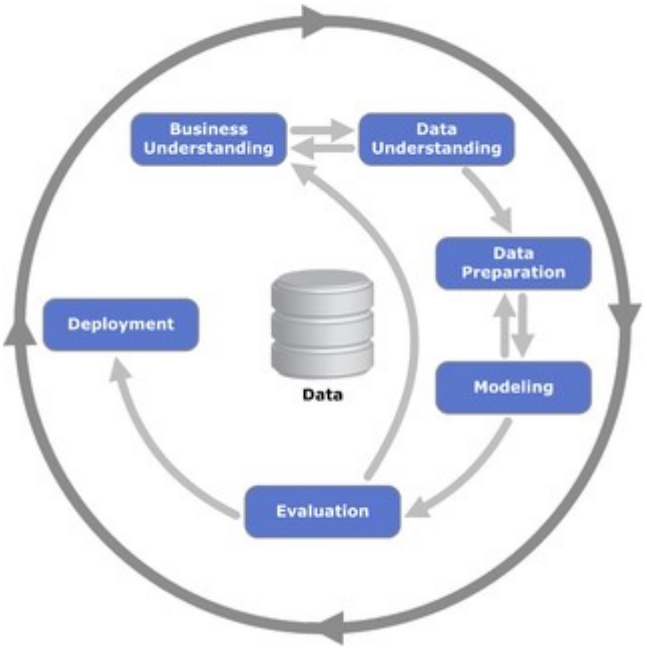
BAB III
METODE PENELITIAN

3.1. Metode Penelitian

Penelitian ini menggunakan pendekatan CRISP-DM (*Cross-Industry Standard Process for Data Mining*) sebagai metodologi utama dalam menganalisis pengaruh kualitas lingkungan belajar terhadap prestasi akademik mahasiswa dengan *Big Data Analytics*. CRISP-DM merupakan metode standar dalam proses data mining[17] yang terdiri dari enam tahapan utama: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation*, dan *Deployment*.

3.1.1 Metode



Gambar 3.1 CRISP-DM

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) merupakan suatu kerangka kerja yang bersifat terbuka dan banyak digunakan dalam proyek-proyek data mining di berbagai sektor industri. Metodologi ini dikembangkan untuk memberikan panduan sistematis dan terstruktur dalam mengelola proyek analisis data, mulai dari pemahaman masalah bisnis hingga

penerapan dan evaluasi model. CRISP-DM bersifat iteratif, artinya setiap tahapan dapat kembali dikunjungi dan diperbaiki seiring perkembangan pemahaman dan hasil yang diperoleh.

3.2 Tahapan Penelitian

3.2.1 *Business Understanding*

Pada tahap ini, penelitian berfokus pada pemahaman konteks dan tujuan penelitian, yaitu menganalisis pengaruh kualitas lingkungan belajar terhadap prestasi akademik mahasiswa. Identifikasi variabel yang berpotensi mempengaruhi prestasi akademik dilakukan berdasarkan studi literatur dan faktor lingkungan belajar, baik internal maupun eksternal. Variabel dependen dalam penelitian ini adalah prestasi akademik mahasiswa, yang dapat diukur melalui indikator seperti Indeks Prestasi Kumulatif (IPK), nilai akhir mata kuliah, atau kategori prestasi (tinggi/sedang/rendah).

Sementara itu, variabel independennya mencakup aspek lingkungan belajar. Lingkungan internal meliputi kenyamanan ruang kelas, ketersediaan fasilitas pembelajaran, akses ke perpustakaan, kualitas internet kampus, interaksi dengan dosen, dan dukungan akademik dari institusi. Adapun lingkungan eksternal mencakup kondisi tempat tinggal, dukungan keluarga, waktu belajar di luar kampus, gangguan sosial di sekitar lingkungan tinggal, akses internet pribadi, serta kepemilikan perangkat belajar seperti laptop atau smartphone. Variabel-variabel tersebut digunakan untuk melihat sejauh mana kualitas lingkungan belajar dapat mempengaruhi pencapaian akademik mahasiswa.

3.2.2 *Data Understanding*

Tahap ini mencakup eksplorasi awal terhadap dua dataset utama yang digunakan dalam penelitian, yaitu data survei lingkungan belajar mahasiswa dan data akademik yang merepresentasikan prestasi akademik melalui indikator seperti IPK, kehadiran, dan jumlah mata kuliah yang diambil. Proses ini diawali dengan memahami struktur data, termasuk banyaknya fitur, jenis variabel

(numerik, kategorikal, maupun ordinal), serta distribusi nilai dalam masing-masing fitur. Langkah ini penting untuk mendeteksi adanya ketidaksesuaian format data, nilai ekstrem (outlier), atau pola distribusi yang tidak normal, yang dapat mempengaruhi kualitas hasil analisis dan pemodelan.

Selain itu, eksplorasi juga melibatkan pengecekan terhadap nilai yang hilang (missing values), identifikasi korelasi antar variabel, serta interpretasi awal terhadap kecenderungan data. Melalui pemahaman yang mendalam ini, peneliti dapat menentukan teknik pra proses yang sesuai, seperti pengisian data hilang, transformasi variabel, atau teknik encoding untuk data kategorikal. Dengan demikian, tahap data understanding berperan sebagai fondasi yang krusial sebelum masuk ke proses persiapan data dan pemodelan, memastikan bahwa data yang digunakan telah siap secara teknis dan konseptual untuk dianalisis lebih lanjut.

