

**SISTEM PREDIKSI IPK DAN LAMA STUDI MAHASISWA
INFORMATIKA UKDW**

Skripsi



oleh:

**RIZKY CHANDRA WIJAYA
71190443**

**PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA**

2023

**SISTEM PREDIKSI IPK DAN LAMA STUDI MAHASISWA
INFORMATIKA UKDW**

Skripsi



Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

RIZKY CHANDRA WIJAYA

71190443

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA

2023

PERNYATAAN KEASLIAN SKRIPSI

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

SISTEM PREDIKSI IPK DAN LAMA STUDI MAHASISWA INFORMATIKA UKDW

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 3 Juli 2023



RIZKY CHANDRA WIJAYA
71190443

DUTA WACANA

HALAMAN PERSETUJUAN

HALAMAN PERSETUJUAN

Judul Skripsi : SISTEM PREDIKSI IPK DAN LAMA STUDI
MAHASISWA INFORMATIKA UKDW
Nama Mahasiswa : RIZKY CHANDRA WIJAYA
N I M : 71190443
Matakuliah : Skripsi (Tugas Akhir)
Kode : TI0366
Semester : Genap
Tahun Akademik : 2022/2023

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 3 Juli 2023

Dosen Pembimbing I

Dosen Pembimbing II


Agata Filiana, S.Kom., M.Sc.


Gloria Virginia, S.Kom., MAI, Ph.D.

DUTA WACANA

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS
SECARA ONLINE
UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA**

Saya yang bertanda tangan di bawah ini:

NIM : 71190443
Nama : Rizky Chandra Wijaya
Prodi / Fakultas : Informatika
Judul Tugas Akhir : Sistem Prediksi IPK Dan Lama Studi Mahasiswa
Informatika UKDW

bersedia menyerahkan Tugas Akhir kepada Universitas melalui Perpustakaan untuk keperluan akademis dan memberikan **Hak Bebas Royalti Non Eksklusif** (*Non-exclusive Royalty-free Right*) serta bersedia Tugas Akhirnya dipublikasikan secara online dan dapat diakses secara lengkap (*full access*).

Dengan Hak Bebas Royalti Non eksklusif ini Perpustakaan Universitas Kristen Duta Wacana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk *database*, merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenar-benarnya.

Yogyakarta, 04 Juli 2023

Yang menyatakan,



(71190443 – Rizky Chandra Wijaya)

HALAMAN PENGESAHAN

HALAMAN PENGESAHAN

SISTEM PREDIKSI IPK DAN LAMA STUDI MAHASISWA INFORMATIKA UKDW

Oleh: RIZKY CHANDRA WIJAYA / 71190443

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 16 Juni 2023

Yogyakarta, 3 Juli 2023
Mengesahkan,

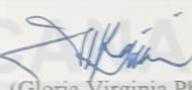
Dewan Penguji:

1. Agata Filiana, S.Kom., M.Sc.
2. Gloria Virginia, S.Kom., MAI, Ph.D.
3. Antonius Rachmat C., S.Kom., M.Cs.
4. Kristian Adi Nugraha, S.Kom., M.T.

Dekan

Ketua Program Studi


(Restyandito, S.Kom., MSIS., Ph.D.)


(Gloria Virginia, Ph.D.)

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS
SECARA ONLINE
UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA**

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS
SECARA ONLINE
UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA

Saya yang bertanda tangan di bawah ini:

NIM : 71190443
Nama : Rizky Chandra Wijaya
Prodi / Fakultas : Informatika
Judul Tugas Akhir : Sistem Prediksi IPK Dan Lama Studi Mahasiswa
Informatika UKDW

bersedia menyerahkan Tugas Akhir kepada Universitas melalui Perpustakaan untuk keperluan akademis dan memberikan **Hak Bebas Royalti Non Eksklusif** (*Non-exclusive Royalty-free Right*) serta bersedia Tugas Akhirnya dipublikasikan secara online dan dapat diakses secara lengkap (*full access*).

Dengan Hak Bebas Royalti Noneklusif ini Perpustakaan Universitas Kristen Duta Wacana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk *database*, merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenar-benarnya.

Yogyakarta, 09 Juli 2023

Yang menyatakan,



(71190443 – Rizky Chandra Wijaya)

DAFTAR ISI

PERNYATAAN KEASLIAN SKRIPSI.....	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PENGESAHAN.....	v
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS SECARA ONLINE.....	vi
UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	x
DAFTAR GAMBAR	xi
INTISARI.....	xiii
ABSTRACT.....	xiv
BAB I.....	1
PENDAHULUAN	1
1.1. Latar Belakang Masalah.....	1
1.2. Perumusan Masalah.....	2
1.3. Batasan Masalah.....	3
1.4. Tujuan Penelitian.....	4
1.5. Manfaat Penelitian.....	4
1.6. Sistematika Penulisan.....	4
BAB II.....	6
TINJAUAN PUSTAKA DAN DASAR TEORI	6
2.1 Tinjauan Pustaka	6
2.2 Landasan Teori	8
2.2.1 Kurikulum	8
2.2.2 Profil Lulusan Prodi Informatika	8
2.2.3 Beban Studi dan Lama Studi.....	8
2.2.4 Yudisium.....	9

2.2.5	Struktur Mata Kuliah.....	9
2.2.6	<i>Data Warehouse</i>	14
2.2.7	<i>Data Mining</i>	14
2.2.8	Regresi.....	16
2.2.9	<i>Super Vector Regression (SVR)</i>	18
2.2.10	Metode Evaluasi.....	18
BAB III		22
METODOLOGI PENELITIAN.....		22
3.1	Kebutuhan Sistem.....	22
3.1.1	Kebutuhan Non Fungsional.....	22
3.1.2	Kebutuhan Fungsional	22
3.2	Metode Penelitian.....	23
3.2.1	Persiapan Awal.....	24
3.2.2	Prapemrosesan (<i>Preprocessing</i>).....	25
3.2.3	Pelatihan Model Pembelajaran Mesin.....	25
3.2.4	Pengujian (<i>Testing</i>)	26
3.2.5	Alur Kerja Sistem.....	26
3.2.6	Rancangan Antar Muka Sistem.....	34
3.2.7	Evaluasi	36
BAB IV		37
IMPLEMENTASI DAN PEMBAHASAN.....		37
4.1	Implementasi Sistem	37
4.1.1	Data Latih.....	37
4.1.3	Implementasi Antar Muka.....	40
4.1.4	Pemodelan	43
4.2	Pengujian dan Analisis	44
4.2.1	Komparasi Model Regresi IPK dan Lama Studi.....	44

4.2.2	Proses Seleksi Fitur (OLS).....	48
4.2.5	Pengaruh metode SMOTE untuk melakukan <i>oversampling</i> profil .	57
4.3	Pembuatan Model Regresi Tiap Profil Lulusan	58
4.4	Pembahasan	60
BAB V.....		62
KESIMPULAN DAN SARAN.....		62
5.1	Kesimpulan.....	62
5.2	Saran.....	63
DAFTAR PUSTAKA		64
LAMPIRAN A		66
KODE SUMBER PROGRAM		66
LAMPIRAN B		86
KARTU KONSULTASI DOSEN 1.....		86
LAMPIRAN C		87
KARTU KONSULTASI DOSEN 2.....		87
LAMPIRAN D.....		88
LAMPIRAN LAIN-LAIN		88



DAFTAR TABEL

Tabel 2. 1 Daftar hubungan antara mata kuliah dan profil lulusan.....	10
Tabel 3. 1 Tabel keterangan model.....	27
Tabel 4. 1 Hasil k-fold cross validation.....	44
Tabel 4. 2 Hasil k-fold cross validation (lanjutan).....	45
Tabel 4. 3. Keterangan variabel independen.....	48
Tabel 4. 4. Keterangan variabel independent (lanjutan).....	49
Tabel 4. 5. Keterangan variabel independent (lanjutan).....	50
Tabel 4. 6 Nilai VIF.....	56
Tabel 4. 7 Hasil percobaan model tiap profil lulusan.....	58
Tabel L. 1 Daftar mata kuliah wajib.....	88
Tabel L. 2 Daftar mata kuliah pilihan.....	89



DAFTAR GAMBAR

Gambar 2.1 Diagram alur CRISP-DM.....	15
Gambar 2.2 Contoh dari plot model regresi linier	17
Gambar 3. 1 Skema Relasi Tabel.....	24
Gambar 3.2 Diagram Alir Tahap Prapemrosesan	25
Gambar 3.3. Diagram alir proses pelatihan model.....	26
Gambar 3. 4 Diagram alir proses prediksi dalam sistem	34
Gambar 3. 5 Rancangan awal halaman menu prediksi sistem.....	35
Gambar 3. 6 Rancangan awal halaman hasil prediksi.....	35
Gambar 4. 1 Distribusi IPK data latih.....	37
Gambar 4. 2 Distribusi lama studi data latih.....	38
Gambar 4. 3 Distribusi tipe SMA asal pada data latih.....	39
Gambar 4. 4 Distribusi jurusan SMA asal pada data latih	39
Gambar 4. 5 Distribusi Provinsi SMA asal mahasiswa data latih.....	40
Gambar 4. 6 Distribusi data profil lulusan mahasiswa data latih.....	40
Gambar 4. 7 Halaman <i>login</i>	41
Gambar 4. 8 Tampilan <i>dashboard</i> sistem	42
Gambar 4. 9 Tampilan menu awal sistem prediksi	42
Gambar 4. 10 Hasil prediksi IPK dan lama studi.....	43
Gambar 4. 11 <i>Pseudocode</i> sistem prediksi	43
Gambar 4. 12 Visualisasi hasil analisis R^2 <i>cross validation</i>	45
Gambar 4. 13 Visualisasi hasil analisis MAE model IPK	46
Gambar 4. 14 Visualisasi hasil analisis R^2 model lama studi	47
Gambar 4. 15 Visualisasi hasil analisis MAE model lama studi	47
Gambar 4. 16 Regresi OLS IPK.....	51
Gambar 4. 17 Hasil OLS IPK setelah seleksi fitur.....	51
Gambar 4. 18 Regresi OLS lama studi	52
Gambar 4. 19 Hasil OLS lama studi setelah seleksi fitur	53
Gambar 4. 20 Visualisasi hasil koefisien model regresi IPK.....	54
Gambar 4. 21 Visualisasi hasil koefisien model regresi lama studi.....	54

Gambar 4. 22 Distribusi normal residu model regresi IPK.....	55
Gambar 4. 23 Distribusi normal residu model prediksi lama studi.....	55
Gambar 4. 24 Visualisasi hasil analisis model IPK SMOTE.....	57
Gambar 4. 25 Visualisasi hasil analisis model lama studi SMOTE.....	58
Gambar Lampiran 1 Mata kuliah konversi per Kurikulum.....	93



INTISARI

SISTEM PREDIKSI IPK DAN LAMA STUDI MAHASISWA INFORMATIKA UKDW

Oleh

RIZKY CHANDRA WIJAYA

71190443

Dalam mengukur kualitas atau mutu dari seorang lulusan mahasiswa dari suatu perguruan tinggi terdapat dua indikator yang sering digunakan sebagai tolak ukur yaitu Indeks Prestasi Kumulatif (IPK) dan lama studi mahasiswa. Berangkat dari pentingnya faktor tersebut, Program Studi Informatika UKDW mempunyai *data warehouse* yang menyimpan data hasil studi lulusan mahasiswa informatika dari tahun 2013 hingga 2018. Dengan adanya data tersebut maka penelitian ini dibuat agar dapat memprediksi IPK dan lama studi mahasiswa yang sedang menempuh studi dengan memanfaatkan data lulusan mahasiswa Informatika UKDW.

Penelitian ini membandingkan 2 model regresi untuk memprediksi IPK dan lama studi yaitu regresi linier multi dan *Super Vector Machine* (SVR). Untuk mengevaluasi model regresi penulis menggunakan uji nilai R^2 untuk melihat seberapa baik model regresi dapat memprediksi variabel dependen dan uji nilai MAE untuk melihat tingkat eror yang dihasilkan model regresi.

Berdasarkan hasil analisis model regresi yang paling baik untuk memprediksi IPK dengan menggunakan SVR, sedangkan untuk memprediksi lama studi berdasarkan analisis model yang paling baik adalah regresi linier multi. Hasil dari model regresi tersebut kemudian di implementasikan ke dalam sistem berbasis web dengan menggunakan *framework* laravel.

Kata-kata kunci : IPK, lama studi, regresi linier multi, SVR, R^2 , MAE

ABSTRACT

PREDICTION SYSTEM FOR GPA AND STUDY DURATION OF DUTA WACANA CHRISTIAN UNIVERSITY INFORMATICS STUDENTS

By

RIZKY CHANDRA WIJAYA

71190443

In measuring the quality or quality of a graduate student from a tertiary institution, two indicators are often used as benchmarks, namely the Grade Point Average (GPA) and the student's study duration. Departing from the importance of these factors, the UKDW Informatics Study Program has a data warehouse that stores data on the study results of informatics student graduates from 2013 to 2018. With this data, this research was made in order to be able to predict the GPA and length of study of students who are currently studying by utilizing data UKDW Informatics graduate students.

This study compares 2 regression models to predict GPA and length of study, namely multiple linear regression and Super Vector Machine (SVR). To evaluate the regression model, the authors used the R^2 value test to see how well the regression model could predict the dependent variable and the MAE value test to see the error level produced by the regression model.

Based on the results of the analysis the best regression model to predict GPA using SVR, while to predict the length of study based on the best model analysis is multiple linear regression. The results of the regression model are then implemented into a web-based system using the Laravel framework.

Keywords: GPA, study duration, multiple linear regression, SVR, R^2 , MAE

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Dalam mengukur kualitas atau mutu dari seorang lulusan mahasiswa dari suatu perguruan tinggi terdapat dua indikator yang sering digunakan sebagai tolak ukur yaitu Indeks Prestasi Kumulatif (IPK) dan lama studi mahasiswa tersebut. Kedua indikator tersebut kerap dikaitkan dengan kriteria orang yang mengikuti seleksi untuk menjadi karyawan di dunia kerja profesional. Banyak perusahaan atau institusi khususnya yang berada di Indonesia mewajibkan calon pelamar kerja untuk memiliki IPK minimal 3,00 bahkan tidak sedikit institusi baik swasta ataupun pemerintahan yang mewajibkan calon pelamar kerja memiliki IPK minimal 3,50 dan masa studi tidak melebihi 5 tahun.

Indeks Prestasi Kumulatif (IPK) dan lama studi mahasiswa yang bersangkutan juga menjadi faktor penting dalam memperoleh gelar sarjana. Pada Universitas Kristen Duta Wacana khususnya Program Studi Informatika, pada kurikulum terbaru yang diterapkan yaitu Kurikulum 2021 terdapat persyaratan yang wajib dipenuhi mahasiswa saat hendak mendaftarkan diri untuk yudisium yaitu adalah $IPK \geq 2,00$ dan lama studi ≤ 14 Semester. Selain sebagai persyaratan untuk yudisium, IPK dan lama studi mahasiswa dapat menjadi salah satu indikator keberhasilan seorang mahasiswa dalam mengenyam pendidikan yaitu dengan mendapatkan gelar *cumlaude* dengan syarat memperoleh $IPK \geq 3,50$ dan lama studi ≤ 8 semester.

Dalam rangka memberi masukan terkait akademik mahasiswa berdasarkan histori studi mahasiswa secara manual, penelitian ini memanfaatkan *data warehouse* yang menyimpan informasi terkait identitas mahasiswa, hasil pembelajaran mahasiswa berupa nilai-nilai mata kuliah yang diambil per semester, data yudisium, data kurikulum, dan data lainnya yang berkaitan dengan proses pembelajaran yang dilakukan di Prodi Informatika UKDW. Dengan tersedianya data tersebut penulis melakukan pengujian korelasi antara data proses pembelajaran mahasiswa dengan IPK dan lama studi mahasiswa yang

bersangkutan. Maka dari itu penulis ingin mencari tahu seberapa tinggi tingkat korelasi antara data pembelajaran mahasiswa beserta variabel atau faktor lainnya dengan IPK dan lama studi.

Berangkat dari permasalahan tersebut, penelitian ini bertujuan untuk dapat memberikan informasi terkait prediksi IPK dan lama studi mahasiswa yang sedang mengenyam pendidikan di Prodi Informatika UKDW. Informasi ini diharapkan dapat membantu mahasiswa yang bersangkutan agar dapat mengetahui perkiraan IPK yang akan diperoleh di akhir studinya beserta perkiraan lama waktu yang diperlukan untuk menyelesaikan studinya. Selain berguna untuk mahasiswa penelitian ini juga diharapkan dapat membantu dosen wali dan juga pihak-pihak yang berperan penting dalam proses pendidikan mahasiswa untuk mengetahui bagaimana proses studi mahasiswa yang bersangkutan.

Dalam proses prediksi IPK dan lama studi penulis akan menggunakan pembelajaran mesin dengan membandingkan model dengan menggunakan metode *Multiple Linear Regression* (regresi linear multi) dan *Super Vector Regression* (SVR). Kedua metode ini sudah banyak digunakan dalam penelitian dan dapat menyelesaikan permasalahan prediksi hasil angka kontinu dengan fitur yang berjumlah lebih dari 1, metode ini juga cocok untuk melakukan pelatihan dan testing terhadap data kategorikal maupun kontinu. Untuk mendapatkan model linier yang paling baik (*best-fit*) maka penulis menggunakan metode analisis model dengan *Ordinary Linear Square* (OLS) untuk melakukan estimasi koefisien, menentukan korelasi fitur dengan hasil prediksinya dari model linier yang didapat, dan menguji distribusi (*variance*) data.

1.2. Perumusan Masalah

Berikut perumusan masalah dalam penelitian ini:

1. Parameter apa saja yang penting dan seberapa tinggi tingkat korelasinya dalam memprediksi IPK dan lama studi mahasiswa Prodi Informatika UKDW?

2. Apakah model pembelajaran mesin dengan menggunakan metode regresi linier multi lebih baik dibandingkan dengan SVR dalam memprediksi IPK dan lama studi?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini antara lain:

1. Seluruh data yang digunakan sebagai data latih berasal dari data lulusan mahasiswa angkatan 2013 hingga 2018 yang sudah dinyatakan lulus yudisium.
2. Dikarenakan keterbatasan data terkait proses pembelajaran mahasiswa yang terdapat pada *data warehouse* Prodi Informatika, variabel independen atau fitur yang digunakan dalam penelitian untuk memprediksi IPK dan lama studi adalah profil lulusan mahasiswa, provinsi SMA asal mahasiswa, tipe SMA asal mahasiswa, jurusan SMA asal mahasiswa, mata kuliah wajib dan beberapa nilai mata kuliah pilihan wajib profil yang dapat diambil mahasiswa selama masa studi hingga semester 4 untuk masing-masing profil lulusan yaitu: *User Interface Designer / User Experience (UI/UX)*, *Intelligent System Developer (ISD)*, *Database Administrator (DA)*, dan *Network Service Supervisor (NSS)*. Penulis meyakini bahwa variabel di atas memiliki korelasi terhadap IPK dan lama studi, sehingga penulis ingin meneliti lebih dalam terkait korelasi antara variabel independen tersebut dengan IPK dan lama studi.
3. Acuan kurikulum yang digunakan untuk sistem adalah Kurikulum 2021 sebagai kurikulum terbaru yang digunakan hingga saat penelitian dilakukan. Dikarenakan terdapat perbedaan dalam bobot SKS, struktur mata kuliah, dan nama mata kuliah, maka diperlukan mata kuliah dari Kurikulum 2011 revisi 2013, Kurikulum 2017, dan Kurikulum 2017 revisi 2019 menjadi Kurikulum 2021. Penyetaraan tersebut dilakukan agar data hasil pembelajaran yang digunakan dalam pelatihan dan prediksi sistem relevan terhadap acuan Kurikulum 2021.

