	float64	32	DE_4sks	347 non-null	float64
16	B_4sks	347 non-null	float64				

dtypes: float64(33)

Gambar B. 11 Fitur yang digunakan dalam model decision tree

Output

Output dari model yang kami hasilkan memiliki 2 nilai, yakni 1 untuk mahasiswa yang lulus tepat waktu dan 0 untuk mahasiswa yang tidak lulus tepat waktu.

Train Test Split

Data model kami yang berupa data mahasiswa angkatan 2010-2015, kami bagi dengan metode split dengan ukuran *data test* sebesar 20%.

Scaling dan Resample

Agar skala antara fitur seragam, kami melakukan transformasi dengan *robust scaller* pada *data train* ataupun *test.* Kemudian karena terjadi ketidakseimbangan antara data yang lulus tepat waktu dan data yang lulus tidak tepat waktu pada data train, kami melakukan *resample* dengan teknik *smote tomek* pada data train.

Evaluasi Model

Berikut adalah hasil evaluasi dari model-model yang sudah dibuat.

Logistic Regression

Classification report

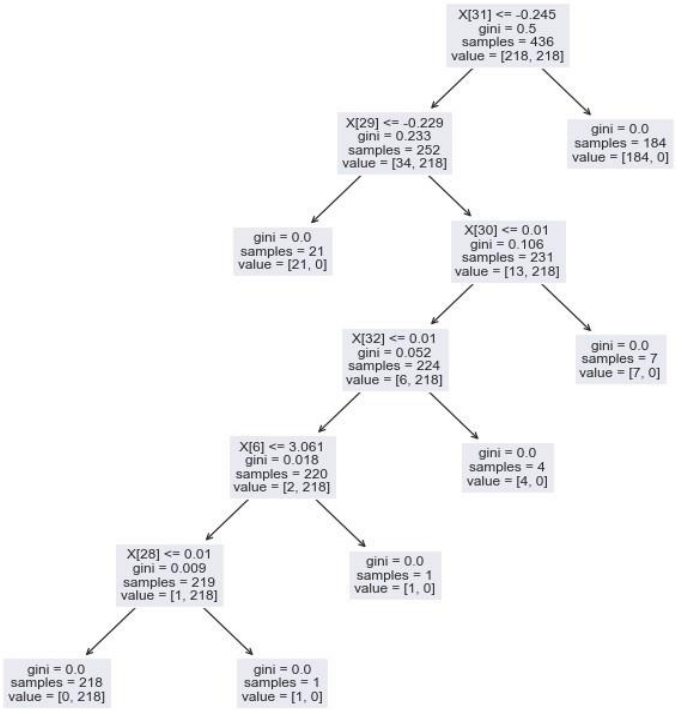
	precision	recall	f1-score	support
0	0.98	0.94	0.96	49
1	0.93	0.97	0.95	38
accuracy			0.95	87
macro avg	0.95	0.96	0.95	87
weighted avg	0.96	0.95	0.95	87

Confusion Matrix

```
[[46  3]
 [ 1 37]]
```

Decision Tree

Diperoleh hasil keputusan dari *decision tree* sebagai berikut



Gambar B. 12 Pohon keputusan dari model decision tree yang dibangun

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	49
1	1.00	1.00	1.00	38
accuracy			1.00	87
macro avg	1.00	1.00	1.00	87
weighted avg	1.00	1.00	1.00	87

Confusion Matrix

```
[[49  0]
 [ 0 38]]
```

RFC method

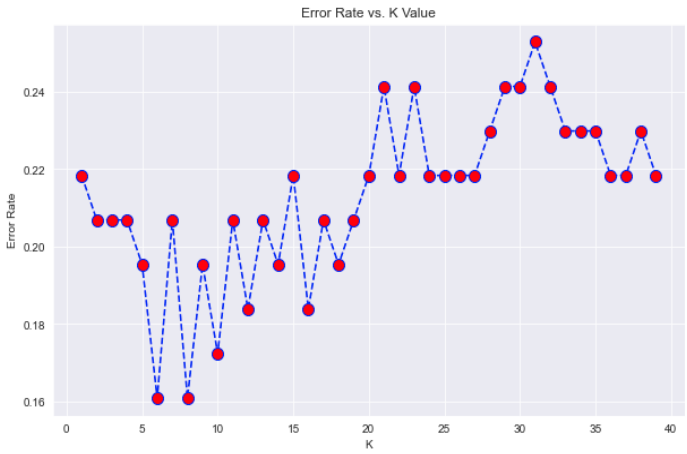
Classification Report

	precision	recall	f1-score	support
0	0.96	1.00	0.98	49
1	1.00	0.95	0.97	38
accuracy			0.98	87
macro avg	0.98	0.97	0.98	87
weighted avg	0.98	0.98	0.98	87