	NIM	ANGKATAN	SEMESTER	KODE_MK	NAMA_MK	NILAI_ANGKA	NILAI_APLHABET
0	10109298971	2010	1011	EM100	EM100 Dasar-dasar Bisnis	57.0	C
1	10109298971	2010	1011	EM180	EM180 Matematika Bisnis	70.0	B
2	10109298971	2010	1011	TI100	TI100 Algoritma dan Pemrograman	57.0	C
3	10109298971	2010	1011	TI101	TI101 Matematika Diskrit	59.0	C
4	10109298971	2010	1011	TI110	TI110 Pengantar Teknologi Multimedia	74.0	B
...	...	...	...	...	...	...	...
16951	14109299053	2014	1411	EM190	Pengantar Manajemen & Bisnis	76.0	B+
16952	14109299053	2014	1411	IF110	Pengantar Teknologi Multimedia	98.0	A
16953	14109299053	2014	1411	IF140	Logika Pemrograman	73.0	B
16954	14109299053	2014	1411	IS100	Sistem Informasi Dalam Manajemen	81.0	A-
16955	14109299053	2014	1411	IS110	Matematika Bi	NaN	NaN

16956 rows × 7 columns

Gambar 3.2 dataset

Gambar di atas menampilkan cuplikan dari dataset nilai akademik mahasiswa yang digunakan dalam penelitian ini. Dataset tersebut terdiri dari 7 kolom, yaitu:

1. NIM: Nomor Induk Mahasiswa, berfungsi sebagai identitas unik setiap mahasiswa.
2. ANGKATAN: Tahun masuk mahasiswa ke perguruan tinggi.

3. SEMESTER: Kode semester saat mata kuliah tersebut diambil.
4. KODE_MK dan NAMA_MK: Mewakili kode dan nama lengkap dari mata kuliah.
5. NILAI_ANGKA: Nilai akhir mata kuliah dalam bentuk numerik.
6. NILAI_APLHABET: Nilai akhir mata kuliah dalam bentuk huruf (misalnya A, B+, C, dsb).

Dari tampilan tersebut, terlihat bahwa terdapat nilai kosong atau *missing value* pada beberapa entri, khususnya pada kolom NILAI_ANGKA dan NILAI_ALPHABET. Hal ini menunjukkan bahwa sebelum data digunakan dalam pemodelan, perlu dilakukan proses pembersihan dan penanganan nilai yang hilang agar tidak mengganggu hasil analisis. Selain itu, struktur data ini memperlihatkan bahwa mahasiswa dapat memiliki banyak baris data sesuai jumlah mata kuliah yang mereka ambil selama studi, sehingga perlu dilakukan agregasi terlebih dahulu untuk memperoleh indikator seperti rata-rata nilai, kehadiran, atau jumlah mata kuliah per mahasiswa.

3.2.3 *Data Preparation*

3.2.3.1 Teknik Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan melalui dua jalur utama, yaitu data akademik mahasiswa dan data survei lingkungan belajar. Data akademik diperoleh dari sistem informasi akademik kampus yang mencatat informasi seperti nilai akhir per mata kuliah, kehadiran, dan jumlah mata kuliah yang diambil oleh setiap mahasiswa. Data ini bersifat numerik dan objektif, sehingga sangat ber dalam analisis kuantitatif untuk memprediksi prestasi akademik, khususnya IPK. Sementara itu, data survei dikumpulkan melalui penyebaran kuesioner kepada mahasiswa aktif dari berbagai jurusan dan angkatan. Survei ini bertujuan untuk menangkap persepsi mahasiswa terhadap

aspek-aspek lingkungan belajar, seperti dukungan keluarga, fasilitas kampus, manajemen waktu, dan kesehatan.

Kedua jenis data tersebut kemudian disiapkan untuk dianalisis secara terpisah. Data akademik digunakan dalam proses supervised learning untuk membangun model prediksi IPK, sedangkan data survei dianalisis dengan pendekatan unsupervised learning untuk pengelompokan mahasiswa berdasarkan karakteristik lingkungan belajar mereka. Dengan menggabungkan dua pendekatan analisis ini, penelitian dapat mengidentifikasi pola umum yang mempengaruhi prestasi akademik serta mengungkapkan faktor-faktor lingkungan yang paling dominan dalam membentuk pengalaman belajar mahasiswa. Pendekatan ini memberikan dasar yang kuat untuk menyusun kebijakan pendidikan berbasis data yang lebih terarah dan berdampak langsung pada peningkatan kualitas pembelajaran.

A.2 Dataset Historis Mahasiswa (2010 - 2024)

Data historis mahasiswa diperoleh dari dataset lama yang didapatkan dari data internal kampus melalui tim Laboratorium FTI sejak tahun 2010 hingga tahun 2024. Dataset ini mencakup informasi akademik mahasiswa, seperti:

1. Indeks Prestasi Semester (IPS) dan Indeks Prestasi Kumulatif (IPK)
2. Data kehadiran perkuliahan
3. Data aktivitas akademik (penggunaan e-learning, partisipasi diskusi, dsb.)
4. Kondisi sosial ekonomi mahasiswa (jika tersedia)

Dataset ini digunakan untuk memahami pola dan tren prestasi akademik mahasiswa dalam jangka waktu yang lebih panjang serta bagaimana faktor lingkungan belajar dapat mempengaruhi capaian akademik mereka.

3.3.2 Survei Mahasiswa Saat Ini

Selain data historis, penelitian ini juga mengumpulkan data melalui survey yang ditujukan kepada mahasiswa saat ini melalui Google form (

https://forms.gle/F6wLvDtkTqej42pLA) periode pengisian dari tanggal 10 Mei 2025 sampai dengan tanggal 30 Juni 2025 untuk mendapatkan perspektif langsung mengenai kualitas lingkungan belajar mereka. Survey ini mencakup beberapa aspek utama, antara lain:

1. Kondisi lingkungan belajar di rumah dan kampus
2. Ketersediaan fasilitas akademik (laboratorium, perpustakaan, internet, dsb.)
3. Metode pembelajaran dan tingkat keterlibatan dalam perkuliahan
4. Faktor eksternal seperti kondisi ekonomi dan sosial
5. Tingkat kepuasan mahasiswa terhadap lingkungan belajar mereka

Survei ini dilakukan menggunakan kuesioner daring yang disebarkan kepada mahasiswa aktif dari berbagai program studi. Data yang dikumpulkan melalui survei ini akan digunakan untuk memperkaya analisis terhadap pengaruh lingkungan belajar terhadap prestasi akademik mahasiswa saat ini.