4. Jumlah SKS tempuh mahasiswa yang digunakan sebagai data untuk diprediksi sistem minimal berjumlah 50 SKS dan sudah menempuh lama studi minimal 4 semester. Jika persyaratan tersebut tidak terpenuhi maka besar kemungkinan hasil prediksi sistem akan tidak akurat.
5. Sistem yang dikembangkan pada penelitian ini adalah sistem berbasis web.

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk mengembangkan sistem prediksi IPK dan lama studi mahasiswa Prodi Informatika UKDW dengan membandingkan hasil dari metode regresi linier multi dengan SVR untuk melihat hubungan antara variabel independen terhadap prediksi IPK dan lama studi mahasiswa Prodi Informatika UKDW.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Memberikan informasi kepada mahasiswa yang bersangkutan agar dapat membuat rencana studi yang lebih baik dengan adanya hasil prediksi IPK dan lama studi yang diperoleh dari sistem.
2. Mempermudah dosen wali Prodi Informatika UKDW, Kepala Prodi Informatika UKDW, Dekanat, dan pihak-pihak lain yang berperan dalam proses pembelajaran pada Prodi Informatika UKDW dalam memantau hasil pembelajaran mahasiswa.

1.6. Sistematika Penulisan

Sistematika Penulisan laporan ini dibagi ke dalam 5 bagian yaitu: Bab 1 adalah bagian Pendahuluan yang menjelaskan tentang latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, dan manfaat penelitian. Bab 2 adalah bagian Tinjauan Pustaka dan Landasan Teori yang menjelaskan tentang penelitian terkait dan berbagai landasan teori yang digunakan penulis dalam melakukan penulisan dan penelitian ini. Bab 3 adalah bagian Metodologi

Penelitian yang menjelaskan proses yang dilakukan penulis dalam membuat sistem prediksi untuk IPK dan Lama studi. Bab 4 adalah bagian Implementasi dan Pembahasan yang menjelaskan terkait hasil penelitian yang sudah dilakukan penulis dan bagaimana penulis mengimplementasikan hasil pembelajaran mesin ke dalam sistem berbasis web dengan menggunakan laravel. Bab 5 adalah bagian Kesimpulan dan Saran yang berisi tentang kesimpulan dari keseluruhan penelitian yang sudah dilakukan dan saran terkait penulis lain yang akan melakukan pengembangan dari penelitian ini.



BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian terkait prediksi IPK dan lama studi sudah banyak dilakukan dengan berbagai macam metode dan macam-macam data atau fitur yang digunakan sebagai data latih. Metode untuk pembelajaran mesin yang paling banyak ditemukan untuk memprediksi IPK dan lama studi adalah klasifikasi (*Naive Bayes classifier*, *Decision Tree*, *Random Forest*, dan *SVM*) dan Regresi (regresi linier sederhana, regresi linier multi, dan *SVR*).

Amalia (2020) menerapkan *data mining* dengan metode *naive bayes* untuk memprediksi tingkat kelulusan siswa SMA. Data yang digunakan pada penelitian ini diambil dari data Daftar Kolektif Hasil Ujian Sekolah Berstandar Nasional (DKHUSBN) tahun ajaran 2018/2019 di provinsi Kalimantan Tengah. Fitur yang digunakan untuk memprediksi tingkat kelulusan adalah nilai bahasa Indonesia, bahasa inggris, matematika, dan IPA. Pada tahap evaluasi digunakan nilai akurasi dan *recall* untuk melihat ketepatan sistem dalam memprediksi kelulusan siswa. Berdasarkan hasil evaluasi sistem tersebut dapat memprediksi tingkat kelulusan dengan baik dengan nilai akurasi 82%.

Kurniatullah dan Pramudi (2020) membuat sistem prediksi masa studi mahasiswa dengan menggunakan regresi linier multi berbasis web dengan menggunakan bahas pemrograman PHP. Dalam penelitiannya digunakan 9 variabel untuk melakukan prediksi yaitu jumlah SKS semester 1 hingga 4, Indeks Prestasi Semester 1 hingga 4, dan jumlah mata kuliah yang mengulang. Data yang digunakan pada penelitian ini didapat dari *data warehouse* Fakultas Ilmu komputer Universitas Dian Nuswantoro dengan mengambil data sejumlah 1753 *records* dari data mahasiswa angkatan 2008 hingga 2012, kemudian dilakukan prapemrosesan data dengan hasil akhir didapatkan sejumlah 1000 *records*. Pada penelitian ini menggunakan pembelajaran mesin untuk mendapatkan sejumlah 10 rumus *slope* untuk regresi linier multi yang kemudian dilakukan perkalian matriks

untuk nilai *slope* dengan koefisien regresi, kemudian hasil perkalian matriks tersebut dikalikan dengan matriks identitas dan selanjutnya dilakukan perkalian *inverse* matriks dengan kolom vektor untuk mendapatkan rumus prediksi lama masa studi. Evaluasi yang digunakan dalam penelitian ini menggunakan nilai MAE dan RMSE yang dihasilkan model, tujuan dari evaluasi ini adalah untuk menemukan model dengan nilai MAE dan RMSE yang paling kecil.

Uyanik dan Guler (2013), diteliti hubungan antara 5 variabel independen yaitu nilai akhir dan evaluasi dari mata kuliah, psikologis akademik, kurikulum yang berlaku, dosen pengajar, dan metode mengajar dengan variabel dependen nilai KPPS. Langkah awal yang dilakukan adalah melakukan uji statistik ANOVA untuk melihat seberapa signifikan variabel independen dengan membandingkan *mean* dengan *variance* yang terdapat pada distribusi data. Kemudian dilakukan uji *variance increase factors* (VIFs) dengan menggunakan standar nilai untuk setiap variabel independen jika memiliki nilai VIF di atas 10 maka dapat disimpulkan bahwa ada korelasi antara variabel independen dengan melihat pertimbangan koefisien korelasi dari uji VIFs. Langkah selanjutnya adalah melihat hasil dari *condition indexes* (CI), jika nilai CI variabel independen lebih dari 30 maka dapat dikatakan bahwa variabel tersebut memiliki korelasi dengan variabel independen lainnya. Langkah terakhir yang dilakukan dalam penelitian ini adalah melakukan evaluasi terhadap model regresi dengan melihat nilai koefisien dan *p-value*. Semakin besar nilai koefisien yang dihasilkan model maka dapat disimpulkan bahwa variabel tersebut memiliki korelasi semakin besar dengan variabel dependen,

Dewi dan Widiastuti (2020) memprediksi GPA mahasiswa dengan menggunakan metode *Super Vector Regression* (SVR). Variabel yang digunakan dalam penelitian ini adalah nilai 8 mata kuliah dasar yang diambil pada semester 1 dan nilai indeks prestasi semester 1. Peneliti menganggap bahwa semester 1 merupakan semester yang penting karena pada semester itu diambil mata kuliah dasar seperti kalkulus, algoritma pemrograman, dan *programming* yang harus ditempuh seluruh mahasiswa. Pada tahap prapemrosesan, dilakukan normalisasi dengan mengubah nilai variabel independen menggunakan *StandardScaler*,

setelah itu baru dilakukan pelatihan dengan SVR dengan membandingkan *kernel* RBF dan *polynomial*. Kesimpulan dari penelitian ini adalah standarisasi data berpengaruh positif dalam mengurangi tingkat eror MAE dan RMSE dan penggunaan *kernel* RBF lebih baik dibandingkan dengan *kernel polynomial*.

2.2 Landasan Teori

2.2.1 Kurikulum

Kurikulum adalah seperangkat rencana dan pengaturan mengenai capaian pembelajaran lulusan, bahan kajian, proses, dan penilaian yang digunakan sebagai pedoman penyelenggaraan program studi (Tim Penyusun Kurikulum 2021 Prodi Informatika UKDW, 2021). Prodi Informatika UKDW secara rutin melakukan penyusunan dan pembaharuan kurikulum setiap 5 tahun sekali untuk perubahan besar, sedangkan untuk perubahan kecil dan peninjauan pelaksanaan kurikulum dilakukan setiap 2 tahun sekali. Prodi Informatika UKDW pada saat ini menerapkan Kurikulum 2021 sebagai acuan dalam penyelenggaraan studi.

2.2.2 Profil Lulusan Prodi Informatika

Terkait dengan profil profesional mandiri atau bidang kerja lulusan, dalam Kurikulum 2021, Prodi Informatika UKDW menetapkan 4 profil profesional yaitu: *User Interface/User Experience Designer (UI/UX)*, *Intelligent System Developer (ISD)*, *Database Administrator (DA)*, dan *Network Service Supervisor (NSS)*.

2.2.3 Beban Studi dan Lama Studi

Dalam proses pendidikan yang dilaksanakan Prodi Informatika UKDW, khususnya dengan diberlakukannya kurikulum 2021 untuk mendapatkan gelar sarjana (S-1) maka mahasiswa sekurang-kurangnya harus menempuh 144 SKS. Lama masa studi maksimal adalah 7 tahun akademik atau 14 semester, tidak termasuk cuti studi (Menristekdikti, 2015).

2.2.4 Yudisium

Yudisium adalah proses sidang dewan dosen program studi untuk memutuskan kelulusan mahasiswa atas seluruh beban studi yang telah ditempuhnya (Tim Penyusun Kurikulum 2021 Prodi Informatika UKDW, 2021). Berikut adalah beberapa aturan yang menentukan kelulusan yudisium:

1. Telah menyelesaikan seluruh proses Skripsi
2. Telah menempuh minimal 144 SKS dengan aturan pembagian SKS tempuh adalah 12 SKS mata kuliah wajib humaniora, 85 SKS mata kuliah wajib Prodi Informatika, 18 SKS mata kuliah pilihan wajib profil, \pm 23 SKS mata kuliah pilihan bebas prodi, dan atau maksimal 6 SKS mata kuliah pilihan bebas non Prodi.
3. Memenuhi poin keaktifan yang disyaratkan bidang kemahasiswaan
4. Bebas dari sanksi akademik dan administratif

2.2.5 Struktur Mata Kuliah

Berdasarkan Kurikulum 2021 Prodi Informatika UKDW menerapkan aturan terkait pembagian kelompok mata kuliah menjadi 2 kelompok besar yaitu Mata Kuliah Wajib dan Mata Kuliah Pilihan.

2.2.5.1 Mata Kuliah Wajib

Dalam proses studi, setiap mahasiswa wajib untuk menempuh seluruh mata kuliah wajib yang disediakan oleh Prodi Informatika UKDW. Prodi Informatika membagi mata kuliah wajib ke dalam 2 kategori yaitu mata kuliah wajib DIKTI (Humaniora) dan mata kuliah wajib keilmuan (Prodi). Pada Tabel L.1 menunjukkan daftar mata kuliah wajib dan prasyaratnya.

2.2.5.2 Mata Kuliah Pilihan

Berdasarkan sifat mengikat dari mata kuliah wajib yang mengharuskan setiap mahasiswa menempuh seluruh mata kuliah wajib tersebut, mata kuliah pilihan bersifat opsional sesuai dengan minat dan bakat mahasiswa yang bersangkutan. Untuk memfasilitasi minat dan bakat mahasiswa Prodi Informatika membagi mata kuliah pilihan ke dalam 3 kategori yaitu:

1. Mata kuliah pilihan wajib profil

Mata kuliah ini dibentuk untuk memperkuat kemampuan mahasiswa sesuai dengan pilihan profil profesional lulusan. Sesuai dengan ketentuan Kurikulum 2021, setiap mahasiswa wajib menempuh minimal 6 mata kuliah pilihan wajib profil atau setara dengan 18 SKS mata kuliah pilihan wajib profil sesuai dengan pilihan profil profesional yang dipilih mahasiswa.

2. Mata kuliah pilihan bebas Prodi

Kelompok mata kuliah ini dapat dipilih oleh mahasiswa secara bebas tanpa melihat keterkaitannya dengan profil profesional lulusan. Dalam Kurikulum 2021, setiap mahasiswa dapat mengambil mata kuliah pilihan bebas Prodi kurang lebih sebanyak 8 mata kuliah atau setara dengan 23 SKS.

3. Mata kuliah pilihan bebas non bidang

Kelompok mata kuliah ini adalah mata kuliah yang disediakan atau diselenggarakan oleh unit atau Prodi di luar Prodi Informatika. Berdasarkan aturan dari Kurikulum 2021, mata kuliah pilihan bebas non bidang yang akan diakui maksimal sejumlah 2 mata kuliah atau setara dengan 6 SKS.

2.2.5.3 Hubungan Antara Mata Kuliah dan Profil Lulusan

Mata kuliah yang ada merupakan satu kesatuan yang mendukung dalam rangka mencapai profil profesional mandiri (Tim Penyusun Kurikulum 2021 Prodi Informatika UKDW, 2021). Hubungan antara mata kuliah dengan profil lulusan digambarkan oleh Tabel 2.1.

Tabel 2. 1 Daftar hubungan antara mata kuliah dan profil lulusan

No	Nama Mata Kuliah	UI/UX	ISD	DA	NSS
1	Pendidikan Agama Kristen	X	X	X	X
2	Bahasa Indonesia	X	X	X	X
3	Interaksi Manusia dan Komputer	X			
4	Teknologi Komputer	X	X	X	X
5	Praktikum Teknologi Komputer	X	X	X	X

Tabel 2.1 Daftar hubungan antara mata kuliah dan profil lulusan

No	Nama Mata Kuliah	UI/UX	ISD	DA	NSS
6	Matematika Teknik		X		X
7	Logika Matematika		X	X	X
8	Pendidikan Kewarganegaraan	X	X	X	X
9	Jaringan Komputer				X
10	Praktikum Jaringan Komputer				X
11	Algoritma dan Pemrograman	X	X	X	X
12	Praktikum Algoritma dan Pemrograman	X	X	X	X
13	Matematika Diskrit	X	X	X	
14	Arsitektur dan Organisasi Komputer			X	X
15	Statistik	X	X	X	X
16	Struktur Data	X	X	X	X
17	Praktikum Struktur Data	X	X	X	X
18	Infrastruktur LAN				X
19	Praktikum Infrastruktur LAN				X
20	Sistem Basis Data			X	
21	Praktikum Sistem Basis Data			X	
22	Riset Operasi		X		
23	Sistem Operasi			X	X
24	Pendidikan Pancasila	X	X	X	X
25	Pemrograman Web	X			
26	Praktikum Pemrograman Web	X			
27	Kecerdasan Buatan		X		
28	Keamanan Komputer	X		X	X
29	Etika Profesi Teknologi Informasi	X	X	X	X
30	Rekayasa Perangkat Lunak Berorientasi Obyek	X			
31	Praktikum Rekayasa Perangkat Lunak Berorientasi Obyek	X			

Tabel 2.1 Daftar hubungan antara mata kuliah dan profil lulusan

No	Nama Mata Kuliah	UI/UX	ISD	DA	NSS
32	Manajemen Proyek Teknologi Informasi	X			
33	Riset Teknologi Informasi	X	X	X	X
34	Kuliah Kerja Nyata	X	X	X	X
35	Kerja Praktik	X	X	X	X
36	Skripsi	X	X	X	X
37	Pola Desain Antarmuka Pengguna	X			
38	Desain Eksperimental	X			
39	Desain dan Evaluasi Antarmuka	X			
40	Pemodelan Proses Bisnis	X			
41	<i>Test Engineering</i>	X			
42	<i>Internet of Things</i>		X		X
43	<i>Machine Learning</i>		X		
44	Jaringan Syaraf Tiruan		X		
45	<i>Knowledge-Based System</i>		X		
46	Pemrosesan Bahasa Natural		X		
47	Pemrosesan Citra Digital		X		
48	Pemrosesan Sinyal Digital		X		
49	<i>Game Engine</i>		X		
50	Pemrograman Perangkat Bergerak Berbasis Android	X	X	X	
51	Pemrograman Perangkat Bergerak Berbasis iOS	X	X	X	
52	Pemrograman Perangkat Bergerak Berbasis Hybrid	X	X	X	
53	Pemrograman Desktop	X	X	X	
54	Pemrograman Web Lanjut	X	X	X	
55	Administrasi Basis Data			X	
56	<i>Data Warehouse</i>			X	

Tabel 2.1 Daftar hubungan antara mata kuliah dan profil lulusan

No	Nama Mata Kuliah	UI/UX	ISD	DA	NSS
57	Basis Data Terdistribusi			X	
58	Keamanan Basis Data			X	
59	Administrasi Basis Data Non Relasional			X	
60	<i>Enterprise Network</i>				X
61	<i>Cloud Infrastructure</i>				X
62	Pengantar Keamanan Jaringan				X
63	Jaringan Nir Kabel				X
64	Otomasi Jaringan				X
65	Teknologi WAN				X
66	Keamanan Jaringan				X
67	Algoritma Graf		X		
68	Analisis Data Statistik	X	X	X	X
69	Analisis Proses Bisnis	X		X	
70	Bahasa Inggris Informatika	X	X	X	X
71	<i>Competitive Programming</i>		X		
72	<i>Deep Learning</i>	X	X	X	X
73	Desain Game		X		
74	Digital Humanities	X	X		
75	<i>E-Commerce</i>	X		X	
76	<i>E-Government</i>	X		X	
77	<i>Game Audio</i>		X		
78	Grafika Game		X		
79	Kompresi Data		X	X	
80	Komunikasi Bisnis	X	X	X	X
81	Manajemen Kepemimpinan	X	X	X	X
82	Manajemen Konten Web	X		X	
84	Praktikum Keahlian Khusus SAP			X	
85	Program Kreativitas mahasiswa	X	X	X	X

Tabel 2.1 Daftar hubungan antara mata kuliah dan profil lulusan

No	Nama Mata Kuliah	UI/UX	ISD	DA	NSS
86	Proyek Informatika Merdeka	X	X	X	X
87	Semantic Web		X		
88	Sistem Informasi Geografis	X	X	X	X
89	Sistem Pakar	X	X		
90	<i>Technopreneurship</i> dan Manajemen Inovasi	X	X	X	X
91	Teknik Animasi	X	X		
92	<i>Forensic Text</i>		X		
93	<i>UX Writing</i> dan <i>Storytelling</i>	X			
94	Visualisasi Data	X		X	

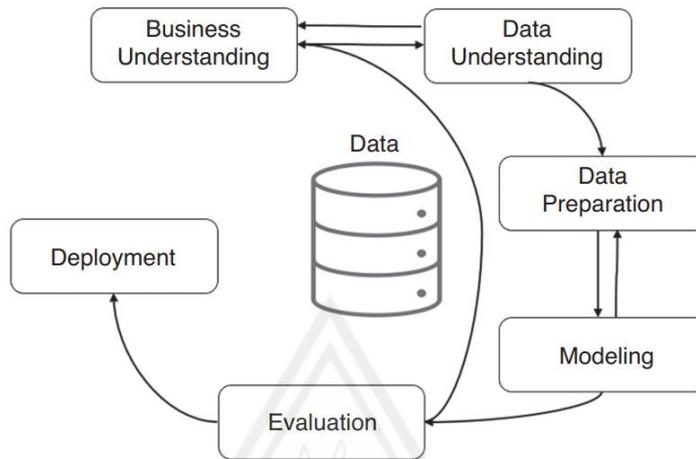
2.2.6 Data Warehouse

Data warehouse adalah repositori yang berisi kumpulan informasi dari berbagai macam sumber, dan disimpan dalam satu skema (Lindstedt, 2015). *Data warehouse* dibangun melalui beberapa proses yaitu *cleaning*, integrasi *data*, transformasi *data*, *data loading*, *data periodic*, dan pembaharuan *data*. *Data warehouse* pada umumnya memiliki model *multidimensional* struktur data yang biasa disebut *data cube*. Pada setiap dimensi memiliki korelasi terhadap kumpulan atribut yang ada pada skema.