Confusion Matrix

```
[[49  0]
 [ 2 36]]
```

KNN



Gambar B. 13 Plot metode KNN

Berdasarkan gambar di atas kami memutuskan untuk memakai *n neighbors* = 25 karena nilai *n neighbors* yang cukup besar dan error rate yang lebih kecil dibanding nilai *n neighbors* lainnya

Classification Report

	precision	recall	f1-score	support
0	0.92	0.67	0.78	49
1	0.69	0.92	0.79	38
accuracy			0.78	87
macro avg	0.80	0.80	0.78	87
weighted avg	0.82	0.78	0.78	87

Confusion Matrix

```
[[33 16]
 [ 3 35]]
```

SVC

Classification Report

	precision	recall	f1-score	support
0	1.00	0.94	0.97	49
1	0.93	1.00	0.96	38
accuracy			0.97	87
macro avg	0.96	0.97	0.97	87
weighted avg	0.97	0.97	0.97	87

Confusion Matrix

```
[[46  3]
 [ 0 38]]
```

Analisa Validasi Model

Berdasarkan hasil evaluasi model yang kami berikan di atas, terlihat hampir setiap model memiliki hasil laporan klasifikasi dan confusion matrix yang cukup baik, kecuali KNN yang memiliki nilai *recall* yang cukup buruk pada kategori 0, yaitu 0,67. Selain itu nilai presisi pada model KKN cukup buruk pada kategori 1, yaitu sebesar 0,69. Oleh karena itu untuk analisis selanjutnya, kami tidak akan memakai model KNN. Kami memperoleh model *decision tree* sebagai model terbaik untuk data ini, karena berhasil memprediksi semua tebakan data test dengan benar dan memiliki nilai yang hampir sempurna untuk segala aspek metrik.

Feature Importances

Metode ini berfungsi untuk mengidentifikasi faktor-faktor mana saja yang merupakan akar penting dalam membuat suatu keputusan. Beberapa model yang digunakan untuk mengidentifikasi sebagai berikut:

Logistic Regression

	features_names	importances
29	mean_sks	2.017178
10	B+_3sks	0.971496
7	A_3sks	0.919766
8	A-_3sks	0.804567
16	B_4sks	0.702494
14	A_4sks	0.566329
12	C_3sks	0.562943
11	B-_3sks	0.294711
0	A_2sks	0.200115
9	B_3sks	0.183264
19	C_4sks	0.159319
3	B+_2sks	0.109179
1	A-_2sks	0.088238
21	A_6sks	0.016945
25	B-_6sks	0.012554
17	B+_4sks	0.006080
27	C+_6sks	0.001080
20	C+_4sks	-0.000067
26	C_6sks	-0.000161
15	A-_4sks	-0.000712
18	B-_4sks	-0.001636
22	A-_6sks	-0.008322
23	B_6sks	-0.013179
24	B+_6sks	-0.015102
28	E_6sks	-0.018319
32	DE_4sks	-0.064678
2	B_2sks	-0.150710
13	C+_3sks	-0.280144
6	C+_2sks	-0.284481
4	B-_2sks	-0.487158
5	C_2sks	-0.548013
31	DEF_3sks	-5.167258

Gambar B. 14 Tingkat kepetingan dari suatu fitur pada model regresi linear

Decision Tree

	features_names	importances
31	DEF_3sks	0.731225
29	mean_sks	0.156706
30	DEF_2sks	0.058736
32	DE_4sks	0.035234
28	E_6sks	0.009091
6	C+_2sks	0.009009

