JUDUL SKRIPSI

SKRIPSI

Oleh :

NAMA

NPM

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"

JAWA TIMUR

2020

LEMBAR PENGESAHAN

SKRIPSI

Judul : Klasifikasi Karakteristik Mahasiswa Berpotensi Drop-Out Menggunakan Algoritma K-Nearest Neighbors yang Ditingkatkan dengan Principal Component Analysis

Oleh : ADISTY REGINA

NPM : 21081010246

Telah Diseminarkan Dalam Ujian Skripsi Pada :

Hari Kamis, Tanggal 9 Juni 2016

Mengetahui

Dosen Pembimbing	Dosen Penguji
1.	1.
Wahyu S.J Saputra, S.Kom, M.Kom	**Intan Yuniar Purbasari, S.Kom, M.Sc**
NPT : 3 8606 10 0295 1	**NPT : 3 8606 04 0198 1**
2.	2.
Budi Nugroho, S.Kom, M.Kom	**Faisal Muttaqin, S.Kom, MT**
NPT : 3 8009 05 0205 1	**NPT : 3 8512 13 0351 1**

Menyetujui

Dekan	Koordinator Program Studi
Fakultas Ilmu Komputer	**Teknik Informatika**
Dr. Ir. Ni Ketut Sari, MT	**Budi Nugroho, S.Kom, M.Kom**
NPT : 19650731 1199203 2 001	**NPT : 3 8009 05 0205 1**

OPTIMASI DETEKSI LOKASI PLAT NOMOR KENDARAAN MENGGUNAKAN METODE MAXIMALLY STABLE EXTREMAL REGIONS

Nama Mahasiswa	**: Hendra Maulana**
NPM	**0334010261**
Program Studi	**: Teknik Informatika**
Dosen Pembimbing	**: Fetty Tri Anggraeny, S.Kom., M.Kom**
	Wahyu SJ Saputra, S.Kom, M.Kom

Abstrak

Pengenalan plat nomor kendaraan sangat mendukung sistem infrastruktur cerdas, seperti: aplikasi pembayaran jalan tol dan parkir, aplikasi monitoring jalan tol, aplikasi monitoring lalu lintas, dan lain sebagainya. Meskipun telah banyak metode pengenalan plat yang telah menunjukkan kinerja menjanjikan, namun beberapa metode mungkin gagal dalam situasi yang lebih kompleks karena kompleksitas seperti variasi posisi dan orientasi plat, berbagai latar belakang, dan benda-benda non-plat.

Beberapa penelitian telah dilakukan untuk meningkatkan kinerja deteksi lokasi plat yang difokuskan untuk menemukan deskripsi fitur visual yang kuat untuk berbagai latar belakang dan orientasi plat itu sendiri. Untuk efisiensi pencocokan visual yang lebih tinggi, beberapa detektor keypoint cepat dan deskripsi yang sesuai telah dilakukan penelitian, seperti seperti fitur FAST, SURF, BRISK, Harris Corner. Dan ada juga fitur MSER yang metode pencarian keypoint berdasarkan extremal regionnya.

MSER sangat efisien untuk mendeteksi karakter dengan setengah atau daerah tertutup sepenuhnya, terutama pada karakter yang memiliki lubang, seperti 0, 6, 8, 9, A, B, D, P, Q. Namun MSER kurang maksimal dalam mendeteksi karakter yang mempunyai sudut, seperti 1, 7, H, I, dan lain sebagainya. Pada penelitian ini diusulkan penggabungan ekstraksi fitur metode Maximally Stable Extremal Regions (MSER) yang pencarian keypoint dilakukan berdasarkan extremal region dan Harris Corner yang metode pencarian keypoint dilakukan berdasarkan titik

pojok (corner detection) diharap dapat meningkatkan akurasi dan waktu komputasional yang optimal pada deteksi lokasi plat nomer kendaraan berdasarkan beragam situasi.

Kata kunci: *Ekstraksi fitur, FAST, SURF, BRISK, Harris Corner, MSER*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

Bagian utama skripsi terdiri dari beberapa komponen atau bab yang tersusun dengan alur yang logis. Pendahuluan merupakan komponen/bab pertama yang harus menjelaskan apa yang akan dikerjakan dalam skripsi dan mengapa ini perlu dikerjakan.