2.2.7 Data Mining

Data mining adalah proses analisis sejumlah data untuk menemukan pola, kelompok, data yang tidak biasa dijumpai, dan keterkaitan antar data. Salah satu metode sebagai standar untuk proses *data mining* yang populer digunakan adalah CRISP-DM.

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah salah satu jenis proses bisnis yang umum dipakai. Terdapat 6 tahapan utama pada metodologi ini. Alur dari tahapan tersebut dapat dilihat pada Gambar 2.1.



Gambar 2.1 Diagram alur CRISP-DM

(<https://www.djkn.kemenkeu.go.id/files/images/2022/06/skema.jpg>)

1. *Business Understanding*

Pada tahapan ini, dilakukan pemahaman mendalam mengenai masalah yang dihadapi agar mampu mendefinisikan masalah untuk dianalisis. Selain itu, ditentukan pula capaian apa yang ingin dicapai setelah melakukan analisis data.

2. *Data Understanding*

Pada tahapan ini, data yang telah dikumpulkan akan diidentifikasi dan diteliti, serta dinilai kualitas dan korelasi antar atributnya untuk mencari data yang relevan dengan capaian yang ingin dicapai.

3. *Data Preparation*

Pada tahapan ini, tujuan utama yang ingin ditempuh adalah mempersiapkan data bersih untuk pembuatan model. Data akan dibersihkan dari nilai-nilai nul atau kosong. Dapat pula dilakukan transformasi tipe data jika diperlukan, Contohnya adalah tipe data *string* menjadi *integer*. Selain itu kolom tambahan dapat dibentuk apabila diperlukan sesuai dengan data yang ada.

4. *Modeling*

Pada tahapan ini, fokus utama adalah membangun model pembelajaran mesin sesuai capaian yang ingin diperoleh, sebagaimana tertera pada tahapan *business understanding*. Tahapan ini mencakup pemilihan algoritma, pembuatan data latih dan data uji, dan pembuatan model pembelajaran mesin.

5. *Evaluation*

Pada tahapan ini dilakukan evaluasi menyeluruh terhadap hasil model dan memastikan bahwa capaian yang ingin diraih telah terpenuhi tanpa melewati faktor penting.

6. *Deployment*

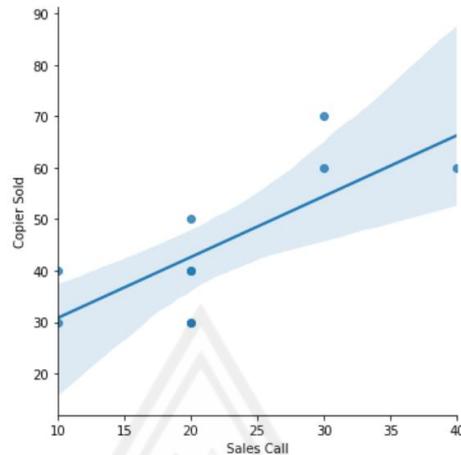
Pada tahapan ini model diimplementasikan untuk capaian yang diinginkan.

2.2.8 Regresi

Regresi biasa digunakan dalam bidang statistika di mana kegunaan dari regresi adalah untuk mengetahui hubungan antara variabel independen dengan variabel dependen (Paoella, 2019). Selain digunakan dalam statistika, regresi juga digunakan sebagai salah satu metode yang paling banyak digunakan dalam proses pembuatan model pada *data mining*. Dalam konsep pembelajaran mesin regresi dapat dibagi ke dalam 3 kategori yaitu regresi linier sederhana, regresi linier multi, dan regresi logistik.

2.2.8.1 Regresi Linier Sederhana

Karakteristik dari Regresi linier adalah model ini hanya dapat melakukan prediksi untuk satu variabel dependen dan satu variabel independen. Terdapat 3 syarat penting yang harus dipenuhi untuk dapat menghasilkan model regresi linier yang baik yaitu relasi antar variabel harus linier, variabel harus memiliki *interval scale*, dan variabel harus merupakan bilangan kontinu. (Bangdiwala, 2018)



Gambar 2.2 Contoh dari plot model regresi linier

Gambar 2.2 merupakan contoh dari hasil *plotting* dari model regresi linier. Model regresi linier digunakan untuk memprediksi *Copier sold* (variabel dependen) terhadap *Sales Call* (variabel independen). Perhitungan yang digunakan dalam menentukan garis plot tersebut dapat dilihat pada Persamaan 2.1 di bawah.

$$\hat{Y} = a + bX \quad (2.1)$$

\hat{Y} = target atau variabel dependen

a = konstanta

b = koefisien regresi

X = variabel independen

2.2.8.2 Regresi Linier Multi

Regresi linier multi merupakan salah satu pengembangan dari metode regresi linier. Penggunaan model regresi linier tidak dapat menjawab permasalahan yang dalam kebanyakan kasus dikarenakan variabel independen yang diteliti berjumlah lebih dari 1 dan biasa ditemukan multikolinieritas antar variabel sehingga tidak memungkinkan untuk mengambil satu variabel independen saja (Kurniatullah & Pramudi, 2017). Contoh dari penggunaan model regresi linier multi adalah prediksi gaji tahunan karyawan berdasarkan latar belakang pendidikan, usia, lingkungan hidup karyawan (Guido & Muller, 2016).

Rumus perhitungan persamaan regresi linier multi dapat dilihat pada Persamaan 2.3.

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k \quad (2.2)$$

\hat{Y} = target atau variabel dependen

a = konstanta

b = koefisien regresi

X = variabel independen

Dalam contoh memprediksi gaji tahunan karyawan yang dilakukan oleh (Guido & Muller, 2016) hubungan antara variabel dependen dengan variabel independen tidak terlihat dengan jelas. Meskipun demikian, model regresi linier multi dapat menunjukkan bahwa variabel independen memiliki pengaruh dalam hasil prediksi target. Pemilihan memilih variabel independen dalam proses pelatihan model regresi linier multi adalah hal yang penting untuk dilakukan.

2.2.9 *Super Vector Regression (SVR)*

SVR adalah bentuk model regresi dengan memanfaatkan bantuan dari *kernelized vector* untuk mendefinisikan sebuah *hyperplane* dalam memprediksi atau mengklasifikasikan variabel dependen (Guido & Muller, 2016). Penggunaan *kernelized support vector* pada penelitian ini ditujukan untuk melihat hubungan non linier antara variabel independen terhadap variabel dependen. Pada model SVR terdapat lima buah *kernel* yang tersedia dalam *library sklearn* yaitu *linear*, *poly*, *RBF*, *sigmoid*, dan *precomputed*. Pemilihan *kernel* yang digunakan dalam penelitian ini adalah menggunakan *kernel poly* dan *RBF*, pemilihan tersebut berdasarkan dari karakteristik bentuk *hyperplane* yang dihasilkan dari kedua *kernel* tersebut memiliki bentuk non linier.

2.2.10 Metode Evaluasi

Proses evaluasi data dan model dibagi ke dalam 3 tahapan. Tahap pertama yaitu analisis multikolinearitas dengan menggunakan uji VIF. Tahap kedua yaitu

dilakukan analisis dengan metode estimator OLS untuk mendapatkan nilai ketepatan prediksi sistem, tingkat korelasi data, dan standar eror untuk tiap variabel independen dalam memprediksi variabel dependen. Tahap ketiga dilakukan analisis untuk tiap model yang dibuat dengan membandingkan nilai MAE dan R^2 .

2.2.10.1 Ordinary Least Squares (OLS)

OLS adalah salah satu metode estimasi untuk model regresi linier. Kegunaan dari OLS ini adalah untuk membuat estimasi garis regresi yang optimal dan sederhana untuk meminimalkan jarak eror dari prediksi dengan hasil aktualnya (Guido & Muller, 2016). Rumus dalam perhitungan OLS untuk model regresi linier dapat dilihat pada Persamaan 2.3. Selain itu dengan menggunakan OLS penulis dapat menghitung *p-value* dari hasil garis regresi yang dihasilkan OLS. Hasil dari *p-value* menghasilkan suatu asumsi terhadap hipotesis seberapa signifikan variabel independen mempengaruhi variabel dependen.

$$m = \left| \frac{\sum_{i=0}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^N (x_i - \bar{x})^2} \right| \quad (2.3)$$

$m = slope$

$y = target$ atau variabel dependen

$x = variabel$ independen

$\bar{x} = mean$ variabel independen

$\bar{y} = mean$ variabel dependen

2.2.10.2 MAE

Mean Absolute Error (MAE) adalah salah satu metode dalam menganalisis eror yang terjadi pada model regresi. Nilai MAE didapatkan dari nilai rata-rata absolut dari perbedaan antara data prediksi oleh pembelajaran mesin dengan data sesungguhnya. Semakin kecil nilai MAE maka semakin kecil jarak kesalahan antara data prediksi dengan data sebenarnya (Paoella, 2019). Dalam menentukan MAE digunakan Persamaan 2.5.

$$MAE = \left| \frac{\sum_{i=1}^N predicted\ value_i - actual\ value_i}{N} \right| \quad (2.5)$$

2.2.10.3 R^2 dan *adjusted-R²*

Nilai R^2 merupakan salah satu indikator yang penting untuk mengevaluasi model regresi. R^2 merepresentasikan kemampuan model untuk memprediksi variabel dependen berdasarkan variabel independen. Jika nilai R^2 semakin mendekati 1 maka model tersebut dianggap dapat menjelaskan hubungan masing-masing variabel independen terhadap variabel dependen.

Terdapat satu kelemahan dalam penggunaan nilai R^2 untuk mengevaluasi model yaitu semakin banyak variabel independen yang digunakan. Meskipun variabel yang ditambahkan dalam model tersebut tidak memiliki hubungan apapun terhadap variabel dependen, nilai R^2 akan tetap bertambah. Untuk mengatasi hal tersebut penulis menggunakan uji *adjusted-R²*. Dalam pengujian *adjusted-R²* jumlah variabel akan dimasukkan ke dalam pembagi sehingga nilai R^2 tidak akan bertambah jika model tidak dapat menjelaskan hubungan variabel independen dan variabel dependen dengan lebih baik. Dalam menentukan R^2 dan *adjusted-R²* dapat dilihat pada persamaan 2.6 dan 2.7.

$$R^2 = 1 - \frac{SSR}{SST} \quad (2.6)$$

SSR = *Sum of Square Residuals*

SST = *Sum of Square Total*

$$Adjusted-R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \quad (2.7)$$

N = jumlah data

p = jumlah variabel independen

2.2.10.4 Variance Inflation Factor (VIF)

VIF digunakan untuk mengukur tingkat *variance* variabel independen yang dihasilkan terhadap model regresi. Semakin tinggi nilai VIF menandakan bahwa semakin tinggi kemungkinan variabel independen tersebut memiliki korelasi terhadap variabel independen lainnya (multikolinearitas). Terdapat batasan nilai yang dapat digunakan sebagai acuan untuk menentukan tingkat multikolinearitas suatu variabel. Jika nilai $VIF < 5$ maka kemungkinan multikolinearitas variabel independen cukup rendah. Jika nilai $VIF \geq 5$ dan nilai $VIF \leq 10$ maka kemungkinan variabel tersebut memiliki korelasi yang tinggi terhadap variabel independen lainnya. Jika nilai $VIF > 10$ maka dapat dipastikan variabel independen tersebut memiliki korelasi yang tinggi terhadap variabel independen lainnya. Perhitungan VIF dilakukan dengan Persamaan 2.8.

$$VIF = \frac{1}{1-R^2} \quad (2.8)$$

2.2.10.5 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE merupakan salah satu teknik dalam metode *oversampling*, dengan melakukan pendekatan yang mengarah pada membuat data sintetis dari kelas data yang memiliki distribusi data sedikit jika (*minority class*) (Lusa, 2013). SMOTE biasa digunakan dalam model pembelajaran mesin klasifikasi data. Dalam model klasifikasi banyaknya kelas atau kelompok data sangat berpengaruh dalam hasil klasifikasi, hal tersebut dikarenakan model akan memiliki kecenderungan untuk mengklasifikasikan fitur yang ada ke dalam kelas dengan distribusi data tertinggi (*majority class*).

BAB III

METODOLOGI PENELITIAN

3.1 Kebutuhan Sistem

3.1.1 Kebutuhan Non Fungsional

3.1.1.1 Perangkat Keras

Perangkat keras yang digunakan untuk menjalankan sistem adalah laptop MSI seri GL63 8SC dengan spesifikasi sebagai berikut:

- Prosesor: Intel i7 8th Gen
- RAM: 16 GB
- Penyimpanan: 500GB SSD
- Sistem Operasi: Windows 10 64-bit

3.1.1.2 Perangkat Lunak

Untuk membangun sistem digunakan beberapa perangkat lunak (*software*) adalah sebagai berikut:

- Python versi 3.10
- Jupyter *Notebook* (Anaconda3) versi 6.4.8
- PHP *Framework* Laravel versi 9
- XAMPP versi 8.0.28
- Google Chrome
- Microsoft Excel
- Visual Studio Code

3.1.2 Kebutuhan Fungsional

Keperluan terkait kebutuhan pustaka (*library*), modul, dan paket (*package*) dalam pembangunan sistem adalah sebagai berikut:

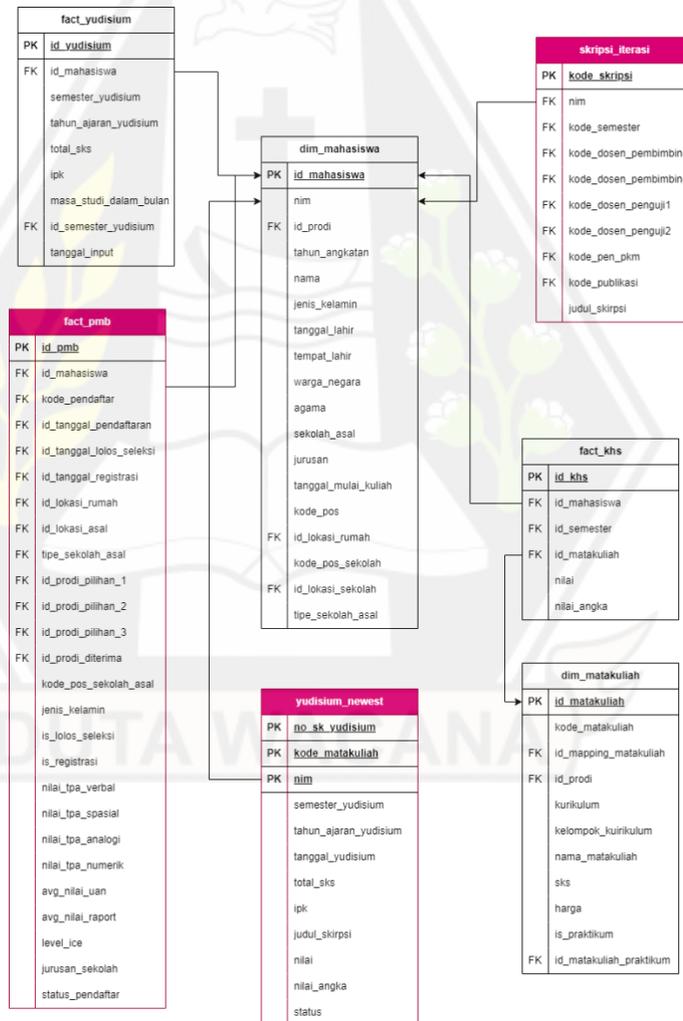
- Numpy versi 1.21.5
Numpy digunakan sebagai pustaka (*library*) yang digunakan dalam perhitungan pada pengolahan data dan saat analisis data.
- Pandas versi 1.4.2
Pandas digunakan untuk mengubah *file* yang berisi data dengan ekstensi csv, json, sql, dan lainnya menjadi *data frame* agar dapat mudah diolah dengan menggunakan pustaka lainnya di python.
- Sklearn versi 1.0.2
Sklearn digunakan untuk pustaka (*library*) utama yang digunakan dalam model. Sklearn digunakan dalam pembuatan data latih dan prediksi dengan *split_test*, pengaplikasian algoritma linier regresi multi, SVR dan evaluasi model seperti penggunaan modul *mean_average_error* untuk menghitung MAE.
- Matplotlib versi 3.5.1
Matplotlib digunakan untuk pustaka (*library*) yang digunakan untuk melihat persebaran data dengan *chart* yang digunakan untuk melihat signifikansi variabel.
- Seaborn versi 0.11.2
Seaborn digunakan untuk menampilkan sebaran data yang lebih bervariasi dibandingkan matplotlib yaitu dalam bentuk *boxplot*, *pairplot*, dan bentuk lainnya yang tidak dimiliki pustaka matplotlib
- Statsmodel versi 0.13.2
Statsmodel digunakan untuk menampilkan hasil analisa dari model yang dibuat oleh estimator OLS. Hasil analisa tersebut meliputi beberapa hal penting seperti nilai *covariance*, *R-squared*, koefisien variabel dependen, dan lainnya.

3.2 Metode Penelitian

Penelitian ini menerapkan metode atau pendekatan CRISP-DM dalam membentuk sistem prediksi. Tahapan tersebut antara lain:

3.2.1 Persiapan Awal

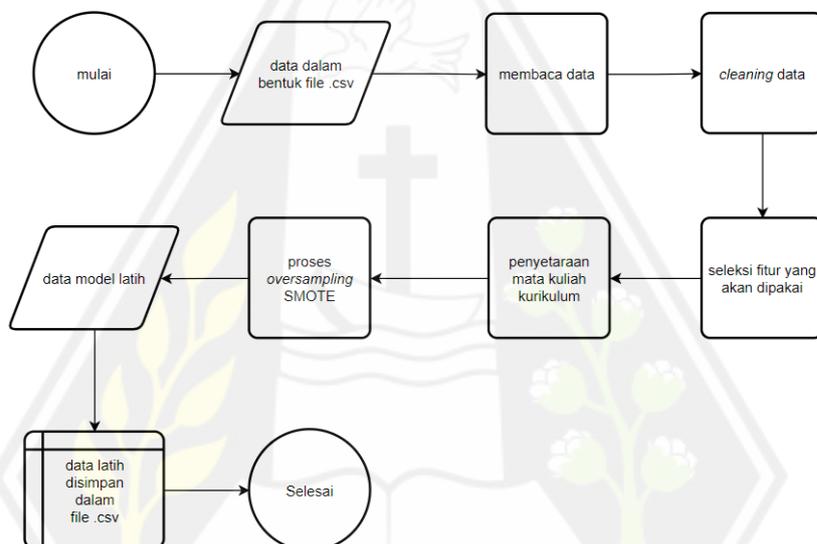
Pada tahap ini penulis melakukan proses pengumpulan data primer dan data sekunder. Dalam proses pengumpulan data primer dilakukan studi literatur, dan tanya jawab dengan dosen pembimbing dan Kepala Program Studi Informatika UKDW. Data sekunder didapatkan penulis dari *data warehouse* yang dimiliki Prodi Informatika UKDW. Bentuk data sekunder yang didapat adalah hasil *dump* berbentuk *file .csv*. Skema relasi antar tabel yang digunakan dalam penelitian ini dapat dilihat pada Gambar 3.1. Pada Gambar 3.1 tabel yang memiliki *header* dengan warna merah muda merupakan data *raw* berbentuk tabel *.xlsx*.