Gambar B. 15 Tingat kepentingan dari suatu fitur pada model regresi linear

RFC method

	features_names	importances
31	DEF_3sks	0.302003
29	mean_sks	0.131515
7	A_3sks	0.091731
0	A_2sks	0.089003
13	C+_3sks	0.078104
14	A_4sks	0.037431
12	C_3sks	0.033283
11	B-_3sks	0.030369
30	DEF_2sks	0.023813
5	C_2sks	0.022022
4	B-_2sks	0.021939
10	B+_3sks	0.021172
9	B_3sks	0.015707
8	A-_3sks	0.015504
2	B_2sks	0.015402
32	DE_4sks	0.012954
3	B+_2sks	0.008845
16	B_4sks	0.008594
1	A-_2sks	0.007420
19	C_4sks	0.006425
6	C+_2sks	0.006351
15	A-_4sks	0.006204
21	A_6sks	0.003259
24	B+_6sks	0.002963
17	B+_4sks	0.002069
23	B_6sks	0.002003
28	E_6sks	0.001773
22	A-_6sks	0.000880
18	B-_4sks	0.000825
25	B-_6sks	0.000409
20	C+_4sks	0.000031
26	C_6sks	0.000000
27	C+_6sks	0.000000

Gambar B. 16 Tingkat kepentingan dari fitur pada model RFC

SVC method

	features_names	importances
29	mean_sks	2.131230
9	B_3sks	0.392042
7	A_3sks	0.386067
12	C_3sks	0.304116
8	A-_3sks	0.235113
10	B+_3sks	0.217483
11	B-_3sks	0.193217
13	C+_3sks	0.154450
21	A_6sks	0.084552
0	A_2sks	0.049106
1	A-_2sks	0.041931
17	B+_4sks	0.040918
25	B-_6sks	0.038650
16	B_4sks	0.029174
14	A_4sks	0.020263
22	A-_6sks	0.019608
15	A-_4sks	0.016078
2	B_2sks	0.014048
3	B+_2sks	0.013677
24	B+_6sks	0.012523
19	C_4sks	0.008520
4	B-_2sks	0.007801
6	C+_2sks	0.002051
18	B-_4sks	0.000000
20	C+_4sks	0.000000
26	C_6sks	0.000000
27	C+_6sks	0.000000
5	C_2sks	-0.006306
23	B_6sks	-0.007769
28	E_6sks	-0.019608
32	DE_4sks	-0.080816
30	DEF_2sks	-0.178525
31	DEF_3sks	-3.970097

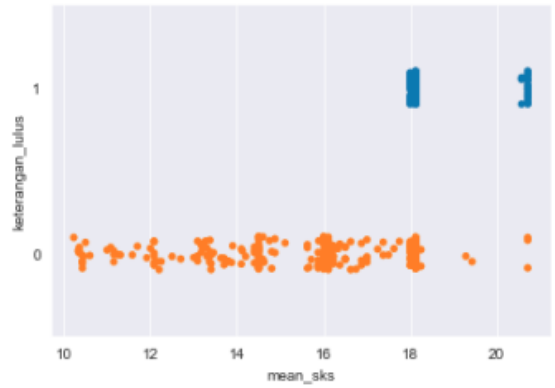
Gambar B. 17 Tingkat kepentingan dari fitur pada model SVC

Analisa *Features Importance*

Pada model RFC dan DCC tree terlihat bahwa persentase indeks D/E/F pada mata kuliah 3 SKS (DEF_3 sks) dan rata-rata SKS setiap semester (mean_sks) memiliki tingkat pengaruh yang penting terhadap model yang dibangun. Berbeda dengan model logistic regression dan SVC yang faktor kelulusannya ditentukan oleh nilai pada seluruh mata kuliah dengan 3 SKS. Sebagai contoh, indeks A/A- pada mata kuliah dengan beban 3 SKS akan menunjang kelulusan tepat waktu, sedangkan indeks D/E/F akan memicu kelulusan tidak tepat waktu. Selanjutnya akan dianalisa hubungan dari kedua faktor yang dominan pada setiap model, yaitu proporsi indeks D/E/F pada mata kuliah 3 SKS dan juga rata-rata pengambilan SKS per semester terhadap waktu kelulusan mahasiswa.