Sebelum pemodelan, data harus dipersiapkan dengan proses seperti:

1. Pembersihan Data: Mengatasi nilai yang hilang, duplikasi, atau inkonsistensi dalam dataset.

```
df.fillna(0, inplace=True)
df
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16956 entries, 0 to 16955
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   NIM             16956 non-null  int64
 1   ANGKATAN        16956 non-null  int64
 2   SEMESTER        16956 non-null  int64
 3   KODE_MK         16956 non-null  object
 4   NAMA_MK         16956 non-null  object
 5   NILAI_ANGKA     16956 non-null  float64
 6   NILAI_APLHABET  16956 non-null  object
dtypes: float64(1), int64(3), object(3)
memory usage: 927.4+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292 entries, 0 to 291
Data columns (total 44 columns):
 #   Column                                                                                                  Non-Null Count  Dtype
---  ------                                                                                                  --------------  -----
 0   Timestamp                                                                                               292 non-null    object
 1   Nama                                                                                                    292 non-null    object
 2   Usia                                                                                                    292 non-null    object
 3   Jenis Kelamin                                                                                           292 non-null    object
 4   Program Studi                                                                                           292 non-null    object
 5   Semester                                                                                                292 non-null    object
 6   Seberapa sering Anda mendapatkan dukungan dari keluarga dalam hal akademik?                             292 non-null    int64
 7   Bagaimana kondisi ekonomi keluarga Anda mempengaruhi prestasi akademik Anda?                            292 non-null    int64
 8   Apakah tingkat pendidikan orang tua Anda mempengaruhi cara Anda belajar?                                292 non-null    int64
 9   Berapa jumlah anggota keluarga yang tinggal serumah dengan Anda?                                        292 non-null    object
 10  Apakah Anda tinggal bersama kedua orang tua? Ya / Tidak                                                 292 non-null    object
 11  Seberapa sering Anda berdiskusi tentang masalah akademik dengan orang tua/wali?                         292 non-null    int64
 12  Apakah Anda mendapatkan dukungan finansial yang cukup dari keluarga untuk keperluan kuliah?            292 non-null    int64
 13  Apakah keluarga mendukung Anda berkuliah di jurusan yang saat ini Anda jalani?                          292 non-null    object
 14  Seberapa sering Anda berpartisipasi dalam kegiatan organisasi atau ekstrakurikuler?                    292 non-null    int64
 15  Seberapa besar pengaruh teman-teman Anda terhadap prestasi akademik Anda?                               292 non-null    int64
 16  Apakah Anda merasa tertekan oleh ekspektasi sosial atau kelompok?                                       292 non-null    int64
 17  Berapa banyak waktu yang Anda habiskan untuk berinteraksi dengan teman sebaya di luar kegiatan kuliah? 292 non-null    int64
 18  Apakah Anda merasa memiliki kelompok pertemanan yang mendukung kegiatan akademik Anda?                  292 non-null    int64
 19  Seberapa sering Anda terlibat dalam kegiatan organisasi di luar kampus?                                 292 non-null    int64
 20  Apakah Anda merasa memiliki cukup waktu luang untuk bersosialisasi dan beristirahat?                    292 non-null    int64
 21  Seberapa lengkap fasilitas pembelajaran yang tersedia di kampus Anda?                                   292 non-null    int64
 22  Seberapa sering Anda menggunakan fasilitas pembelajaran di kampus (perpustakaan, laboratorium, dll.)?  292 non-null    int64
 23  Seberapa sering Anda mengalami gangguan saat belajar (contoh: kebisingan, kurang nyaman)?              292 non-null    int64
 24  Apakah Anda bekerja sambil kuliah?                                                                      292 non-null    object
 25  Jika ya, seberapa besar pengaruh pekerjaan Anda terhadap prestasi akademik Anda?                        199 non-null    float64
 26  Jika ya, apakah Anda bekerja untuk membiayai kuliah Anda?                                               202 non-null    object
 27  Jika tidak, apakah anda mendapatkan dukungan finansial penuh dari keluarga untuk keperluan kuliah?     266 non-null    object
 28  Jika tidak, apakah anda sudah merasakan cukup dengan dukungan finansial yang Anda dapatkan?             260 non-null    float64
 29  Apakah Anda sering mengalami masalah kesehatan yang mengganggu proses belajar?                          292 non-null    int64
 30  Apakah Anda memiliki akses yang baik terhadap layanan kesehatan?                                        292 non-null    int64
 31  Apakah Anda suka berolahraga?                                                                           292 non-null    object
 32  Seberapa sering Anda berolahraga?                                                                       292 non-null    int64
 33  Apakah Anda memiliki kegiatan di luar kuliah yang mempengaruhi waktu belajar Anda?                      292 non-null    int64
 34  Seberapa baik Anda mengelola waktu antara kuliah, pekerjaan, dan kegiatan lain?                         292 non-null    int64
 35  Apakah Anda pernah tidak hadir kuliah karena sulit membagi waktu antara kuliah dengan kegiatan lain?    292 non-null    int64
 36  Seberapa sering Anda mendapatkan bimbingan akademik dari dosen?                                         292 non-null    int64
 37  Seberapa besar pengaruh bimbingan akademik dari dosen terhadap prestasi akademik Anda?                  292 non-null    int64
 38  Seberapa efektif menurut Anda fasilitas belajar yang tersedia di kampus (perpustakaan, laboratorium, dll.)? 292 non-null    int64
 39  Apakah Anda merasa puas dengan kualitas pengajaran dosen di jurusan Anda?                               292 non-null    int64
 40   Apakah Anda merasa beban tugas kuliah yang diberikan terlalu berat?                                    292 non-null    int64
 41  Apakah Anda merasa adanya persaingan yang tidak sehat di antara sesama mahasiswa?                       292 non-null    int64
 42  Apakah ada faktor eksternal lain yang menurut Anda mempengaruhi prestasi akademik Anda?                 292 non-null    object
 43  Saran atau masukan untuk meningkatkan prestasi akademik mahasiswa di kampus Anda.                       246 non-null    object
```