Gambar 3. 1 Skema Relasi Tabel

3.2.2 Prapemrosesan (*Preprocessing*)

Pada tahap prapemrosesan dilakukan proses pembersihan (*cleaning*) data untuk menghilangkan fitur-fitur yang tidak akan digunakan dalam model regresi dan untuk mengurangi *noise* atau data-data yang dapat merusak model. Penyetaraan mata kuliah dilakukan untuk Kurikulum 2011 revisi 2013, Kurikulum 2017, dan Kurikulum 2017 revisi 2019. Setelah itu penulis memberikan label kepada datum kategorikal menjadi data numerik agar dapat digunakan dalam model regresi. Langkah-langkah yang dilakukan dalam tahap prapemrosesan dapat dilihat pada Gambar 3.2.



Gambar 3.2 Diagram Alir Tahap Prapemrosesan

3.2.3 Pelatihan Model Pembelajaran Mesin

Setelah penulis mendapatkan data latih yang sudah dianggap baik, data latih tersebut digunakan untuk membuat model pembelajaran mesin dengan menggunakan dua metode yaitu regresi linier multi dan *Super Vector Regression* (SVR). Pada tahapan ini model akan melakukan proses pembelajaran dari fitur-fitur yang terdapat pada data latih untuk dapat memprediksi IPK dan lama studi. Pada penelitian ini untuk membuat model pembelajaran mesin regresi linier multi dan SVR, penulis menggunakan pustaka yang tersedia pada *package scikit-learn*.

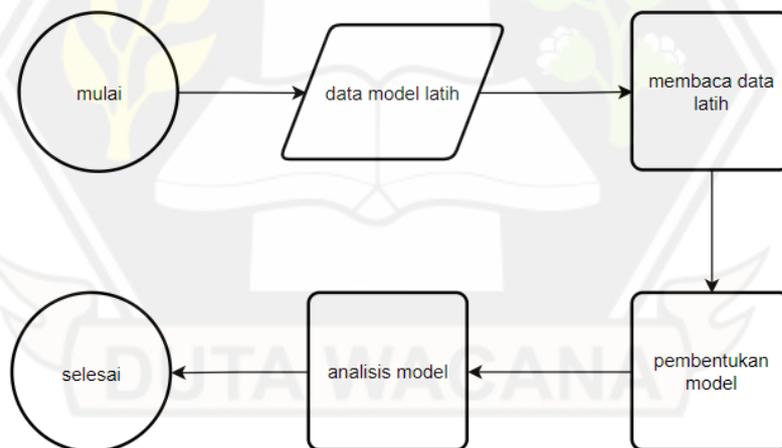
3.2.4 Pengujian (*Testing*)

Tahap pengujian ini dilakukan dengan menggunakan data dari hasil pemisahan data latih yang disebut dengan data tes dengan persentase jumlah data tes sebesar 20% dari data latih yang digunakan. Data yang digunakan dalam data latih dan data tes diambil secara acak dengan nilai `random_state` sebesar 42. Pada tahapan ini penulis melakukan perbandingan nilai R^2 , *adjusted-R²*, dan MAE untuk setiap model regresi yang sudah dibuat. Dari hasil perbandingan tersebut penulis dapat menentukan model dengan fitur dan metode regresi yang terbaik untuk diimplementasikan pada sistem.

3.2.5 Alur Kerja Sistem

3.2.5.1 Alur Proses Pelatihan Model

Pada penelitian ini, penulis membuat sebanyak 18 model regresi dengan tujuan untuk menemukan model dengan fitur dan hasil yang terbaik. Perbedaan dari masing-masing model dapat dilihat pada Tabel 3.1.



Gambar 3.3. Diagram alir proses pelatihan model

Gambar 3.3 menunjukkan proses yang dilakukan untuk pelatihan pada tiap model dengan menggunakan fitur dan metode regresi yang tertera pada Tabel 3.1.

Tabel 3. 1 Tabel keterangan model

Kelompok Model	Variabel Independen	Variabel Dependen	Metode Regresi
Model 1	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib	IPK	Regresi Linier Multi
Model 2	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib	IPK	SVR – RBF
Model 3	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib	IPK	SVR – Poly
Model 4	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib	Lama Studi	Regresi Linier Multi
Model 5	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib	Lama Studi	SVR – RBF
Model 6	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib	Lama Studi	SVR – Poly

Tabel 3.1 Tabel keterangan model (lanjutan)

Kelompok Model	Variabel Independen	Variabel Dependen	Metode Regresi
Model 7	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib, kelompok_matkul_pilihan_wajib _profil_ui/ux, kelompok_matkul_pilihan_wajib _profil_isd, kelompok_matkul_pilihan_wajib _profil_da, kelompok_matkul_pilihan_wajib _profil_nss	IPK	Regresi Linier Multi
Model 8	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib, kelompok_matkul_pilihan_wajib _profil_ui/ux, kelompok_matkul_pilihan_wajib _profil_isd, kelompok_matkul_pilihan_wajib _profil_da, kelompok_matkul_pilihan_wajib _profil_nss	IPK	SVR – RBF

Tabel 3.1 Tabel keterangan model (lanjutan)

Kelompok Model	Variabel Independen	Variabel Dependen	Metode Regresi
Model 9	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_pilihan_wajib_pr ofil_ui/ux, kelompok_matkul_pilihan_wajib_pr ofil_isd, kelompok_matkul_pilihan_wajib_pr ofil_da, kelompok_matkul_pilihan_wajib_pr ofil_nss	IPK	SVR – Poly
Model 10	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_pilihan_wajib_pr ofil_ui/ux, kelompok_matkul_pilihan_wajib_pr ofil_isd, kelompok_matkul_pilihan_wajib_pr ofil_da, kelompok_matkul_pilihan_wajib_pr ofil_nss	Lama Studi	Regresi Linier Multi

Tabel 3.1 Tabel keterangan model (lanjutan)

Kelompok Model	Variabel Independen	Variabel Dependen	Metode Regresi
Model 11	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_pilihan_wajib_pr ofil_ui/ux, kelompok_matkul_pilihan_wajib_pr ofil_isd, kelompok_matkul_pilihan_wajib_pr ofil_da, kelompok_matkul_pilihan_wajib_pr ofil_nss	Lama Studi	SVR – RBF
Model 12	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_pilihan_wajib_pr ofil_ui/ux, kelompok_matkul_pilihan_wajib_pr ofil_isd, kelompok_matkul_pilihan_wajib_pr ofil_da, kelompok_matkul_pilihan_wajib_pr ofil_nss	Lama Studi	SVR – Poly

Tabel 3.1 Tabel keterangan model (lanjutan)

Kelompok Model	Variabel Independen	Variabel Dependen	Metode Regresi
Model 13	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib, kelompok_matkul_relasi_profil_ui/ux-isd-da, kelompok_matkul_relasi _profil_isd-da-nss, kelompok_matkul_relasi _profil_ui/ux-da-nss, kelompok_matkul_relasi _profil_ui/ux, kelompok_matkul_relasi _profil_isd, kelompok_matkul_relasi _profil_da, kelompok_matkul_relasi_profil_nss	IPK	Regresi Linier Multi
Model 14	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib, kelompok_matkul_relasi_profil_ui/ux-isd-da, kelompok_matkul_relasi _profil_isd-da-nss, kelompok_matkul_relasi _profil_ui/ux-da-nss, kelompok_matkul_relasi _profil_ui/ux, kelompok_matkul_relasi _profil_isd, kelompok_matkul_relasi _profil_da, kelompok_matkul_relasi_profil_nss	IPK	SVR – RBF

Tabel 3.1 Tabel keterangan model (lanjutan)

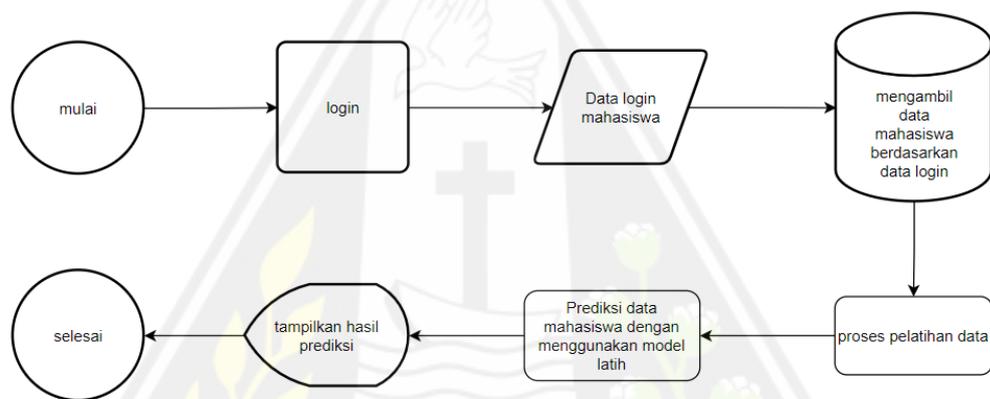
Kelompok Model	Variabel Independen	Variabel Dependen	Metode Regresi
Model 15	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_relasi_profil_ui/ux-isd-da, kelompok_matkul_relasi_profil_isd-da-nss, kelompok_matkul_relasi_profil_ui/ux-da-nss, kelompok_matkul_relasi_profil_ui/ux, kelompok_matkul_relasi_profil_isd, kelompok_matkul_relasi_profil_da, kelompok_matkul_relasi_profil_nss	IPK	SVR – Poly
Model 16	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_relasi_profil_ui/ux-isd-da, kelompok_matkul_relasi_profil_isd-da-nss, kelompok_matkul_relasi_profil_ui/ux-da-nss, kelompok_matkul_relasi_profil_ui/ux, kelompok_matkul_relasi_profil_isd, kelompok_matkul_relasi_profil_da, kelompok_matkul_relasi_profil_nss	Lama Studi	Regresi Linier Multi

Tabel 3.1 Tabel keterangan model (lanjutan)

Kelompok Model	Variabel Independen	Variabel Dependen	Metode Regresi
Model 17	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_relasi_profil_ui/ux-isd-da, kelompok_matkul_relasi _profil_isd-da-nss, kelompok_matkul_relasi _profil_ui/ux-da-nss, kelompok_matkul_relasi _profil_ui/ux, kelompok_matkul_relasi _profil_isd, kelompok_matkul_relasi _profil_da, kelompok_matkul_relasi_profil_nss	IPK	SVR – RBF
Model 18	profil, jurusan_sma_asal, provinsi, sekolah_asal, tipe_sekolah_asal, kelompok_matkul_humaniora, kelompok_matkul_ti_wajib_seluruh profil, kelompok_matkul_relasi_profil_ui/ux-isd-da, kelompok_matkul_relasi _profil_isd-da-nss, kelompok_matkul_relasi _profil_ui/ux-da-nss, kelompok_matkul_relasi _profil_ui/ux, kelompok_matkul_relasi _profil_isd, kelompok_matkul_relasi _profil_da, kelompok_matkul_relasi_profil_nss	Lama Studi	SVR – Poly

3.2.5.2 Alur Proses Prediksi

Proses prediksi sistem IPK dan lama studi dapat dilihat pada Gambar 3.4. Proses dimulai dengan mendapatkan data NIM dan nama mahasiswa setelah mahasiswa berhasil *login*. NIM tersebut akan digunakan untuk mencari data terkait identitas dan hasil studi mahasiswa terkait. Setelah mendapatkan data tersebut akan dilakukan proses pelatihan model regresi untuk model IPK dan model lama studi. Selanjutnya data identitas dan hasil studi mahasiswa akan di prediksi dengan model regresi yang sudah dilatih, kemudian hasil tersebut akan ditampilkan pada halaman web.

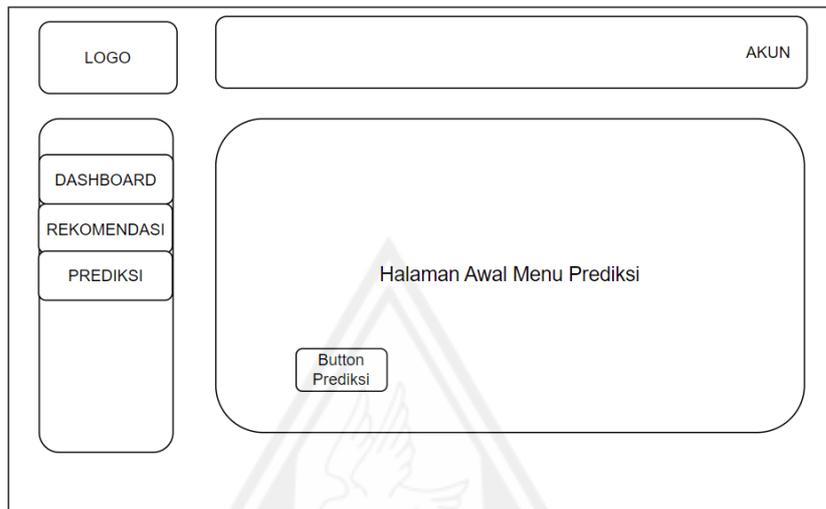


Gambar 3. 4 Diagram alir proses prediksi dalam sistem

3.2.6 Rancangan Antar Muka Sistem

Penelitian ini merupakan kelanjutan dari penelitian (Kurniawan, 2022) dan (Surya, 2022) dengan fokus penelitian yang berbeda; (Kurniawan, 2022) fokus penelitian terhadap rekomendasi topik skripsi, sedangkan (Surya, 2022) fokus penelitian pada rekomendasi profil lulusan. Pada bagian ini, rancangan antar muka yang dibangun fokus pada halaman terkait prediksi IPK dan lama studi mahasiswa.

3.2.6.1 Halaman awal menu prediksi



Gambar 3. 5 Rancangan awal halaman menu prediksi sistem

Gambar 3.5 menunjukkan rancangan antar muka untuk halaman awal setelah mahasiswa berhasil *login* dan memilih menu PREDIKSI pada *side bar*. Setelah mahasiswa menekan tombol prediksi mahasiswa akan diarahkan menuju halaman hasil prediksi.

3.2.6.2 Halaman Hasil Prediksi



Gambar 3. 6 Rancangan awal halaman hasil prediksi

Gambar 3.6 menunjukkan rancangan antar muka untuk halaman hasil prediksi IPK dan lama studi. Hasil prediksi akan dibagi menjadi 4 kategori sesuai dengan profil lulusan yang ada pada Prodi Informatika UKDW yaitu profil UI/UX, ISD, DA, dan NSS.

3.2.7 Evaluasi

Evaluasi yang dilakukan dalam penelitian ini dibagi ke dalam dua tahapan, tahap evaluasi awal untuk menentukan kelompok data dari pembagian mata kuliah dan tahap evaluasi akhir terhadap model prediksi IPK dan lama studi. Pada tahap evaluasi awal penulis menggunakan *k-fold cross validation* untuk mengevaluasi hasil R^2 dan MAE dari kelompok data latih untuk tiap model regresi. Setelah ditemukan kelompok data latih yang akan digunakan, penulis melakukan uji VIF untuk melihat multikolinieritas antar variabel independen. Kemudian pada tahap evaluasi akhir dilakukan perhitungan statistika dengan metode *Ordinary Least Square* (OLS). Pada tahap ini penulis menggunakan nilai R^2 dan *adjusted-R²* yang didapatkan dari metode OLS untuk melihat seberapa baik model dapat memprediksi IPK dan lama studi. OLS juga digunakan penulis untuk melihat tingkat korelasi antara variabel independen terhadap variabel dependen atau variabel prediksi IPK dan lama studi dengan melihat koefisien yang dihasilkan model. Penulis menggunakan nilai MAE sebagai tolak ukur untuk tingkat eror yang dihasilkan model regresi dalam memprediksi IPK dan lama studi.

BAB IV

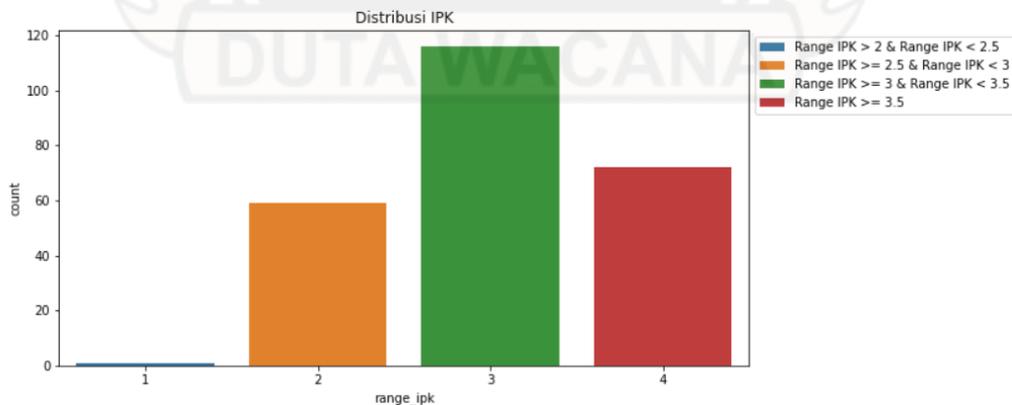
IMPLEMENTASI DAN PEMBAHASAN

1.1 Implementasi Sistem

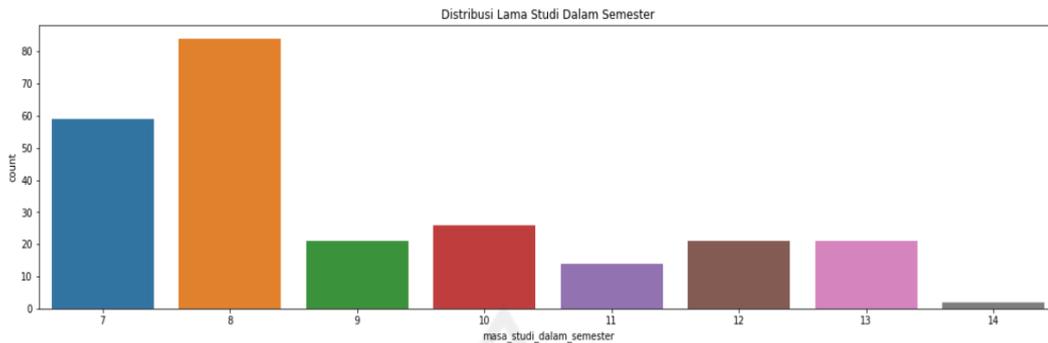
1.1.1 Data Latih

Tabel 3.1 menunjukkan 18 model beserta variabel independen, variabel dependennya, dan model regresi. Jika variabel dependen dan model regresi dihilangkan hanya akan ada 3 kelompok variabel independen unik. Kelompok variabel independen tersebut dipengaruhi akibat adanya perbedaan dalam tindakan penyetaraan mata kuliah dan pelabelannya. Data latih yang akan digunakan dalam sistem dipilih berdasarkan kemampuan variabel independen dalam menjelaskan variabel dependen, di mana hasil tersebut akan diketahui melewati komparasi antar model.