1) **Hubungan rata-rata pengambilan per semester terhadap waktu kelulusan mahasiswa**

Perhatikan *scatter plot* persebaran mahasiswa lulus tepat waktu dengan rata-rata SKS yang diambil setiap semester di bawah ini.

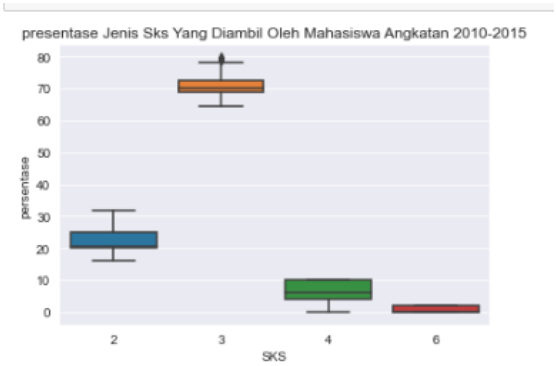


Gambar B. 18 Scatter plot SKS yang diambil mahasiswa terhadap status kelulusan

Berdasarkan gambar B.14 terlihat bahwa mahasiswa yang lulus tepat waktu cenderung mengambil antara 18-19 SKS dan 21-22 SKS. Hal ini mengimplikasikan bahwa pengambilan SKS yang tinggi atau diatas 18 tidak menjadi faktor kepastian bahwa mahasiswa akan lulus tepat waktu.

2) **Hubungan mata kuliah dengan beban 3 SKS terhadap waktu kelulusan mahasiswa**

Berdasarkan analisa mata kuliah pada bagian EDA, terlihat bahwa ketidakseimbangan antara beban mata kuliah, di mana didominasi oleh mata kuliah 3 SKS (gambar B.15). Ketidakseimbangan ini berimplikasi pada faktor bahwa mata kuliah 3 SKS akan lebih menentukan kelulusan mahasiswa berdasarkan metode yang dipakai.



Gambar B. 19 Persebaran jumlah dari setiap SKS yang diambil mahasiswa angkatan 2010-2015

Prediksi/ Implementasi Model

Berdasarkan model yang sudah dibuat dan divalidasi, kami akan memprediksi tren waktu kelulusan mahasiswa untuk mahasiswa angkatan 2016 -2018 dengan model tersebut.

Logistic regression

Gambar B. 20 Hasil prediksi kelulusan untuk angkatan 2016-2018 dengan metode regresi

Decision Tree

Gambar B. 21 Hasil prediksi kelulusan untuk angkatan 2016-2018 dengan metode decision tree

RF

Gambar B. 22 Hasil prediksi kelulusan untuk angkatan 2016-2018 dengan metode RFC

SVC

Gambar B. 23 Hasil prediksi kelulusan untuk angkatan 2016-2018 dengan metode SVC

Pembahasan Prediksi Model

Berdasarkan hasil prediksi dari model-model yang sudah dibangun, dapat dilihat bahwa persentase kelulusan tepat waktu angkatan 2016 -2018 cenderung lebih sedikit daripada angkatan 2010-2015. Kemudian berdasarkan model *decision tree* dan *RFC* dapat dilihat bawah persentase mahasiswa yang lulus tepat waktu akan terus meningkat dari angkatan 2016 -2018. Sebaliknya, berdasarkan model *logistic regression* dan *SVC* persentase mahasiswa yang lulus tepat waktu dari angkatan 2016 ke 2017 akan menurun tetapi dari angkatan 2017 ke 2018 terjadi peningkatan yang pesat pada angka persentase kelulusan tepat waktu.

CONCLUSION AND SUGGESTION

Berdasarkan model yang kami buat, kami menyimpulkan bahwa mata kuliah dengan beban 3 SKS dan rataan SKS yang diambil setiap semester merupakan dua hal yang paling berpengaruh terhadap waktu kelulusan mahasiswa. Oleh karena itu kami menyarankan mahasiswa mengambil 18-22 SKS untuk setiap semesternya dan lebih memfokuskan belajar pada matkul 3 SKS namun tidak melalupakan matkul dengan besar SKS lainnya. Kami juga menyarankan agar universitas melakukan pembenahan pada matkul klaster 2 yang nilai evaluasi dan nilai ujiannya sangat buruk.