Gambar 3.3 Info Dataset

2. Transformasi Data: Menyesuaikan format data agar sesuai dengan kebutuhan analisis. Dengan memilih hanya kolom yang relevan, seperti NIM, nama mata kuliah, dan nilai angka. Dengan cara ini, data menjadi lebih ringkas dan terfokus, sehingga memudahkan proses analisis dan mengurangi potensi gangguan dari informasi yang tidak diperlukan.

```
df = df[['NIM', 'NAMA_MK', 'NILAI_ANGKA']]
```

Gambar 3.4 Transformasi Data

3. Feature Engineering: Memilih dan mengkombinasikan variabel yang relevan untuk meningkatkan akurasi model analisis. Dengan memproses data nilai mahasiswa menggunakan teknik pivot. Data awal

dikelompokkan berdasarkan NIM dan nama mata kuliah, lalu hanya entri pertama yang dipilih jika terjadi duplikasi. Setelah itu, data diubah menjadi format pivot sehingga setiap baris mewakili satu mahasiswa dan kolom-kolomnya berisi nilai dari masing-masing mata kuliah. Hasilnya, dataset menjadi lebih terstruktur dan siap digunakan untuk analisis lebih lanjut seperti klasifikasi atau regresi.

```
df_aggregated = df.groupby(['NIM', 'NAMA_MK'], as_index=False).first()

# Pivot the DataFrame to make each course a column
df_pivoted = df_aggregated.pivot(index='NIM', columns='NAMA_MK', values='NILAI_ANGKA')

# Reset index to make 'NIM' a column
df_pivoted.reset_index(inplace=True)

# Show the result
print(df_pivoted)
```

Gambar 3.5 Feature Engineering

3.2.4 *Modeling*

3.2.4.1 Analisis Hubungan Lingkungan Belajar dan Prestasi Akademik

Pada tahap ini, pendekatan *Big Data Analytics* digunakan untuk menganalisis dan mengevaluasi hubungan antara lingkungan belajar (misalnya, keaktifan dalam kelas, kehadiran, dan penggunaan platform pembelajaran) dengan prestasi akademik mahasiswa. Proses analisis memanfaatkan teknik statistik dan teknik *machine learning* menggunakan beberapa algoritma metode LR, RF dan kombinasi LR dan RF. untuk menggali pola-pola tersembunyi dalam data.

1. Metode Linear Regression

Regresi linear digunakan untuk mengevaluasi seberapa besar pengaruh variabel-variabel lingkungan belajar seperti kehadiran, jumlah mata kuliah yang diambil, dan rata-rata nilai terhadap capaian akademik mahasiswa yang direpresentasikan dalam bentuk IPK (Indeks Prestasi Kumulatif). Dengan menggunakan regresi linear, peneliti dapat memperoleh gambaran yang lebih jelas

mengenai kontribusi masing-masing variabel terhadap hasil akhir akademik, serta menilai signifikansi hubungan yang terbentuk di antara mereka.

Keunggulan utama dari regresi linear adalah kemampuannya dalam menyajikan interpretasi yang sederhana dan transparan terhadap pengaruh variabel prediktor. Misalnya, koefisien regresi dapat menunjukkan seberapa besar peningkatan IPK yang diharapkan jika variabel tertentu mengalami peningkatan satu satuan. Pendekatan ini sangat ber dalam penelitian pendidikan yang bertujuan untuk memberikan rekomendasi berbasis data bagi kebijakan akademik. Studi oleh El Jihaoui et al. [1] misalnya, menunjukkan bahwa penggunaan regresi linear berhasil menjelaskan sekitar 88,53% variasi nilai akademik mahasiswa melalui faktor-faktor lingkungan, menandakan kekuatan model ini dalam menggambarkan hubungan yang kompleks namun masih bersifat linier.

Di samping itu, regresi linear juga memiliki keunggulan dalam hal efisiensi komputasi dan kemudahan implementasi. Berbeda dengan model yang lebih kompleks seperti *Random Forest* atau *XGBoost*, regresi linear tidak memerlukan parameter tuning yang rumit dan relatif lebih cepat dalam proses pelatihan data. Hal ini menjadikan regresi linear sangat cocok untuk dataset berskala sedang hingga besar yang memiliki struktur data yang cukup sederhana atau hubungan yang cenderung linear. Dalam penelitian oleh Airlangga [4], regresi linear bahkan menunjukkan performa yang lebih unggul dibanding *Random Forest* dalam memprediksi nilai akademik, terutama dalam situasi di mana data memiliki hubungan semi-linear yang dapat ditangkap dengan baik oleh pendekatan linier sederhana.

Selain untuk prediksi, regresi linear juga memiliki fungsi yang lebih luas dalam konteks analisis hubungan antar variabel. Dengan model ini, peneliti dapat mengidentifikasi variabel mana yang paling signifikan dalam mempengaruhi IPK mahasiswa, serta menilai apakah hubungan tersebut bersifat positif atau negatif. Informasi ini penting dalam menyusun intervensi pendidikan, seperti menargetkan program peningkatan kehadiran atau penguatan kemampuan manajemen waktu

bagi mahasiswa yang terindikasi berisiko. Oleh karena itu, regresi linear tidak hanya ber dalam menghasilkan prediksi, tetapi juga dalam membentuk pemahaman yang mendalam tentang dinamika pembelajaran di lingkungan pendidikan tinggi.

Dengan mempertimbangkan kelebihan-kelebihan tersebut, pemilihan regresi linear dalam penelitian ini dirasa sangat tepat. Selain memberikan hasil yang mudah diinterpretasikan, model ini juga mampu mengakomodasi tujuan utama penelitian, yaitu memahami hubungan antara variabel lingkungan belajar dengan prestasi akademik mahasiswa. Walaupun model-model kompleks lain mungkin mampu memberikan akurasi prediksi yang lebih tinggi, regresi linear tetap menjadi pilihan utama ketika analisis diarahkan untuk memperoleh insight yang dapat ditindaklanjuti secara praktis oleh institusi pendidikan.