Distribusi data berdasarkan variabel dependen merupakan salah satu faktor yang terpenting dalam model regresi. Model regresi hanya dapat memberikan hasil prediksi IPK dan lama studi mahasiswa berdasarkan IPK dan lama studi yang ada pada data latih, oleh sebab itu batas atas dan batas bawah dari variabel dependen beserta jumlah data akan mempengaruhi kemampuan model untuk memprediksi. Gambar 4.1 merupakan visualisasi distribusi IPK dengan membagi IPK ke dalam 4 kelompok dengan jarak IPK sebesar 0.5, untuk melihat distribusi lama studi mahasiswa dalam semester dapat dilihat pada Gambar 4.2.



Gambar 4. 1 Distribusi IPK data latih



Gambar 4. 2 Distribusi lama studi data latih

1.1.2 Penyetaraan Mata Kuliah dan Pelabelan Variabel Non Numerik

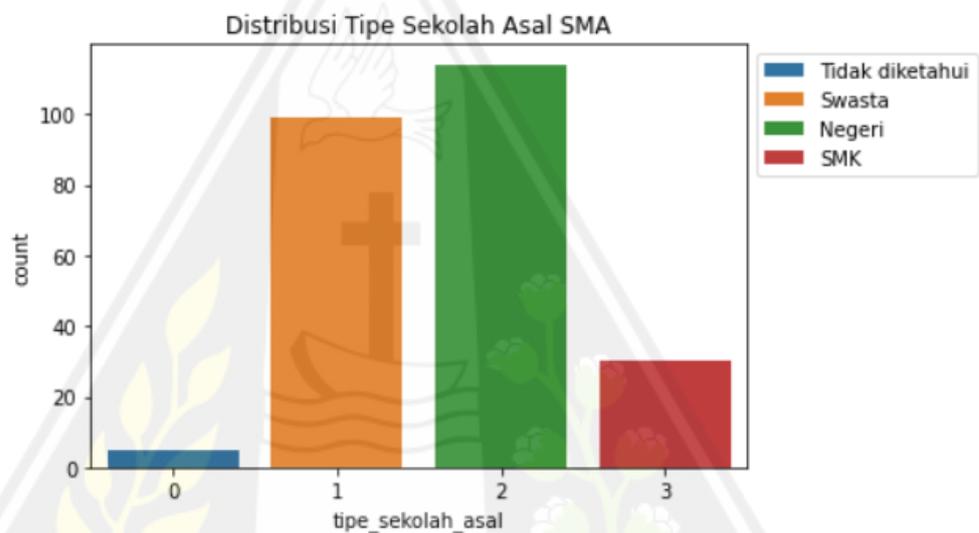
Penyetaraan mata kuliah merupakan tahapan paling penting untuk menghasilkan data latih sistem. Pada penelitian ini data nilai dan data mahasiswa yang digunakan untuk data latih dan data tes adalah data mahasiswa tahun 2013 hingga 2019. Dalam kurun waktu tersebut terdapat 3 kurikulum yang berlaku yaitu Kurikulum 2011 revisi 2013, Kurikulum 2017 dan Kurikulum 2017 revisi 2019. Untuk menjaga konsistensi data nilai mata kuliah dan mempermudah model dalam memprediksi variabel dependen, maka 2 kurikulum tersebut disetarakan ke dalam Kurikulum 2021. Penyetaraan mata kuliah berdasarkan kurikulum dapat dilihat pada Lampiran D.

Pelabelan untuk data `tipe_sekolah` dilakukan dengan mengubah tipe SMA asal yaitu swasta, negeri, tidak diketahui dan SMK secara berurutan menjadi 1, 2, 3, 0. Distribusi data terkait tipe SMA asal dapat dilihat pada Gambar 4.3. Pelabelan untuk jurusan SMA asal dilakukan dengan mengubah jurusan IPA, IPS, Bahasa, keahlian khusus, dan tidak diketahui secara berurutan menjadi 1, 2, 3, 4, 0. Distribusi data terkait jurusan SMA asal dapat dilihat pada Gambar 4.4.

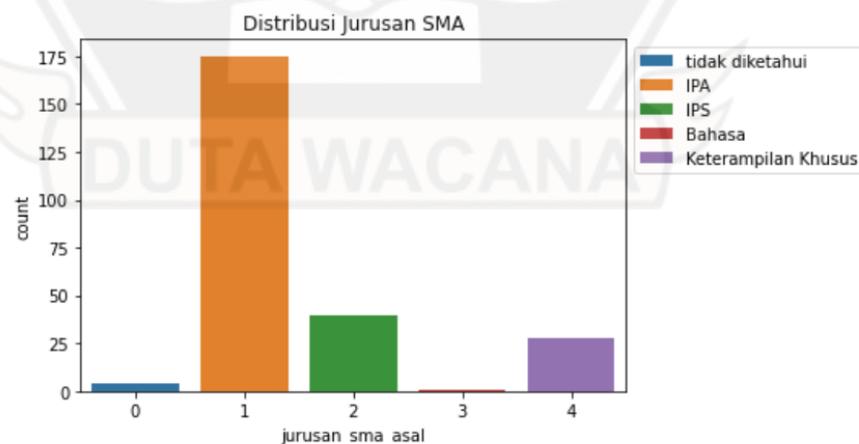
Perlakuan khusus dilakukan untuk labelisasi data variabel `provinsi_asal_sekolah` dengan mengubah nama Provinsi menjadi angka 0 hingga 39, dengan angka 0 adalah label provinsi yang tidak diketahui, angka 1 merupakan Provinsi DIY, angka 39 merupakan label asal sekolah dari luar negeri. Khusus untuk label 2 hingga 38 memiliki aturan labelisasi yaitu semakin kecil

nilai label maka semakin dekat jarak relatif Provinsi tersebut dengan Provinsi DIY. Distribusi data dapat dilihat pada Gambar 4.5

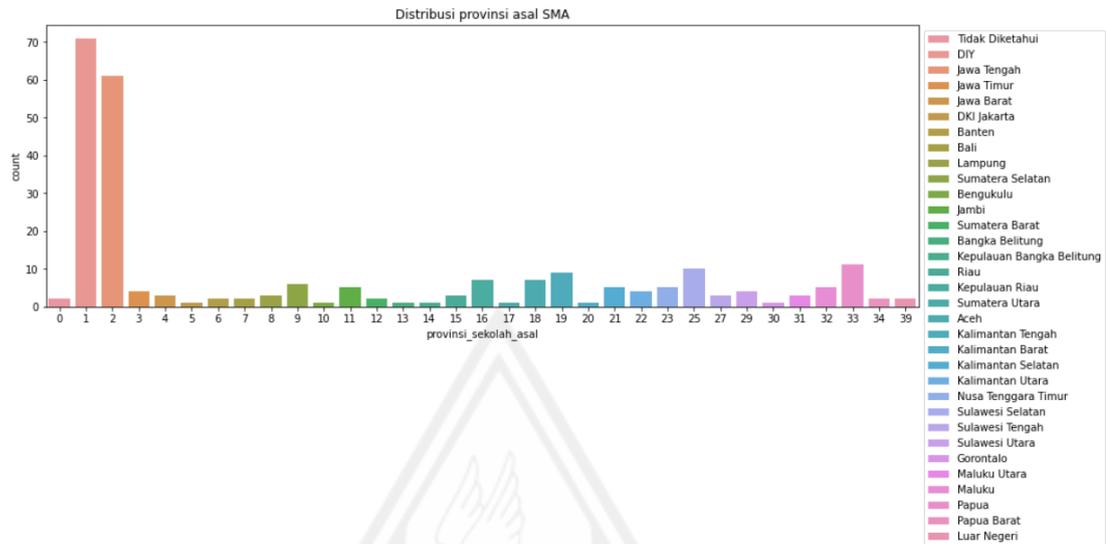
Pada penelitian ini untuk mendapatkan data profil lulusan mahasiswa peneliti melakukan pengumpulan data primer. Pelabelan profil lulusan diperoleh dengan mencocokkan judul skripsi dengan profil lulusan. Data judul skripsi tersebut didapat dari file `skripsi_iterasi`. Distribusi data profil lulusan dapat dilihat pada Gambar 4.6.



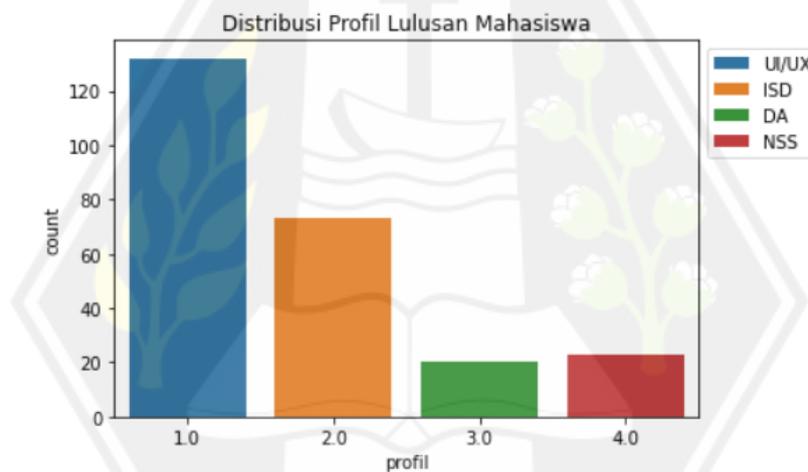
Gambar 4. 3 Distribusi tipe SMA asal pada data latih



Gambar 4. 4 Distribusi jurusan SMA asal pada data latih



Gambar 4. 5 Distribusi Provinsi SMA asal mahasiswa data latih

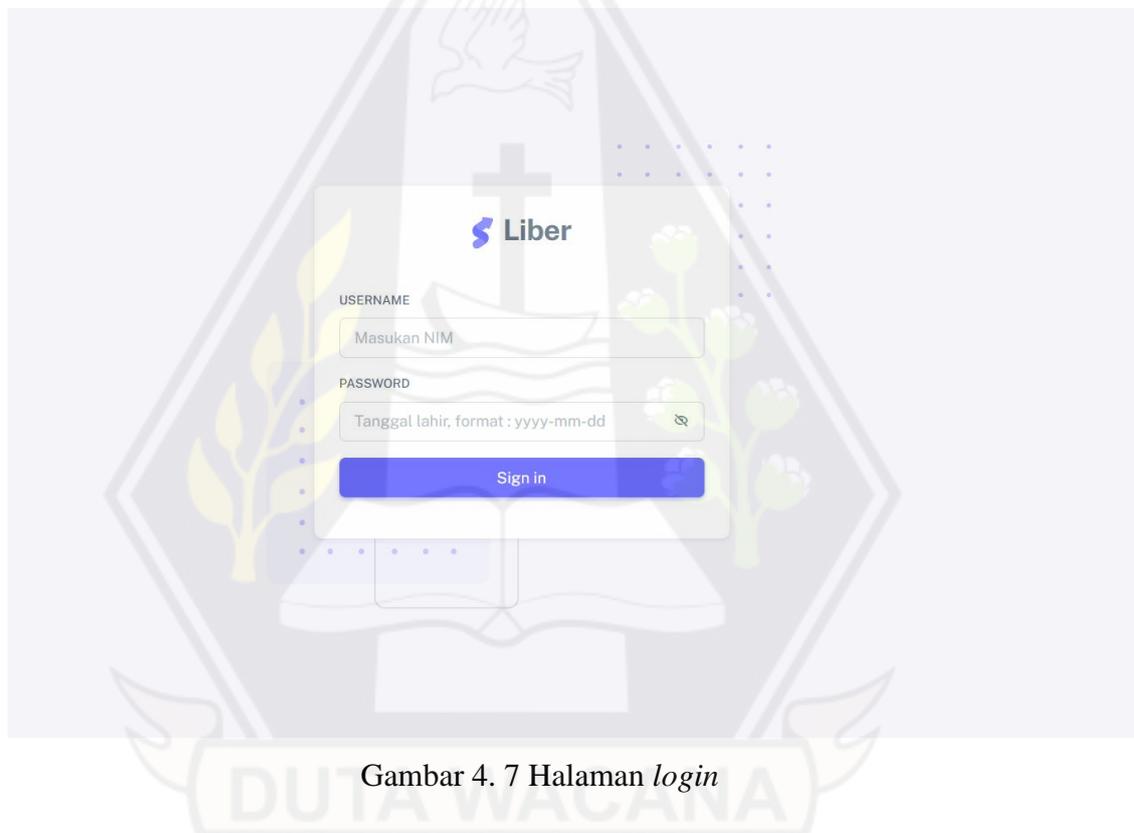


Gambar 4. 6 Distribusi data profil lulusan mahasiswa data latih

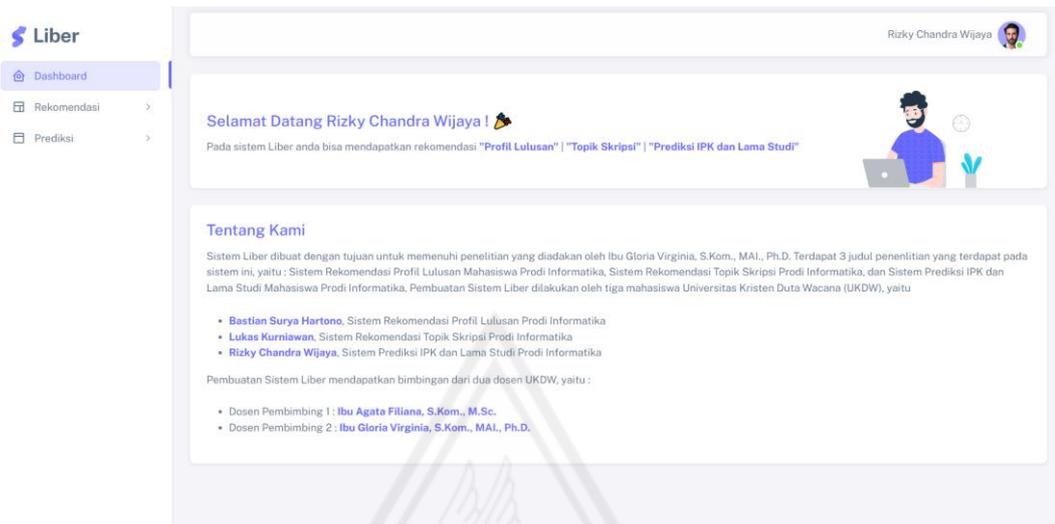
1.1.3 Implementasi Antar Muka

Mahasiswa dapat melihat prediksi IPK dan lama studi dengan cara *login* ke dalam sistem dengan mengisi NIM ke kolom *username* dan tanggal lahir ke dalam kolom *password* dengan format (YYYY-MM-DD). Tampilan halaman *login* dapat dilihat pada Gambar 4.7. Setelah proses *login* maka mahasiswa akan diarahkan ke halaman *dashboard* yang dapat dilihat pada Gambar 4.8. Kemudian mahasiswa dapat memilih menu prediksi pada *side bar*.

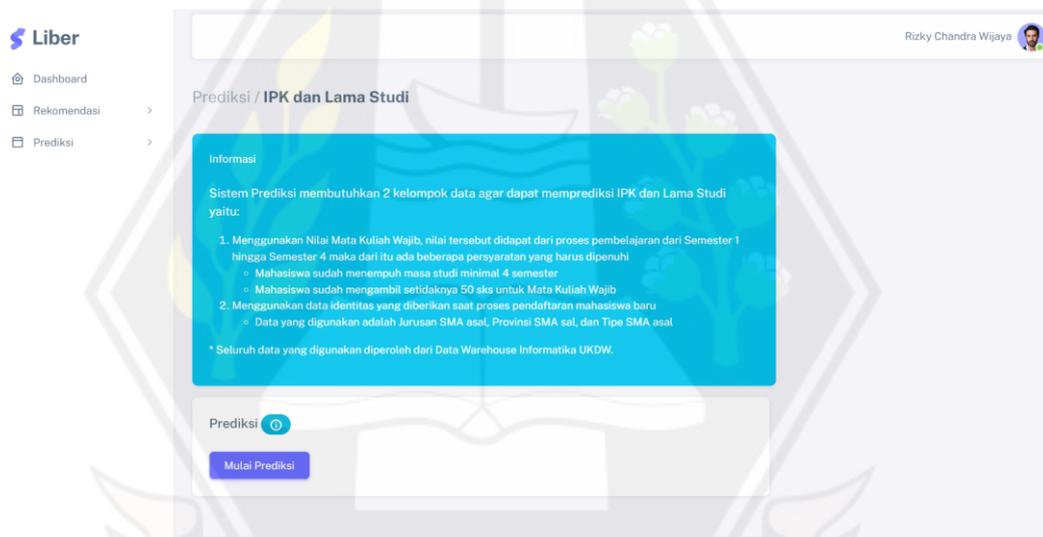
Proses prediksi sistem berjalan ketika mahasiswa menekan tombol mulai prediksi yang dapat dilihat pada Gambar 4.9. Tombol tersebut memanggil *pyscript* yang sudah terintegrasi dengan halaman web untuk mengambil data *id_mahasiswa* dan nama mahasiswa, kemudian *id_mahasiswa* tersebut akan digunakan untuk mengambil data dari file *.csv mahasiswa_cleaned* dan file *.csv khs_cleaned*. Hasil dari sistem prediksi dapat dilihat pada Gambar 4.10. Hasil prediksi mencakup 4 kategori profil lulusan beserta IPK dan lama studi jika mahasiswa tersebut memilih jalur profil tersebut.