2. Metode *Random Forest*

Selain regresi linear, penelitian ini juga memanfaatkan algoritma *Random Forest* untuk mengevaluasi hubungan antar variabel dalam konteks prediksi prestasi akademik mahasiswa. *Random Forest* merupakan metode ensemble yang terdiri dari sejumlah pohon keputusan (decision trees) yang dibangun secara acak dari subset data dan fitur. Berbeda dengan regresi yang hanya menghasilkan model linier, *Random Forest* mampu menangkap pola non-linier dan interaksi kompleks antar variabel. Hal ini menjadikannya sangat efektif dalam menjelaskan hubungan yang mungkin tidak terlihat secara eksplisit melalui analisis linier. Dengan membangun banyak pohon keputusan dan melakukan voting agregat terhadap hasilnya, *Random Forest* meningkatkan akurasi prediksi sekaligus mengurangi risiko overfitting yang sering terjadi pada model tunggal.

Keunggulan utama dari *Random Forest* adalah kemampuannya dalam memberikan visualisasi struktur pengambilan keputusan dan mengidentifikasi tingkat kepentingan (*feature importance*) dari masing-masing variabel. Dalam

konteks penelitian ini, algoritma ini digunakan untuk menentukan seberapa besar kontribusi variabel seperti rata-rata nilai, kehadiran, dan jumlah mata kuliah terhadap pencapaian IPK. Pendekatan ini tidak hanya menghasilkan model prediktif yang kuat, tetapi juga memberikan informasi penting bagi pembuat kebijakan di bidang pendidikan untuk merumuskan strategi intervensi. Sejalan dengan temuan Putri dan Rusdah [2], yang menunjukkan bahwa algoritma C4.5 pendahulu *Random Forest* dapat mencapai akurasi prediksi sebesar 93,9% dalam konteks prediksi kelulusan, *Random Forest* dianggap lebih unggul dalam memberikan interpretasi yang intuitif dan aplikatif.

Penelitian lain oleh Putri et al. [3] juga mendukung efektivitas *Random Forest* dalam konteks pendidikan. Dalam studi tersebut, *Random Forest* dibandingkan dengan berbagai algoritma lainnya, termasuk C4.5 dan *Naive Bayes*. Hasilnya menunjukkan bahwa selain memiliki akurasi yang kompetitif, *Random Forest* dinilai lebih mudah dipahami oleh pengambil keputusan karena mampu memetakan variabel-variabel penting dalam bentuk grafik kontribusi. Kejelasan visualisasi ini membantu institusi pendidikan untuk lebih cepat mengidentifikasi titik-titik lemah dalam sistem pembelajaran dan menetapkan prioritas kebijakan berdasarkan data. Oleh karena itu, penggunaan *Random Forest* dalam penelitian ini tidak hanya berperan sebagai alat analitik, tetapi juga sebagai instrumen praktis untuk mendukung perumusan kebijakan pendidikan berbasis bukti.

Secara keseluruhan, integrasi algoritma *Random Forest* dalam penelitian ini memperkaya dimensi analisis dan membuka peluang untuk mengungkap pola-pola tersembunyi yang tidak terdeteksi oleh pendekatan linier semata. Dengan kemampuannya dalam menangani data yang kompleks dan memberikan interpretasi yang mudah dipahami, *Random Forest* menjadi pilihan tepat dalam menjembatani kebutuhan analisis ilmiah dengan implementasi kebijakan nyata di lingkungan pendidikan tinggi.

3. Kombinasi Regresi dan *Random Forest*

Dalam implementasi nyata, regresi linear dan *Random Forest* dapat digunakan secara bersamaan untuk saling melengkapi dalam proses analisis data pendidikan. Regresi linear berfungsi sebagai alat yang efektif untuk mengukur kekuatan dan arah hubungan antar variabel numerik, seperti hubungan antara rata-rata nilai, kehadiran, dan jumlah mata kuliah dengan capaian IPK mahasiswa. Model ini memberikan hasil yang mudah diinterpretasikan dalam bentuk persamaan matematis dan koefisien, sehingga sangat ber untuk menjelaskan secara langsung bagaimana perubahan pada satu variabel mempengaruhi variabel lainnya.

Sementara itu, *Random Forest* lebih unggul dalam menangani data yang kompleks dan heterogen, terutama ketika digunakan untuk melakukan klasifikasi atau segmentasi kelompok mahasiswa berdasarkan pola lingkungan belajar mereka. Dengan kemampuan menangkap hubungan non-linier dan interaksi antar fitur yang kompleks, algoritma ini dapat mengidentifikasi kelompok mahasiswa yang memiliki karakteristik serupa berdasarkan variabel-variabel lingkungan, seperti dukungan keluarga, aktivitas non-akademik, atau akses terhadap fasilitas kampus. Beberapa penelitian terkini [4]–[6] mendukung pendekatan gabungan ini, di mana regresi linear digunakan untuk mendeteksi pengaruh langsung antar variabel, dan *Random Forest* digunakan untuk analisis prediktif dan segmentasi yang lebih mendalam. Kombinasi keduanya tidak hanya meningkatkan akurasi hasil analisis, tetapi juga memberikan insight yang lebih komprehensif bagi institusi pendidikan dalam merancang intervensi dan kebijakan berbasis data.

3.2.5 *Evaluation*

Evaluasi dilakukan untuk menilai keakuratan dan efektivitas model yang digunakan dalam penelitian. Tujuan dari tahap ini adalah untuk mengukur sejauh mana model yang dibangun mampu memprediksi atau merepresentasikan data secara akurat, serta memastikan bahwa hasil analisis dapat diandalkan untuk pengambilan keputusan. Dalam konteks ini, evaluasi dilakukan terhadap dua

model regresi, yaitu **Linear Regression** dan **Random Forest Regression**, dengan membandingkan nilai R^2 (R-squared) dan *MAE* (Mean Absolute Error). Nilai R^2 digunakan untuk melihat seberapa besar variasi data target (IPK) yang dapat dijelaskan oleh model, sedangkan *MAE* menunjukkan rata-rata selisih absolut antara nilai prediksi dan nilai aktual.

Beberapa fungsi dari pustaka scikit-learn digunakan dalam proses evaluasi ini, antara lain:

1. r2_score(): untuk menghitung nilai koefisien determinasi (R^2), yang menunjukkan tingkat kecocokan antara nilai aktual dan prediksi.
2. mean_absolute_error(): untuk mengukur rata-rata kesalahan absolut dari prediksi model.
   cross_val_score(): digunakan untuk melakukan validasi silang (*cross-validation*) sebanyak 5 kali lipat (*5-fold*) menguji kestabilan dan generalisasi model terhadap data yang berbeda-beda.

Dengan memanfaatkan fungsi-fungsi tersebut, hasil evaluasi menunjukkan bahwa Random Forest Regression memiliki performa yang lebih unggul dibandingkan Linear Regression, baik dari segi akurasi prediksi maupun stabilitas model. Nilai R^2 yang tinggi dan *MAE* yang rendah menjadi indikator bahwa model mampu memberikan hasil prediktif yang kuat dan dapat digunakan sebagai dasar analisis dalam memahami kontribusi variabel akademik terhadap IPK mahasiswa.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan dua sumber utama dalam pengumpulan data, yaitu dataset historis mahasiswa dan survey mahasiswa melalui Google form (https://forms.gle/F6wLvDtkTqej42pLA) periode pengisian dari tanggal 10 Mei 2025, memperoleh gambaran yang komprehensif mengenai pengaruh kualitas lingkungan belajar terhadap prestasi akademik mahasiswa.

3.3.1 Dataset Historis Mahasiswa (2010 - 2024)

Data historis mahasiswa diperoleh dari dataset lama yang tersedia sejak tahun 2010 hingga tahun 2024. Dataset ini mencakup informasi akademik mahasiswa, seperti:

1. Indeks Prestasi Semester (IPS) dan Indeks Prestasi Kumulatif (IPK)
2. Data kehadiran perkuliahan
3. Data aktivitas akademik (penggunaan e-learning, partisipasi diskusi, dsb.)
4. Kondisi sosial ekonomi mahasiswa (jika tersedia)

Dataset ini digunakan untuk memahami pola dan tren prestasi akademik mahasiswa dalam jangka waktu yang lebih panjang serta bagaimana faktor lingkungan belajar dapat mempengaruhi capaian akademik mereka.

3.3.2 Survei Mahasiswa Saat Ini

Selain data historis, penelitian ini juga mengumpulkan data melalui survey yang ditujukan kepada mahasiswa saat ini untuk mendapatkan perspektif langsung mengenai kualitas lingkungan belajar mereka. Survey ini mencakup beberapa aspek utama, antara lain:

1. Kondisi lingkungan belajar di rumah dan kampus
2. Ketersediaan fasilitas akademik (laboratorium, perpustakaan, internet, dsb.)
3. Metode pembelajaran dan tingkat keterlibatan dalam perkuliahan
4. Faktor eksternal seperti kondisi ekonomi dan sosial
5. Tingkat kepuasan mahasiswa terhadap lingkungan belajar mereka

Survei ini dilakukan menggunakan kuesioner daring yang disebarkan kepada mahasiswa aktif dari berbagai program studi. Data yang dikumpulkan melalui survei ini akan digunakan untuk memperkaya analisis terhadap pengaruh lingkungan belajar terhadap prestasi akademik mahasiswa saat ini.

3.4 Teknik Analisis Data

Teknik analisis data dalam penelitian ini mengikuti prinsip *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*), yaitu suatu kerangka kerja yang sistematis dan umum digunakan dalam proyek analisis data untuk memastikan proses berjalan terstruktur dan menghasilkan insight yang relevan. Model ini terdiri dari enam tahap utama yang saling terhubung dan iteratif, yang memungkinkan peneliti untuk melakukan eksplorasi data secara menyeluruh, mengembangkan model prediktif, serta menginterpretasikan hasil secara komprehensif., dengan tahapan sebagai berikut:

3.4.1 Business & Data Understanding

Tahapan awal dalam penelitian ini mengikuti prinsip *CRISP-DM* dan dimulai dari fase Business Understanding, yaitu memahami permasalahan utama yang ingin diselesaikan. Dalam konteks ini, fokus penelitian adalah mengidentifikasi faktor-faktor yang mempengaruhi prestasi akademik mahasiswa serta pengelompokkan mahasiswa berdasarkan kondisi lingkungan belajarnya. Prestasi akademik sering kali menjadi indikator utama keberhasilan institusi pendidikan tinggi, namun rendahnya IPK tidak selalu berasal dari faktor internal seperti motivasi atau kecerdasan, melainkan juga dapat dipengaruhi oleh faktor eksternal, seperti dukungan keluarga, akses fasilitas kampus, atau tekanan sosial. Oleh karena itu, penelitian ini bertujuan untuk menyusun pemodelan berbasis data mengungkap pola-pola lingkungan belajar serta variabel-variabel akademik yang berkaitan dengan capaian IPK mahasiswa.

Setelah tujuan dirumuskan, tahap berikutnya adalah Data Understanding, yaitu eksplorasi dan pemahaman awal terhadap data yang digunakan. Dalam penelitian ini, terdapat dua jenis dataset utama: data hasil survei lingkungan belajar dan data akademik mahasiswa. Dataset survei terdiri dari 292 responden dan berisi variabel-variabel seperti dukungan keluarga, manajemen waktu, fasilitas kampus, hingga kebiasaan berolahraga, yang bersifat ordinal maupun kategorikal. Sementara itu, dataset akademik berisi informasi dari 206 mahasiswa,

mencakup rata-rata nilai, kehadiran, jumlah mata kuliah yang diambil, serta IPK sebagai variabel target. Tahap ini juga mencakup analisis awal terhadap distribusi data, identifikasi data kosong, serta pengelompokan variabel berdasarkan karakteristiknya untuk mempersiapkan proses transformasi data pada tahap selanjutnya.