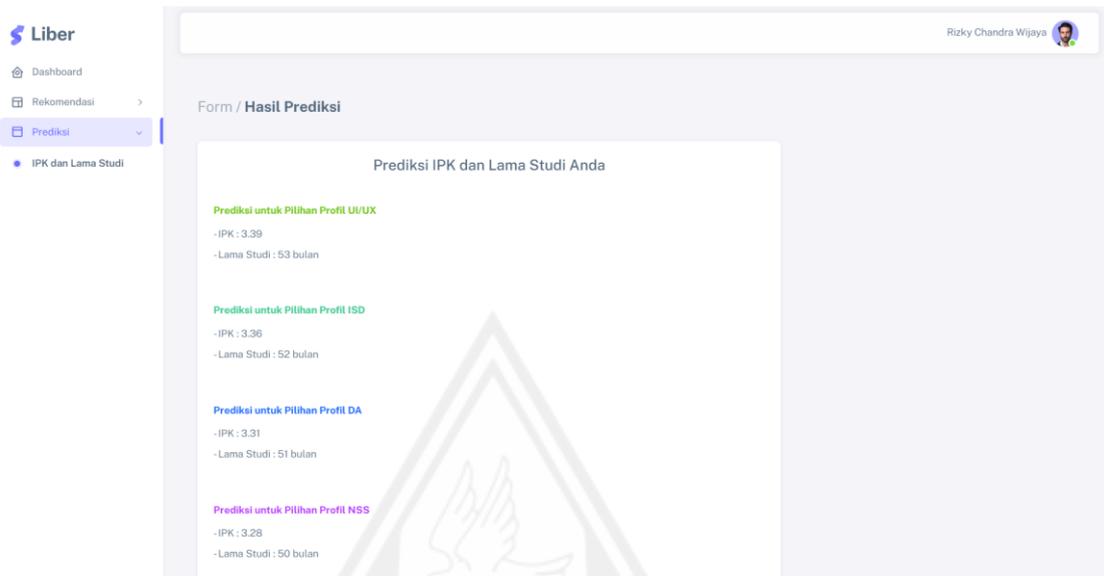
Gambar 4. 7 Halaman *login*



Gambar 4. 8 Tampilan *dashboard* sistem



Gambar 4. 9 Tampilan menu awal sistem prediksi



Gambar 4. 10 Hasil prediksi IPK dan lama studi

1.1.4 Pemodelan

Tahap pemodelan sistem dilakukan sesuai dengan alur proses prediksi yang tertera pada Gambar 3.4. Proses yang terjadi dalam sistem prediksi dapat dilihat lebih lengkap pada Gambar 4.11.

```

1 # get data user login
2 id_mahasiswa_pencari = int(document.getElementById('id_pencari').value)
3 # get data KHS
4 df_khs_user = df_khs[df_khs['id_mahasiswa'] == id_mahasiswa_pencari].copy()
5 # get data mahasiswa
6 df_data_mahasiswa = df_mahasiswa[df_mahasiswa['id_mahasiswa'] == id_mahasiswa_pencari].copy()
7 #transformasi data KHS
8 for k in df_khs_user.index:
9     #membuat dictionary dan memasangkan key id_kelompok_matakuliah beserta nilainya
10     data_dict_khs_user[id_mahasiswa][f'id_kelompok_matkul {k}'] = nilai_data_dict + nilai_khs*sks
11 # melakukan normalisasi data KHS user
12 scaled_prediction = pd.DataFrame(StandardScaler().fit_transform(data_dict_khs_user))
13 #proses training model regresi IPK dengan SVR - RBF
14 regressor_ipk = SVR(kernel = 'rbf')
15 regressor_ipk.fit(X_train_ipk, y_train_ipk)
16 #proses training model regresi lama studi dengan regresi linier multi
17 model = sm.OLS(y_train, X_train).fit()
18 #prediksi IPK untuk tiap profil lulusan dan mengirim value ke halaman web
19 prediksi_ipk_ui_ux.setAttribute("value", round(prediction_ipk_res[0],2))
20 prediksi_ipk_isd = document.getElementById('prediksi_ipk_isd')
21 prediksi_ipk_isd.setAttribute("value", round(prediction_ipk_res[1],2))
22 prediksi_ipk_da = document.getElementById('prediksi_ipk_da')
23 prediksi_ipk_da.setAttribute("value", round(prediction_ipk_res[2],2))
24 prediksi_ipk_nss = document.getElementById('prediksi_ipk_nss')
25 prediksi_ipk_nss.setAttribute("value", round(prediction_ipk_res[3],2))
26 #prediksi lama studi untuk tiap profil lulusan dan mengirim value ke halaman web
27 prediksi_lama_studi_ui_ux = document.getElementById('prediksi_lama_studi_ui_ux')
28 prediksi_lama_studi_ui_ux.setAttribute("value", round(prediction_lama_studi_res[0]))
29 prediksi_lama_studi_isd = document.getElementById('prediksi_lama_studi_isd')
30 prediksi_lama_studi_isd.setAttribute("value", round(prediction_lama_studi_res[1]))
31 prediksi_lama_studi_da = document.getElementById('prediksi_lama_studi_da')
32 prediksi_lama_studi_da.setAttribute("value", round(prediction_lama_studi_res[2]))
33 prediksi_lama_studi_nss = document.getElementById('prediksi_lama_studi_nss')
34 prediksi_lama_studi_nss.setAttribute("value", round(prediction_lama_studi_res[3]))

```

Gambar 4. 11 Pseudocode sistem prediksi

Pada baris 1 sistem akan mengambil data `id_mahasiswa` yang sudah berhasil *login*. Pada baris 4 dan 6 sistem akan mengambil data identitas mahasiswa dan KHS berdasarkan yang `id_mahasiswa` diperoleh. Selanjutnya pada baris 8 dan 10 sistem akan melakukan *looping* untuk mendapatkan seluruh nilai KHS untuk mahasiswa yang bersangkutan. Pada baris ke 12 sistem melakukan normalisasi data KHS dengan menggunakan `StandardScaler()`. Pada baris 14 hingga 17 Sistem melakukan pelatihan model regresi prediksi IPK dan lama studi dengan menggunakan metode SVR (*kernel* RBF). Pada baris ke 19 hingga 34 merupakan tahap terakhir sistem. Tahap terakhir adalah melakukan prediksi model IPK dan lama studi, kemudian hasil tersebut akan dikirim ke halaman hasil prediksi untuk ditampilkan kepada mahasiswa.

1.2 Pengujian dan Analisis

1.2.1 Komparasi Model Regresi IPK dan Lama Studi

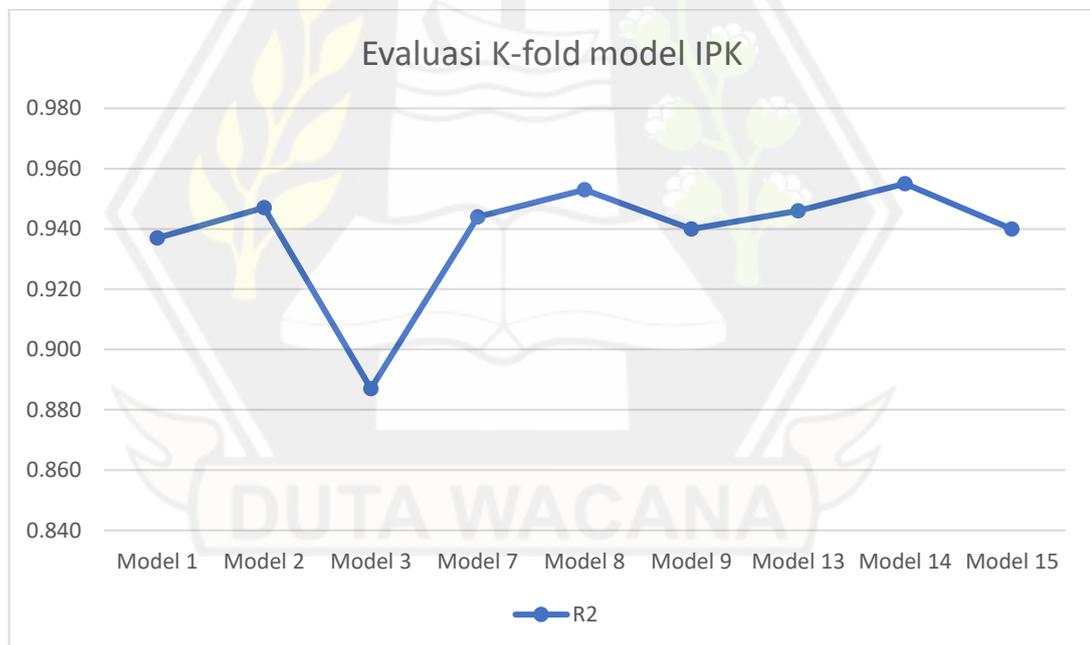
Komparasi model regresi yang tertera pada Tabel 3.1 dilakukan dengan menggunakan *k-fold cross validation* pada masing-masing model. Hasil pengujian dapat dilihat pada Tabel 4.1.

Tabel 4. 1 Hasil k-fold cross validation

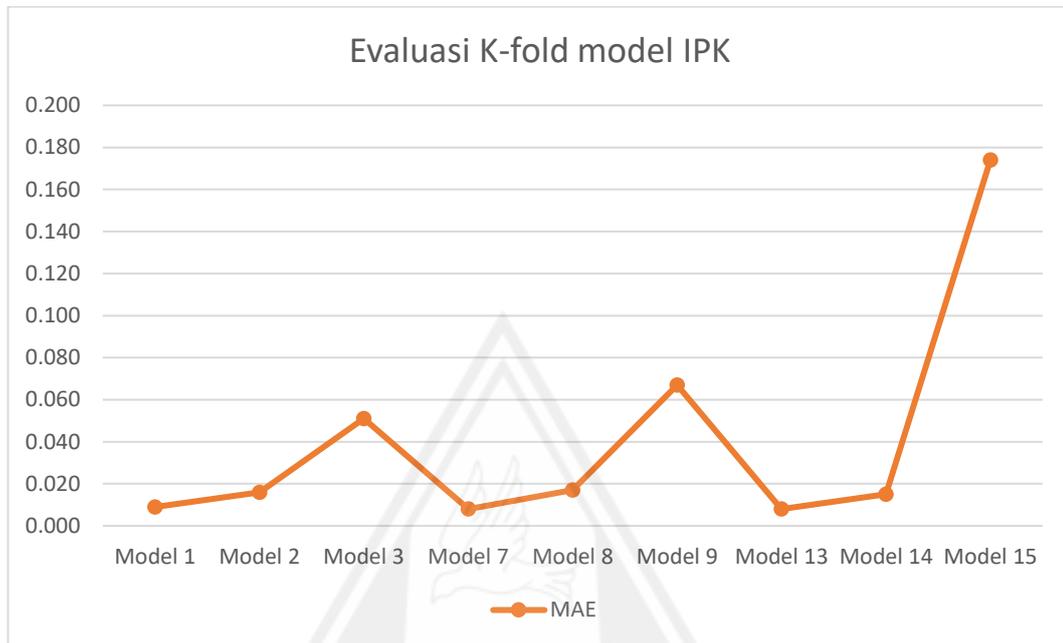
Model Regresi	R ²	MAE	MAE (%)
Model 1 (IPK – Regresi Linier Multi)	0.937	0.009	2.345
Model 2 (IPK – SVR RBF)	0.947	0.016	3.001
Model 3 (IPK – SVR Poly)	0.887	0.051	4.561
Model 4 (Lama Studi – Regresi Linier Multi)	0.364	2.903	14.341
Model 5 (Lama Studi – SVR RBF)	0.464	2.79	12.227
Model 6 (Lama Studi – SVR Poly)	0.357	40.74	17.830
Model 7 (IPK – Regresi Linier Multi)	0.944	0.008	2.266
Model 8 (IPK – SVR RBF)	0.953	0.017	3.090

Tabel 4. 2 Hasil k-fold cross validation (lanjutan)

Model Regresi	R ²	MAE	MAE (%)
Model 9 (IPK – SVR Poly)	0.940	0.067	4.448
Model 10 (Lama Studi – Regresi Linier Multi)	0.398	2.887	14.457
Model 11 (Lama Studi – SVR RBF)	0.560	2.469	11.731
Model 12 (Lama Studi – SVR Poly)	0.439	14.873	15.486
Model 13 (IPK – Regresi Linier Multi)	0.946	0.008	2.265
Model 14 (IPK – SVR RBF)	0.955	0.015	2.963
Model 15 (IPK – SVR Poly)	0.940	0.174	4.739
Model 16 (Lama Studi – Regresi Linier Multi)	0.439	2.743	13.870
Model 17 (Lama Studi – SVR RBF)	0.581	2.600	11.985
Model 18 (Lama Studi – SVR Poly)	0.512	12.491	15.752

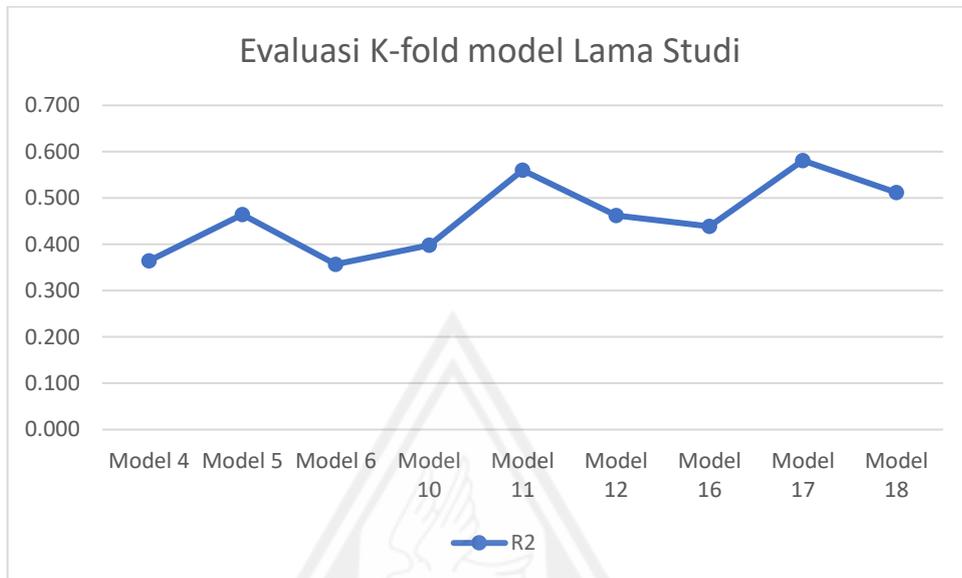


Gambar 4. 12 Visualisasi hasil analisis R² cross validation

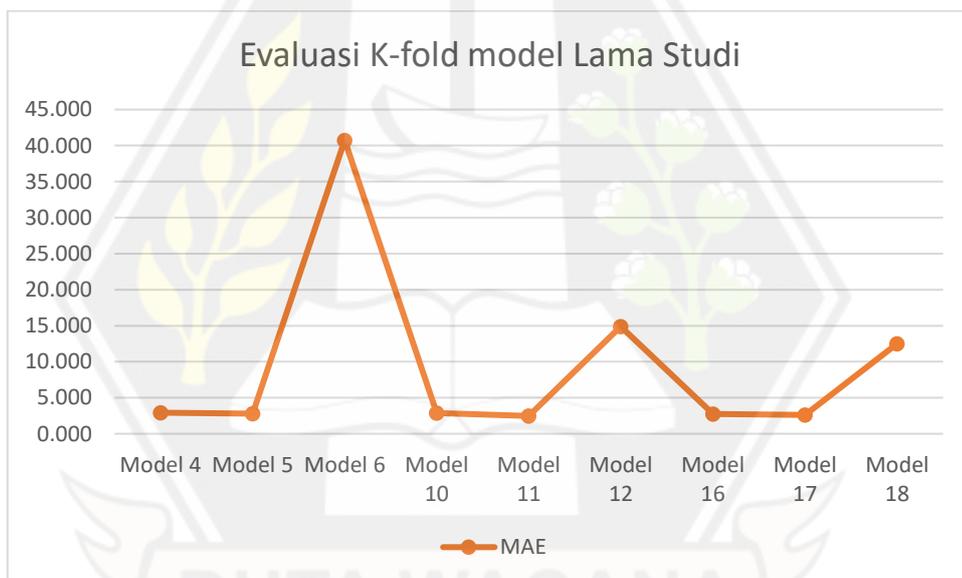


Gambar 4. 13 Visualisasi hasil analisis MAE model IPK

Berdasarkan analisis model regresi IPK pada Gambar 4.12 dan Gambar 4.13, model regresi yang terbaik untuk menjelaskan variabel independen terhadap IPK dengan melihat hasil dari nilai R^2 yang paling tinggi adalah model 14 ($R^2 = 0.955$) dan model 8 ($R^2 = 0.953$). Kemudian dengan melihat hasil MAE pada model 14 (MAE = 0.015) dan model 8 (MAE = 0.017), maka berdasarkan hasil tersebut dapat dipastikan model regresi yang paling baik untuk memprediksi IPK adalah model 14, dengan menggunakan kelompok data profil lulusan, tipe SMA asal, Provinsi SMA asal, jurusan SMA asal, dan mata kuliah wajib berdasarkan relasi antar profil lulusan.



Gambar 4. 14 Visualisasi hasil analisis R² model lama studi



Gambar 4. 15 Visualisasi hasil analisis MAE model lama studi

Berdasarkan Analisis model regresi lama studi pada Gambar 4.14 dan Gambar 4.15, model regresi yang terbaik untuk menjelaskan variabel independen terhadap lama studi adalah model 11 ($R^2 = 0.560$) dan model 17 ($R^2 = 0.581$). Kemudian dengan melihat nilai tingkat eror MAE, maka dapat dipastikan bahwa

model regresi yang terbaik untuk memprediksi lama studi mahasiswa adalah model 17 (MAE = 2.600) merupakan model terbaik.

Kedua model regresi yang terbaik untuk memprediksi IPK dan lama studi menggunakan variabel independen kelompok data latih yang sama. Berdasarkan hasil tersebut, data latih dengan variabel independen `profil`, `jurusan_sma_asal`, `provinsi`, `sekolah_asal`, `tipe_sekolah_asal`, `kelompok_matkul_humaniora`, `kelompok_matkul_ti_wajib`, `kelompok_matkul_relasi_profil_ui/ux-isd-da`, `kelompok_matkul_relasi_profil_isd-da-nss`, `kelompok_matkul_relasi_profil_ui/ux-da-nss`, `kelompok_matkul_relasi_profil_ui/ux`, `kelompok_matkul_relasi_profil_isd`, `kelompok_matkul_relasi_profil_da`, `kelompok_matkul_relasi_profil_nss` dianggap penulis merupakan kelompok data latih paling baik. Metode model regresi pada model 14 dan 17 menggunakan metode SVR dengan *kernel* RBF, oleh karena itu metode SVR dengan kernel RBF merupakan metode regresi yang paling baik dalam memprediksi IPK dan lama studi dengan variabel independen terkait.

1.2.2 Proses Seleksi Fitur (OLS)

Berdasarkan analisis komparasi model, untuk melihat besar korelasi antara variabel independen dan variabel dependen dilakukan analisis model dari hasil regresi menggunakan *Ordinary Least Square* (OLS). Keterangan terkait nama variabel independen yang digunakan dalam model OLS dapat dilihat pada Tabel 4.3.

Tabel 4. 3. Keterangan variabel independen

No	Nama Variabel	Keterangan
1	<code>const</code>	Merupakan konstanta yang dihasilkan oleh model dalam menentukan garis regresi.
2	<code>profil_lulusan</code>	Variabel ini diperoleh dari <i>labeling</i> profil lulusan mahasiswa yang diperoleh dari judul dan topik Skripsi.

Tabel 4. 4. Keterangan variabel independent (lanjutan)

No	Nama Variabel	Keterangan
3	jurusan_sma	Variabel ini diperoleh dari labeling penjurusan yang diambil mahasiswa sewaktu menempuh pendidikan di SMA.
4	tipe_sma	Variabel ini diperoleh dari <i>labeling</i> tipe SMA asal mahasiswa.
5	Kel_matkul_1	Variabel ini diperoleh dari nilai mata kuliah wajib humaniora dari Semester 1 hingga 4.
6	Kel_matkul_2	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap seluruh profil lulusan.
7	Kel_matkul_3	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil UI/UX, ISD, dan DA.
8	Kel_matkul_4	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil ISD, DA, dan NSS.
9	Kel_matkul_5	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil UI/UX, DA, dan NSS.
10	Kel_matkul_6	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil ISD dan NSS.
11	Kel_matkul_7	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil NSS dan DA.