3.4.2 Data Preparation

Tahap *data preparation* merupakan bagian fundamental dalam proses data mining yang bertujuan untuk memastikan bahwa data berada dalam kondisi optimal sebelum digunakan dalam proses pemodelan. Kualitas data yang baik akan meningkatkan akurasi, stabilitas, serta interpretabilitas model yang dihasilkan. Dalam penelitian ini, proses persiapan data mencakup beberapa tahapan penting seperti pembersihan data, transformasi variabel, serta penyesuaian format agar sesuai dengan kebutuhan algoritma yang digunakan.

1. Langkah pertama yang dilakukan adalah penghapusan kolom identitas seperti "Nama", "Program Studi", dan "Timestamp". Kolom-kolom ini dihilangkan karena tidak memberikan kontribusi langsung terhadap analisis, serta berpotensi menimbulkan bias atau pelanggaran privasi. Keberadaan informasi pribadi juga tidak diperlukan dalam proses clustering maupun prediksi IPK, sehingga penghapusannya membantu menjaga fokus model terhadap variabel yang relevan.
2. Selanjutnya, dilakukan konversi tipe data pada variabel-variabel yang memiliki format kategorikal, terutama jawaban dalam bentuk “Ya” dan “Tidak”. Nilai-nilai tersebut diubah menjadi bentuk numerik biner (1 dan 0) agar dapat diolah secara matematis. Proses ini dikenal sebagai encoding sederhana dan menjadi prasyarat penting dalam analisis statistik maupun pemodelan *machine learning*, karena sebagian besar algoritma hanya menerima input numerik.
3. Untuk kategori yang memiliki lebih dari dua nilai unik (seperti jenis kelamin, tingkat semester, atau status pekerjaan), digunakan teknik

one-hot encoding agar tiap kategori dapat direpresentasikan secara terpisah dan tidak dianggap memiliki urutan nilai tertentu. Teknik ini menghasilkan beberapa kolom baru dari satu variabel kategorikal, yang masing-masing menunjukkan kehadiran atau ketidakhadiran suatu kategori dalam setiap entri data. Hal ini membantu model membedakan pengaruh dari masing-masing kategori tanpa memperkenalkan hubungan kuantitatif yang tidak valid.

4. Setelah semua variabel berada dalam bentuk numerik, dilakukan proses *standarisasi* menggunakan metode StandardScaler. Standardisasi bertujuan untuk menyamakan skala seluruh variabel agar memiliki rata-rata 0 dan standar deviasi 1. Ini sangat penting terutama untuk algoritma seperti K-Means dan PCA yang sensitif terhadap perbedaan skala antar fitur. Dengan demikian, seluruh data yang telah melalui proses pembersihan, transformasi, dan normalisasi ini kemudian siap digunakan untuk tahap modeling dan evaluasi selanjutnya.

3.4.3 Feature Selection & Engineering

Selain proses pembersihan dan transformasi data, dilakukan pula tahap pemilihan fitur (*feature selection*) mengurangi *noise*, menurunkan kompleksitas model, serta meningkatkan akurasi prediksi. Pemilihan fitur bertujuan untuk menyaring variabel-variabel yang paling relevan terhadap target atau tujuan analisis, sehingga model yang dibangun dapat bekerja secara lebih efisien dan tidak terganggu oleh informasi yang tidak signifikan. Dalam konteks penelitian ini, teknik seperti *SelectKBest*, yang memilih fitur berdasarkan nilai statistik tertentu, dan algoritma *Random Forest*, yang menyediakan metrik *feature importance*, dapat dimanfaatkan untuk mengidentifikasi fitur-fitur dengan kontribusi terbesar. Di samping itu, jika ditemukan bahwa variabel yang ada belum cukup menggambarkan fenomena yang ingin dianalisis, maka dilakukan pula proses *feature engineering* untuk membentuk fitur-fitur baru yang lebih informatif dan representatif, seperti rasio atau interaksi antar-variabel. Langkah

ini bertujuan untuk menangkap pola tersembunyi dalam data yang mungkin tidak langsung terlihat dari variabel aslinya.

3.4.4 Modeling

Penerapan dua algoritma:" dengan"Pembangunan model *machine learning* menerapkan penggunaan algoritma/ metode RF(*Random Forest*) dan LR(Linear Regression) dengan pertimbangan

1. *Random Forest Regressor*: Penerapan algoritma *Random Forest Regressor* dalam penelitian ini didasarkan pada kemampuannya dalam menangani hubungan non-linier antar variabel prediktor dengan variabel target, dalam hal ini nilai Indeks Prestasi Kumulatif (IPK) mahasiswa. *Random Forest* merupakan metode ensemble learning yang menggabungkan banyak pohon keputusan (*decision trees*) secara paralel dan melakukan prediksi berdasarkan rata-rata hasil dari semua pohon. Karakteristik ini membuat model menjadi lebih stabil dan tahan terhadap overfitting dibandingkan dengan pohon keputusan tunggal. *Random Forest* juga dapat menangani data dengan banyak fitur dan interaksi antar variabel tanpa perlu asumsi linearitas seperti pada model regresi klasik.

Keunggulan lain dari *Random Forest* adalah kemampuannya untuk menghasilkan nilai *feature importance*, yaitu metrik yang menunjukkan sejauh mana masing-masing variabel berkontribusi terhadap hasil prediksi. Dalam konteks penelitian ini, *feature importance* digunakan untuk mengidentifikasi faktor-faktor akademik mana yang paling berpengaruh terhadap IPK mahasiswa. Hasilnya memperlihatkan bahwa rata-rata nilai mata kuliah (rata2_nilai) memiliki kontribusi terbesar, diikuti oleh rata-rata kehadiran (rata2_hadir), dan terakhir jumlah mata kuliah yang diambil (jumlah_mk_diambil). Temuan ini memberikan insight penting dalam menyusun kebijakan pendidikan berbasis data.

Model *Random Forest* selanjutnya dievaluasi menggunakan dua metrik utama, yaitu nilai koefisien determinasi (R^2) dan *Mean Absolute Error* (MAE).

R^2 digunakan untuk mengukur seberapa besar variasi nilai IPK dapat dijelaskan oleh model, sedangkan MAE menunjukkan rata-rata selisih absolut antara nilai IPK yang diprediksi dan nilai aktualnya. Untuk memastikan bahwa model tidak hanya bekerja baik pada data tertentu saja, dilakukan pula proses *cross-validation* sebanyak 5 lipatan (5-fold CV). Hasil evaluasi menunjukkan bahwa *Random Forest* memiliki performa prediksi yang tinggi, stabil, dan unggul dalam menangkap pola kompleks dari data akademik mahasiswa.

2. *Linear Regression*: Model *Linear Regression* diterapkan dalam penelitian ini sebagai model pembanding dengan pendekatan yang lebih sederhana dan interpretable. *Linear Regression* merupakan metode statistik klasik yang digunakan untuk memodelkan hubungan linier antara satu atau lebih variabel bebas dengan variabel terikat. Dalam kasus ini, model berupaya untuk memetakan hubungan linier antara tiga variabel akademik (rata2_nilai, rata2_hadir, jumlah_mk_diambil) dengan IPK mahasiswa sebagai target. Model ini dipilih karena kemampuannya dalam memberikan hasil yang mudah dipahami melalui koefisien regresi yang dapat diinterpretasikan secara langsung.

Setiap koefisien pada model *Linear Regression* menunjukkan pengaruh langsung dari satu variabel terhadap IPK, dengan asumsi variabel lain tetap konstan. Hal ini memungkinkan peng model, seperti dosen atau pihak akademik, untuk memahami arah dan besaran pengaruh dari masing-masing variabel terhadap capaian akademik mahasiswa. Misalnya, koefisien positif pada variabel kehadiran menunjukkan bahwa peningkatan kehadiran berasosiasi dengan peningkatan IPK. Meskipun pendekatan ini memiliki keterbatasan dalam menangkap hubungan non-linier dan interaksi antar fitur, kesederhanaannya menjadi kekuatan tersendiri dalam konteks edukasi.

Evaluasi terhadap model Linear Regression dilakukan dengan metrik yang sama, yaitu R^2 dan MAE, untuk memudahkan perbandingan langsung dengan model *Random Forest.* Selain itu, dilakukan pula 5-*fold cross-validation*

untuk mengevaluasi stabilitas model terhadap variasi data. Hasil yang diperoleh menunjukkan bahwa meskipun *Linear Regression* memiliki performa prediksi yang cukup baik, namun masih berada di bawah *Random Forest* dalam hal akurasi dan kemampuan menangkap kompleksitas data. Kendati demikian, model ini tetap relevan digunakan sebagai baseline dan alat bantu interpretasi yang kuat dalam analisis akademik.

```
# Random Forest Regressor
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

# Linear Regression
lr = LinearRegression()
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
```

Gambar 3.6 Modeling *Random Forest* dan *Linear Regression*

3.4.5 Evaluation

Evaluasi model dilakukan untuk menilai sejauh mana model yang dibangun mampu merepresentasikan hubungan antara variabel input (kualitas lingkungan belajar) dan output (prestasi akademik atau persepsi dukungan dosen) secara akurat dan dapat digeneralisasi. Tujuan utama dari proses evaluasi adalah memastikan bahwa model tidak hanya bekerja baik pada data pelatihan, tetapi juga mampu memberikan prediksi yang andal pada data baru. Menurut Han, Kamber, & Pei (2012), evaluasi model merupakan tahap penting dalam proses data mining karena memberikan dasar untuk mengukur kinerja, menghindari overfitting, dan memilih model terbaik berdasarkan metrik yang relevan.

```
# Evaluasi
print("Random Forest R²:", r2_score(y_test, y_pred_rf))
print("Linear Regression R²:", r2_score(y_test, y_pred_lr))
```

Gambar 3.7 Evaluasi *Modeling*

Evaluasi dalam penelitian ini dilakukan secara komprehensif, meliputi penghitungan kualitas model menggunakan metrik seperti R^2 (koefisien determinasi), MAE (*Mean Absolute Error*), dan CV R^2 (*Cross-Validation* R^2). Selain itu, dilakukan analisis *feature importance* pada algoritma *Random Forest* dan interpretasi koefisien pada *Linear Regression* untuk mengidentifikasi variabel lingkungan belajar yang paling signifikan dalam mempengaruhi persepsi dukungan dosen. Validasi silang dan analisis residual juga digunakan untuk menguji stabilitas dan generalisasi model terhadap data lain, sehingga dapat dipercaya untuk pengambilan kesimpulan dan rekomendasi.

3.4.6 *Deployment*

Tahap *deployment* merupakan proses penerapan model yang telah dibangun agar dapat diakses dan dimanfaatkan oleh peng akhir. Dalam penelitian ini, model yang dikembangkan akan di-*deploy* menggunakan *platform Streamlit*, yaitu *framework open-source* berbasis *Python* yang memungkinkan pembuatan aplikasi web interaktif untuk visualisasi data dan hasil prediksi secara *real-time*.

Model analisis lingkungan belajar dan prediksi IPK yang telah dilatih sebelumnya akan dikemas dalam antarmuka yang sederhana dan *user-friendly*. Pengguna, seperti dosen atau pihak akademik, dapat memasukkan data lingkungan belajar mahasiswa dan memperoleh prediksi IPK secara langsung melalui tampilan web interaktif. Proses input, prediksi, dan visualisasi akan dijalankan secara otomatis oleh sistem yang dibangun menggunakan skrip *Python*.

Aplikasi ini akan di-*hosting* melalui repository *GitHub* yang terhubung langsung dengan akun *Streamlit Cloud*, sehingga dapat diakses secara publik melalui tautan *URL* tanpa memerlukan instalasi lokal. Dengan pendekatan ini, *deployment* bersifat ringan, fleksibel, dan mudah diperbarui kapan saja hanya dengan melakukan pembaruan (*push*) ke dalam repository. Strategi ini juga memungkinkan kolaborasi lanjutan dan pengembangan fitur lebih lanjut secara terbuka dan terkontrol.