Tabel 4. 5. Keterangan variabel independent (lanjutan)

No	Nama Variabel	Keterangan
12	Kel_matkul_8	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil UI/UX.
13	Kel_matkul_9	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil ISD.
14	Kel_matkul_10	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil DA.
15	Kel_matkul_11	Variabel ini diperoleh dari nilai mata kuliah wajib Prodi dari Semester 1 hingga 4 yang memiliki relasi terhadap profil NSS.

Berdasarkan hasil dari estimator OLS pada Gambar 4.16, dapat dilihat *p-value* untuk variabel `profil_lulusan`, `jurusan_sma`, `provinsi_sma`, `tipe_sma`, `kel_matkul_4`. Jika hasil *p-value* sebuah variabel di atas 0,05, *p-value* menunjukkan seberapa yakin model dalam menghitung kemungkinan terdapat data baru yang memiliki nilai diluar batas *confidence level* 95%. Maka berdasarkan *p-value*, peneliti dapat menolak H_0 yaitu koefisien pada rumus regresi memiliki dampak signifikan untuk variabel yang hasil *p-value* diatas 0,05. Maka penulis mencoba membuat model baru dengan menggunakan fitur atau variabel independent yang tidak digunakan dalam model OLS. Hasil analisis model setelah dilakukan fitur seleksi dapat dilihat pada Gambar 4.17.

OLS Regression Results

```

=====
Dep. Variable:          y      R-squared:                0.948
Model:                 OLS    Adj. R-squared:           0.943
Method:                Least Squares    F-statistic:              220.0
Date:                  Wed, 05 Jul 2023    Prob (F-statistic):      6.16e-108
Time:                  07:06:31    Log-Likelihood:          214.12
No. Observations:     198    AIC:                     -396.2
Df Residuals:         182    BIC:                     -343.6
Df Model:              15
Covariance Type:      nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	3.2787	0.006	534.745	0.000	3.267	3.291
profil_lulusan	0.0029	0.006	0.448	0.655	-0.010	0.016
jurusan_sma	-0.0002	0.009	-0.018	0.985	-0.017	0.017
provinsi_sma	-0.0106	0.007	-1.510	0.133	-0.025	0.003
tipe_sma	-5.131e-05	0.008	-0.006	0.995	-0.016	0.016
kel_matkul_1	0.0401	0.007	5.633	0.000	0.026	0.054
kel_matkul_2	0.0520	0.014	3.842	0.000	0.025	0.079
kel_matkul_3	0.0300	0.009	3.508	0.001	0.013	0.047
kel_matkul_4	0.0123	0.008	1.526	0.129	-0.004	0.028
kel_matkul_5	0.0369	0.008	4.399	0.000	0.020	0.053
kel_matkul_6	0.0327	0.007	4.465	0.000	0.018	0.047
kel_matkul_7	0.0377	0.009	4.211	0.000	0.020	0.055
kel_matkul_8	0.1276	0.013	10.086	0.000	0.103	0.153
kel_matkul_9	0.0337	0.010	3.312	0.001	0.014	0.054
kel_matkul_10	0.0241	0.009	2.716	0.007	0.007	0.042
kel_matkul_11	0.0180	0.008	2.362	0.019	0.003	0.033

```

=====
Omnibus:                6.910    Durbin-Watson:           1.940
Prob(Omnibus):          0.032    Jarque-Bera (JB):        8.309
Skew:                   -0.267    Prob(JB):                 0.0157
Kurtosis:                3.850    Cond. No.:                6.39
=====

```

Gambar 4. 16 Regresi OLS IPK

OLS Regression Results

```

=====
Dep. Variable:          y      R-squared:                0.946
Model:                 OLS    Adj. R-squared:           0.943
Method:                Least Squares    F-statistic:              329.8
Date:                  Wed, 05 Jul 2023    Prob (F-statistic):      4.78e-113
Time:                  07:47:53    Log-Likelihood:          211.52
No. Observations:     198    AIC:                     -401.0
Df Residuals:         187    BIC:                     -364.9
Df Model:              10
Covariance Type:      nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	3.2788	0.006	536.297	0.000	3.267	3.291
kel_matkul_1	0.0428	0.007	6.201	0.000	0.029	0.056
kel_matkul_2	0.0609	0.013	4.781	0.000	0.036	0.086
kel_matkul_3	0.0327	0.008	3.984	0.000	0.017	0.049
kel_matkul_5	0.0337	0.008	4.137	0.000	0.018	0.050
kel_matkul_6	0.0339	0.007	4.669	0.000	0.020	0.048
kel_matkul_7	0.0368	0.009	4.182	0.000	0.019	0.054
kel_matkul_8	0.1319	0.012	10.714	0.000	0.108	0.156
kel_matkul_9	0.0335	0.010	3.364	0.001	0.014	0.053
kel_matkul_10	0.0220	0.009	2.504	0.013	0.005	0.039
kel_matkul_11	0.0174	0.008	2.313	0.022	0.003	0.032

```

=====
Omnibus:                12.292    Durbin-Watson:           1.944
Prob(Omnibus):          0.002    Jarque-Bera (JB):        19.376
Skew:                   -0.353    Prob(JB):                 6.20e-05
Kurtosis:                4.360    Cond. No.:                5.85
=====

```

Gambar 4. 17 Hasil OLS IPK setelah seleksi fitur

Setelah melakukan percobaan ulang untuk membuat model berdasarkan seleksi fitur OLS, ternyata ditemukan bahwa nilai R^2 yang semula sebesar 0.948 turun menjadi 0.946. Dengan hasil OLS ini penulis tidak akan menggunakan fitur seleksi pada model IPK. Maka dari itu fitur seleksi dianggap tidak dapat diterapkan untuk model regresi IPK.

Hasil serupa juga ditemukan dalam percobaan fitur seleksi pada model regresi lama studi, hasil analisis dapat dilihat pada Gambar 4.18 dan Gambar 4.19. Model regresi OLS lama studi memiliki nilai R^2 yang semula sebesar 0.403. Setelah melalui proses seleksi fitur untuk membuang variabel dengan nilai *p-value* di atas 5, nilai R^2 mengalami penurunan menjadi 0.366. Hal tersebut menunjukkan bahwa variabel independen data latih awal merupakan variabel data latih yang paling baik untuk memprediksi lama studi. Maka dari itu proses seleksi fitur pada model lama studi tidak dapat diterapkan pada model regresi lama studi

OLS Regression Results						
=====						
Dep. Variable:	y	R-squared:	0.403			
Model:	OLS	Adj. R-squared:	0.353			
Method:	Least Squares	F-statistic:	8.178			
Date:	Wed, 05 Jul 2023	Prob (F-statistic):	4.92e-14			
Time:	07:06:32	Log-Likelihood:	-359.40			
No. Observations:	198	AIC:	750.8			
Df Residuals:	182	BIC:	803.4			
Df Model:	15					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	8.9848	0.111	80.906	0.000	8.766	9.204
profil_lulusan	-0.1750	0.116	-1.503	0.135	-0.405	0.055
jurusan_sma	0.1059	0.155	0.683	0.496	-0.200	0.412
provinsi_sma	-0.0194	0.128	-0.152	0.879	-0.271	0.232
tipe_sma	-0.0193	0.144	-0.134	0.893	-0.304	0.265
kel_matkul_1	-0.3972	0.129	-3.082	0.002	-0.651	-0.143
kel_matkul_2	0.2637	0.245	1.075	0.284	-0.220	0.748
kel_matkul_3	0.0186	0.155	0.120	0.905	-0.287	0.325
kel_matkul_4	0.2115	0.146	1.448	0.149	-0.077	0.500
kel_matkul_5	-0.1178	0.152	-0.775	0.439	-0.418	0.182
kel_matkul_6	-0.0704	0.133	-0.531	0.596	-0.332	0.191
kel_matkul_7	-0.4111	0.162	-2.536	0.012	-0.731	-0.091
kel_matkul_8	-0.6402	0.229	-2.793	0.006	-1.092	-0.188
kel_matkul_9	-0.1140	0.184	-0.619	0.537	-0.477	0.249
kel_matkul_10	-0.1891	0.161	-1.176	0.241	-0.506	0.128
kel_matkul_11	0.1257	0.138	0.912	0.363	-0.146	0.398

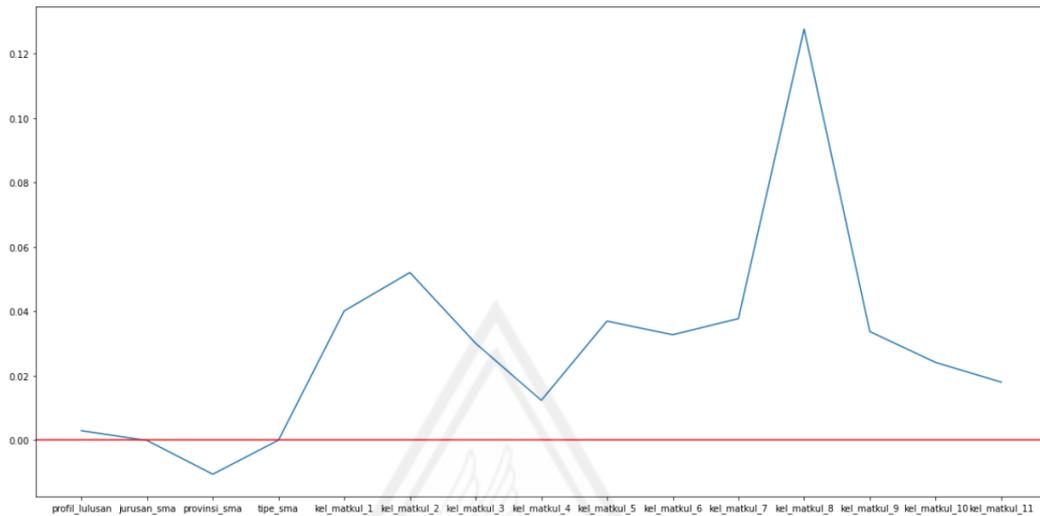
Omnibus:	20.393	Durbin-Watson:	1.909			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.385			
Skew:	0.794	Prob(JB):	8.35e-06			
Kurtosis:	3.558	Cond. No.	6.39			
=====						

Gambar 4. 18 Regresi OLS lama studi

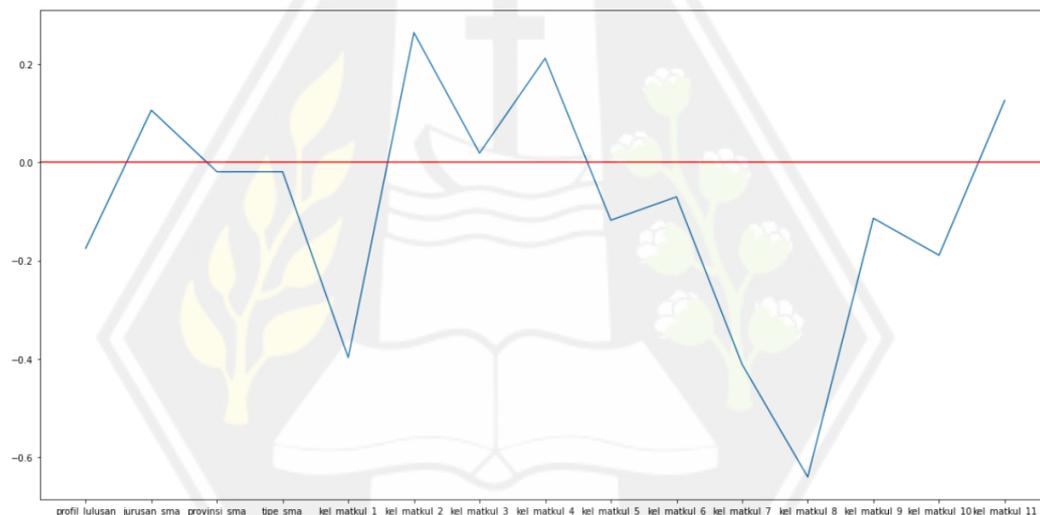
OLS Regression Results						
Dep. Variable:	y	R-squared:	0.366			
Model:	OLS	Adj. R-squared:	0.356			
Method:	Least Squares	F-statistic:	37.34			
Date:	Wed, 05 Jul 2023	Prob (F-statistic):	4.29e-19			
Time:	07:47:54	Log-Likelihood:	-365.29			
No. Observations:	198	AIC:	738.6			
Df Residuals:	194	BIC:	751.7			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	8.9946	0.110	81.715	0.000	8.778	9.212
kel_matkul_1	-0.3502	0.118	-2.968	0.003	-0.583	-0.117
kel_matkul_7	-0.4638	0.153	-3.035	0.003	-0.765	-0.162
kel_matkul_8	-0.5830	0.161	-3.613	0.000	-0.901	-0.265
Omnibus:	23.995	Durbin-Watson:	1.873			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28.700			
Skew:	0.891	Prob(JB):	5.86e-07			
Kurtosis:	3.550	Cond. No.	2.65			

Gambar 4. 19 Hasil OLS lama studi setelah seleksi fitur

Selain digunakan dalam proses seleksi fitur, OLS digunakan peneliti dalam menentukan seberapa besar pengaruh variabel independen terhadap variabel dependen. Hal tersebut dapat ditemukan berdasarkan hasil OLS pada bagian kolom `coef` dan melihat *p-value* kurang dari 0,05. Berdasarkan pengamatan yang dilakukan dengan melihat hasil visualisasi diagram garis pada Gambar 4.20, bahwa variabel independen yang paling berpengaruh terhadap IPK adalah `kel_matkul_8`. Variabel tersebut memiliki dampak positif yang memiliki arti semakin besar nilai variabel tersebut maka semakin besar IPK yang akan diperoleh mahasiswa. Terkait dengan model lama studi dengan menggunakan OLS dapat dilihat pada Gambar 4.21, variabel `kel_matkul_8` merupakan variabel yang paling signifikan dan memiliki dampak negatif terhadap lama studi mahasiswa, yang berarti bahwa semakin besar nilai variabel tersebut, maka semakin kecil lama studi mahasiswa.



Gambar 4. 20 Visualisasi hasil koefisien model regresi IPK

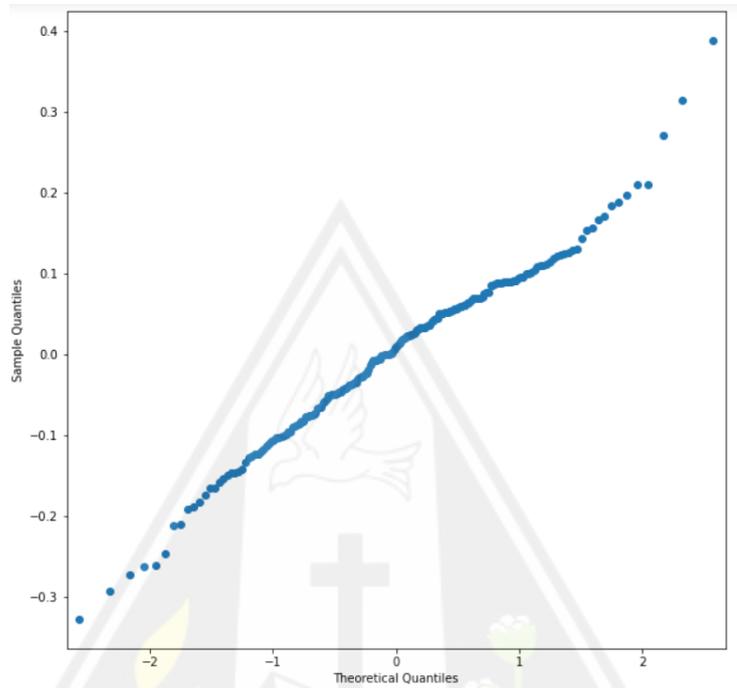


Gambar 4. 21 Visualisasi hasil koefisien model regresi lama studi

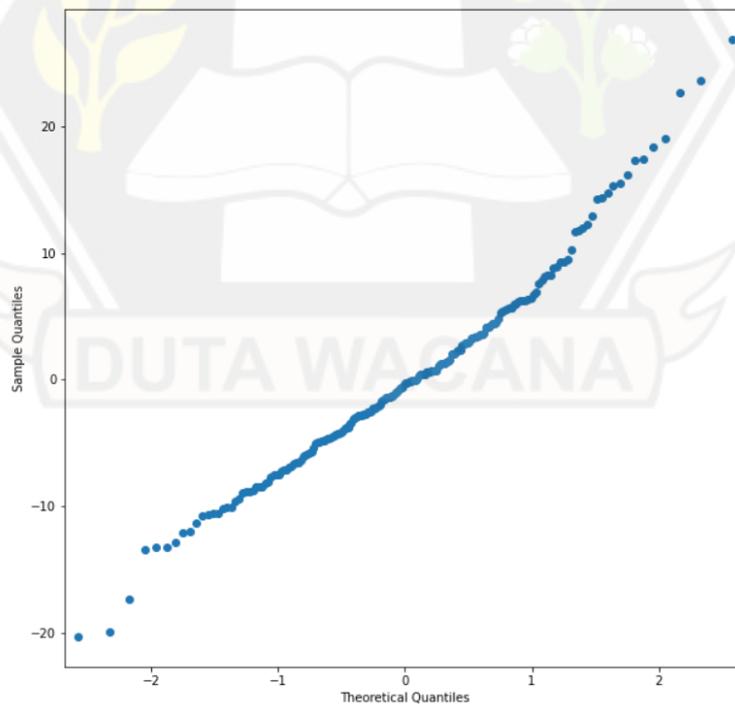
1.2.3 Distribusi Normal Residu

Residu merepresentasikan perbedaan antara hasil dari prediksi model terhadap hasil sebenarnya. Sehingga dapat dikatakan bahwa semakin tinggi nilai residu, maka semakin besar tingkat keraguan yang dimiliki model dalam memprediksi variabel dependen. Selain dari nilai residu, model regresi yang baik harus memiliki distribusi residu yang normal dan memiliki suatu pola sehingga dapat dikatakan bahwa model regresi tersebut memenuhi kondisi

homoscedasticity. Hasil dari *scatter plot* distribusi residu untuk model regresi IPK dan lama studi dapat dilihat pada Gambar 4.22 dan 4.23.



Gambar 4. 22 Distribusi normal residu model regresi IPK



Gambar 4. 23 Distribusi normal residu model prediksi lama studi

Pada Gambar 4.22 dan 4.23 menunjukkan pola distribusi normal yang baik, walaupun ada beberapa titik *outlier* yang terlihat pada masing-masing model. Tingkat residu yang dimiliki model lama studi memiliki tingkat residu tertinggi pada nilai 40, hal tersebut dapat diartikan bahwa variabel independen yang digunakan model belum dapat menjelaskan dengan baik lama studi mahasiswa.

1.2.4 Korelasi Berganda (Multikolinearitas)

Dalam menentukan tingkat estimasi multikolinearitas penulis melakukan uji VIF. Semakin tinggi nilai VIF maka kemungkinan variabel tersebut multikolinearitas terhadap variabel lainnya semakin tinggi. Hasil dari tes VIF dapat dilihat pada Tabel 4.4.

Tabel 4. 6 Nilai VIF

No	Variabel Independen	VIF score
1	Profil lulusan	1.146
2	Jurusan SMA asal	1.893
3	Provinsi SMA asal	1.259
4	Tipe SMA asal	1.747
5	Kelompok mata kuliah Humaniora	1.439
6	Kelompok mata kuliah TI Wajib semua profil	5.055
7	Kelompok mata kuliah TI Wajib profil UI/UX, ISD, DA	2.011
8	Kelompok mata kuliah TI Wajib profil ISD,DA,NSS	1.771
9	Kelompok mata kuliah TI Wajib profil UI/UX, DA, NSS	1.931
10	Kelompok mata kuliah TI Wajib profil ISD, NSS	1.511
11	Kelompok mata kuliah TI Wajib profil NSS,DA	2.208
12	Kelompok mata kuliah TI Wajib UI/UX	4.408
13	Kelompok mata kuliah TI Wajib ISD	2.952
14	Kelompok mata kuliah TI Wajib DA	2.160
15	Kelompok mata kuliah TI Wajib NSS	1.618

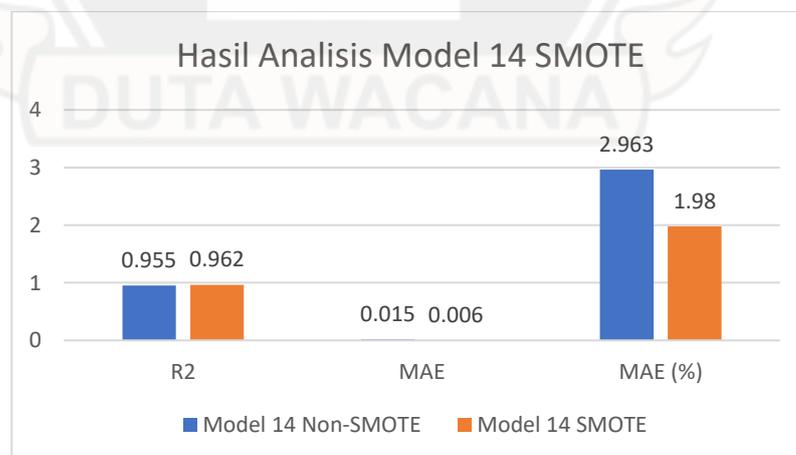
Terdapat 2 variabel independen yang memiliki nilai $VIF \geq 3$, kedua variabel independen tersebut adalah variabel dengan nomor 6 dan 12 pada Tabel 4.2. Berdasarkan hasil uji tersebut menandakan ada kemungkinan variabel tersebut memiliki multikolinearitas terhadap variabel independen lainnya. Walaupun demikian skor tersebut masih ≤ 10 dan tidak terlalu tinggi, sehingga variabel tersebut tetap dapat digunakan dalam perhitungan model regresi.

1.2.5 Pengaruh metode SMOTE untuk melakukan *oversampling* profil

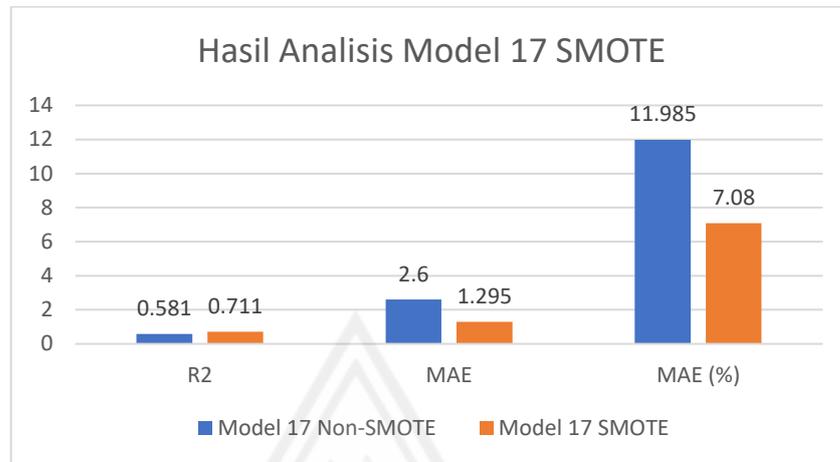
Penggunaan teknik *oversampling* dengan metode SMOTE kepada data latih untuk model regresi IPK dan lama studi memberikan efek positif terhadap model. Dapat dilihat pada Gambar 4.24 Hasil R^2 untuk model regresi IPK dengan menggunakan metode SMOTE membuat peningkatan nilai R^2 sebesar 0,007. Dampak positif juga terlihat pada hasil MAE yang mengalami penurunan sebesar 0,009 dan MAE (%) sebesar 0,983 %.

Hasil yang serupa ditemukan dalam hasil analisis model lama studi pada Gambar 4.25. Hasil R^2 dan *adjusted-R*² untuk model regresi lama studi dengan menggunakan metode SMOTE membuat peningkatan nilai R^2 sebesar 0,13. Dampak positif juga terlihat pada hasil MAE yang mengalami penurunan sebesar 1,305 dan MAE (%) sebesar 4,905 %.

Dikarenakan hasil uji R^2 dan MAE untuk model 14 dan model 17 semakin baik setelah menggunakan metode SMOTE, maka penulis akan menggunakan metode SMOTE dalam mengolah data latih bagi model 14 dan model 17.



Gambar 4. 24 Visualisasi hasil analisis model IPK SMOTE



Gambar 4. 25 Visualisasi hasil analisis model lama studi SMOTE

1.3 Pembuatan Model Regresi Tiap Profil Lulusan

Dikarenakan hasil dari analisis metode SMOTE pada model 14 dan model 17 memberikan dampak positif, penulis melakukan percobaan untuk membuat model regresi untuk tiap profil lulusan. Percobaan ini guna mendapatkan variabel independen yang signifikan untuk tiap profil lulusan dalam memprediksi IPK dan lama studi. Hasil dari percobaan ini dapat dilihat pada Tabel 4.5.

Tabel 4. 7 Hasil percobaan model tiap profil lulusan

Variabel Dependen	Metode	R2	MAE	MAE (%)
Prediksi IPK Profil UI/UX	Regresi Linier Multi	0.569	0.007	1.572
	SVR - RBF	0.537	0.014	2.385
	SVR - Poly	0.540	0.015	2.405
Prediksi IPK Profil ISD	Regresi Linier Multi	0.580	0.003	0.981
	SVR - RBF	0.550	0.009	1.764
	SVR - Poly	0.569	0.005	1.324
Prediksi IPK Profil DA	Regresi Linier Multi	0.594	0.001	0.557
	SVR - RBF	0.573	0.003	1.108
	SVR - Poly	0.572	0.003	1.151

Tabel 4. 5 Hasil percobaan model tiap profil lulusan (lanjutan)

Variabel Dependen	Metode	R2	MAE	MAE (%)
Prediksi IPK Profil NSS	Regresi Linier Multi	0.582	0.003	0.902
	SVR - RBF	0.560	0.004	1.374
	SVR - Poly	0.566	0.008	1.312
Prediksi Lama Studi Profil UI/UX	Regresi Linier Multi	0.274	2.436	9.452
	SVR - RBF	0.119	2.406	9.158
	SVR - Poly	0.158	2.619	9.403
Prediksi Lama Studi Profil ISD	Regresi Linier Multi	0.396	0.990	6.389
	SVR - RBF	0.107	1.599	5.964
	SVR - Poly	0.194	1.760	6.452
Prediksi Lama Studi Profil DA	Regresi Linier Multi	0.540	0.055	1.675
	SVR - RBF	0.455	0.122	2.254
	SVR - Poly	0.493	0.082	1.867
Prediksi Lama Studi Profil NSS	Regresi Linier Multi	0.499	0.390	4.205
	SVR - RBF	0.334	0.849	6.794
	SVR - Poly	0.459	0.809	5.133

Berdasarkan hasil uji dari pembuatan model regresi bagi tiap profil lulusan, hasil yang diperoleh kurang baik. Terlihat untuk seluruh model regresi tidak ada yang memiliki nilai $R^2 \geq 0,6$. Hasil pengujian nilai R^2 yang rendah menunjukkan bahwa model tidak dapat menjelaskan dengan baik korelasi antara variabel independen terhadap variabel dependen. Terdapat beberapa faktor yang mempengaruhi hasil analisis model pada percobaan ini yaitu, jumlah data yang sedikit bagi tiap model ($n = 162$) dan penggunaan metode *oversampling* yang membuat tidak ditemukan keunikan antar data yang dibuat secara sintetis oleh metode SMOTE. Dikarenakan hasil pengujian yang kurang baik tersebut, penulis tidak menggunakan model regresi bagi tiap profil lulusan.

1.4 Pembahasan

Berdasarkan hasil dari pengujian dan analisis yang sudah dilakukan, metode yang paling baik untuk memprediksi IPK dan lama studi mahasiswa adalah metode *Super Vector Regression* dengan menggunakan *kernel* RBF. Hal tersebut dapat disebabkan oleh garis prediksi pada *kernel* RBF tidak terpaku pada garis lurus atau linier melainkan berbentuk *radial* atau melengkung. Karena kelonggaran tersebut mengakibatkan MAE metode SVR lebih kecil jika dibandingkan regresi linier multi. Pemilihan metode SVR ini menunjukkan bahwa relasi variabel independen dan variabel dependen pada data latih memiliki relasi non linier.

Dalam melakukan uji *k-fold cross validation* ditemukan satu kelemahan dalam model prediksi lama studi yaitu model tidak dapat memprediksi data mahasiswa yang memiliki masa studi di atas 65 bulan atau masa studi di atas 10 Semester. Dengan kata lain model tidak dapat menangani *input* data yang masuk ke dalam kategori *outlier*.

Berdasarkan hasil yang cukup memuaskan dari model prediksi IPK penulis dapat menarik sebuah kesimpulan terkait variabel independen apa saja yang paling berdampak baik itu positif maupun negatif ke dalam hasil prediksi IPK. Berdasarkan *p-value* dan koefisien yang dihasilkan metode estimator OLS, variabel `kel_matkul_8` merupakan kontributor terbesar dalam mempengaruhi hasil IPK. Dengan menganggap variabel lain bernilai konstan maka untuk nilai variabel `kel_matkul_8` akan mempengaruhi penambahan IPK sebesar +0.1143 atau +3%. Mata kuliah yang termasuk ke dalam `kel_matkul_8` adalah mata kuliah wajib Prodi Informatika UKDW yang memiliki relasi khusus terhadap profil UI/UX yaitu Interaksi Manusia dan Komputer (IMK), Pemrograman Web (ProgWeb), Praktikum Pemrograman Web (PrakProgWeb), Rekayasa Perangkat Lunak Berbasis Objek (RPLBO), dan Praktikum Rekayasa Perangkat Lunak Berbasis Objek (PrakRPLBO).

Berbeda dengan hasil analisis model prediksi IPK, model prediksi lama studi memiliki hasil yang kurang baik dengan ditujukannya nilai R^2 yang tidak

terlalu baik, tingkat error dari MAE yang mencapai 1.295, dan kekurangan model prediksi yang tidak dapat memprediksi lama studi mahasiswa di atas 65 bulan atau di atas 10 semester. Oleh sebab itu peneliti tidak dapat dengan yakin untuk memberikan rekomendasi dan nilai mata kuliah signifikan yang mempengaruhi masa studi mahasiswa.



BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Metode regresi linier multi dan *Super Vector Regression* (SVR) telah berhasil diimplementasikan untuk membuat sistem prediksi IPK dan lama studi mahasiswa Informatika UKDW, walaupun hasil yang didapatkan masih jauh dari sempurna. Berdasarkan hasil pengujian dan analisis yang dilakukan penulis, dapat diambil kesimpulan sebagai berikut:

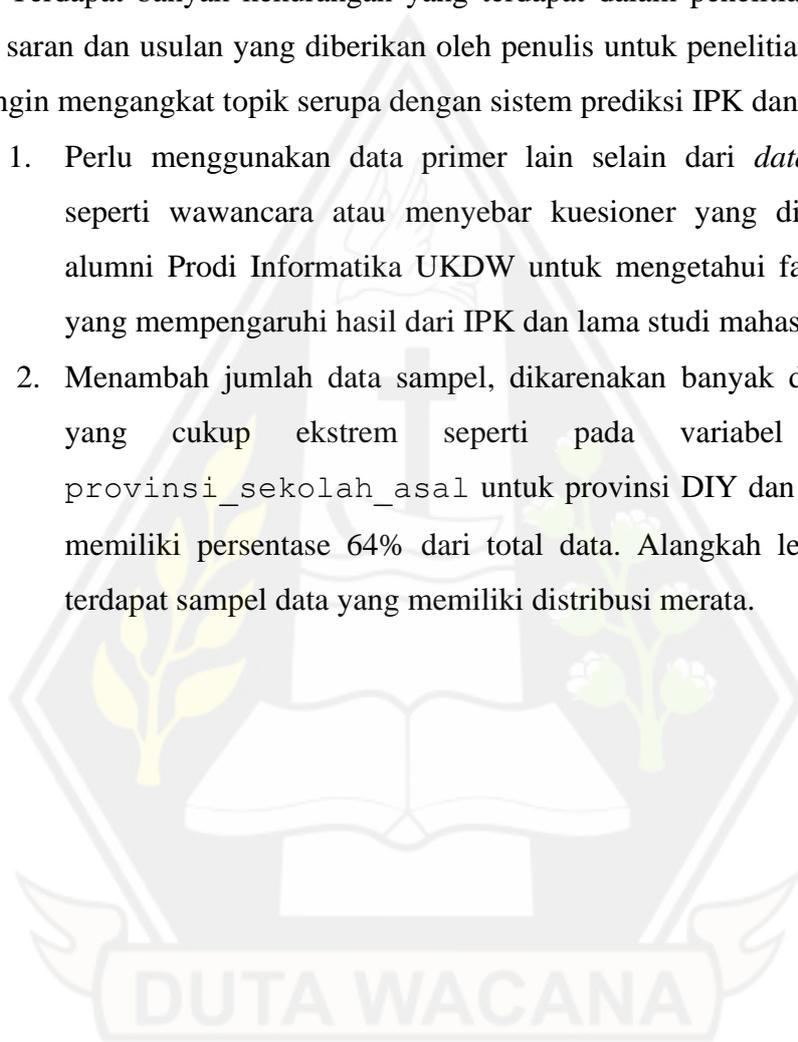
1. Pada penelitian ini model regresi yang terbaik untuk memprediksi IPK adalah metode SVR (*kernel* RBF) dan dengan menggunakan teknik SMOTE yaitu model 14. Model ini memiliki nilai R^2 sebesar 0.962, nilai MAE sebesar 0.006, dan nilai MAE (%) sebesar 1.98 %
2. Pada penelitian ini model regresi yang terbaik untuk memprediksi lama studi adalah metode SVR (*kernel* RBF) dan dengan menggunakan teknik SMOTE yaitu model 17. Model ini memiliki nilai R^2 sebesar 0.711, nilai MAE sebesar 1.295, dan nilai MAE (%) sebesar 7.08 %.
3. Berdasarkan pengujian nilai koefisien regresi dengan metode OLS dapat disimpulkan bahwa variabel independen yang paling berpengaruh terhadap prediksi IPK adalah variabel `kel_matkul_8`. Dengan demikian nilai mata kuliah Interaksi Manusia dan Komputer (IMK), Pemrograman Web (ProgWeb), Praktikum Pemrograman Web (PrakProgWeb), Rekayasa Perangkat Lunak Berbasis Objek (RPLBO), dan Praktikum Rekayasa Perangkat Lunak Berbasis Objek (PrakRPLBO) merupakan mata kuliah yang paling penting untuk mendapatkan IPK yang maksimal.
4. Berdasarkan pengujian terhadap model regresi lama studi, penulis tidak dapat menganalisis korelasi antar data dan menentukan faktor

apa yang terpenting dalam data. Hal tersebut dikarenakan hasil dari nilai R^2 yang cukup rendah dan tingkat eror MAE yang cukup tinggi.

5.2 Saran

Terdapat banyak kekurangan yang terdapat dalam penelitian ini, berikut adalah saran dan usulan yang diberikan oleh penulis untuk penelitian kedepannya yang ingin mengangkat topik serupa dengan sistem prediksi IPK dan lama studi:

1. Perlu menggunakan data primer lain selain dari *data warehouse*, seperti wawancara atau menyebarkan kuesioner yang ditujukan pada alumni Prodi Informatika UKDW untuk mengetahui faktor apa saja yang mempengaruhi hasil dari IPK dan lama studi mahasiswa.
2. Menambah jumlah data sampel, dikarenakan banyak distribusi data yang cukup ekstrem seperti pada variabel independen *provinsi_sekolah_asal* untuk provinsi DIY dan Jawa Tengah memiliki persentase 64% dari total data. Alangkah lebih baik jika terdapat sampel data yang memiliki distribusi merata.



DAFTAR PUSTAKA

- Amalia, R. (2020). Penerapan Data Mining Untuk Memprediksi Hasil Kelulusan Siswa Menggunakan Metode Naive Bayes. *JUISI*, 33-42.
- Bangdiwala, S. I. (2018). Regression: simple linear. *International Journal of Injury Control and Safety*, 113-115.
doi:<https://doi.org/10.1080/17457300.2018.1426702>
- Dewi, K. E., & Widiastuti, N. I. (2020). Support Vector Regression for GPA Prediction. *IOP Conf. Series: Materials Science and Engineering*, 879.
doi:10.1088/1757-899X/879/1/012112
- Ghosal, S., Sengupta, S., Majumder, M., & Sinha, B. (2020, March 27). Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 311-315.
doi:<https://doi.org/10.1016/j.dsx.2020.03.017>
- Guido, S., & Muller, A. (2016). *Introduction to Machine Learning*. United States of America: O'Reilly Media, Inc.
- Kurniatullah, B. D., & Pramudi, Y. T. (2017). Estimation of Students' Graduation Using Multiple Linear Regression Method. *Journal of Applied Intelligent System*, Vol2, hal 29-36.
- Kurniawan, L. (2022). *Sistem Rekomendasi Topik Skripsi Mahasiswa Informatika UKDW*. Yogyakarta: Prodi Informatika UKDW.
- Lindstedt, W. I. (2015). *Data Architecture: A Primer For The Data Scientist*. Waltham, USA: Elsevier Inc.
- Lusa, R. B. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 1471-2105. doi:10.1186/1471-2105-14-106
- Menristekdikti. (2015). *Permenristek no. 44/2015 tentang standar nasional perguruan tinggi*. Indonesia: Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia.
- Paoella, M. S. (2019). *Linear Models and Time Series Analysis*. Oxford, UK: Wiley.

- Surya, B. (2022). *Sistem Rekomendasi Profil Mahasiswa Informatika UKDW*. Yogyakarta: Prodi Informatika UKDW.
- Tim Penyusun Kurikulum 2013 Prodi Informatika UKDW. (2013). *Panduan Akademik Kurikulum 2011 Revisi 2013 Program Studi Informatika UKDW*. Universitas Kristen Duta Wacana.
- Tim Penyusun Kurikulum 2017 Prodi Informatika UKDW. (2017). *Panduan Akademik Kurikulum 2017 Program Studi Informatika UKDW*. Universitas Kristen Duta Wacana.
- Tim Penyusun Kurikulum 2019 Prodi Informatika UKDW. (2019). *Panduan Akademik Kurikulum 2017 Revisi 2019 Program Studi Informatika UKDW*. Universitas Kristen Duta Wacana.
- Tim Penyusun Kurikulum 2021 Prodi Informatika UKDW. (2021). *Panduan Akademik Kurikulum 2021 Program Studi Informatika UKDW*. Universitas Kristen Duta Wacana.
- Uyanik, G. K., & Guler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia*, 234-240.