1.1. Latar Belakang

Angka drop-out di institusi pendidikan tinggi masih menjadi masalah signifikan. Setiap tahun, banyak mahasiswa yang tidak mampu menyelesaikan studi mereka, yang berakibat pada pemborosan sumber daya pendidikan dan menghambat pengembangan individu serta Masyarakat. Penelitian ini menunjukkan bahwa meskipun ada penurunan dalam angka drop-out, tantangan untuk mengidentifikasi mahasiswa yang beresiko tetap ada (Nafisa Tasnim et al, 2019)

Seiring dengan kemajuan teknologi dan peningkatan data yang tersedia di institusi pendidikan, tantangan baru muncul dalam menganalisis dan memprediksi performa mahasiswa, terutama dalam mengidentifikasi mahasiswa dengan potensi drop-out (DO). Fenomena drop-out merupakan masalah penting yang memengaruhi kualitas pendidikan serta reputasi instusi. Di dalam konteks pendidikan tinggi, mengatahui karakteristik mahasiswa yang berpotensi drop-out dapat membantu institusi untuk mengambil Tindakan preventif. Berdasarkan penelitian terdahulu, algoritma klasifikasi data telah digunakan secara luas untuk tujuan ini (Emny Yossy et al, 2020)

Prediksi performa mahasiswa merupakan masalah yang sangat penting dalam meningkatkan proses pembelajaran. Tingkat kinerja siswa mungkin dapat dipengaruhi oleh beberapa faktor pada beberapa tahun belakangan. Prediksi awal dari performa mahasiswa mungkin dapat membantu untuk meningkatkan proses pembelajaran. Prediksi performa mahasiswa dapat di peroleh dengan menggunakan teknik data mining pada Kumpulan data mahasiswa. Pengklasifikasian data adalah teknik yang paling penting dalam penelitian data mining. Klasifikasi data bergantung pada kategorisasi atau pemberian kelas data berdasarkan atribut

prediksi. Pengklasifikasian akan direpresentasikan oleh berbagai macam model. Selain itu, variasi dalam algoritma telah berkembang secara signifikan dari waktu ke waktu, sehingga menghasilkan kemajuan dalam klasifikasi data. Algoritma klasifikasi yang umum digunakan antara lain K-Nearest Neigbor Classifier, Naïve Bayes, Neural Networks, Algoritma Decision Tree (ID3, 34.5, dan Random Forest), dan Support Vector Machines (SVM) (Ihsan Amra & Ashraf Maghari, 2017).

Prediksi kinerja mahasiswa baru berdasarkan kesamaan hubungan, hal ini dapat meningkatkan kinerja mahasi'swa sebelum mereka menyelesaikan studinya. Menurut statistic pangkalan data pendidikan tinggi (PDDIKTI), dari 9.656.252 mahasiswa yang aktif, 11.472 mahasiswa dapat lulus dan 2.397 mahasiswa drop out, dan 1.793 mahasiswa mengundurkan diri. Dalam penelitian ini, penulis akan mengkaji dataset mahasiswa di salah satu perguruan tinggi di Jawa Timur. Makalah ini mengusulkan pengklasifikasian data karakteristik mahasiswa salah satu kampus di Jawa Timur berpotensi drop-out menggunakan algoritma KNN yang ditingkatkan dengan PCA (Principal Component Analysis) sebagai teknik klasifisikasi yang diterapkan pada dataset mahasiswa dari salah satu kampus di Jawa Timur.

Penelitian ini memfokuskan pada data mahasiswa di salah satu kampus di Jawa Timur, dengan menggunakan atribut-atribut yang berkaitan dengan akademik, social, dan demografis mahasiswa. Data ini dikumpulkan dari sistem manajemen kampus. Pengolahan data awal meliputi tahap pre-processing seperti pembersihan data, penghapusan data redundan, serta normalisasi data untuk mengurangi pengaruh variabel-variabel yang tidak relevan (Micheal Arowolo et al, 2020)

Kemampuan untuk memprediksi potensi drop-out secara akurat merupakan langkah peting untuk mencegah mahasiswa meninggalkan pendidikan sebelum menyelesaikan masa studi. Berbagai metode pengolahan data telah diusulkan untuk membantu mengidentifikasi mahasiswa yang berpotensi drop-out berdasarkan berbagai faktor, seperti Indeks Prestasi (IP), jumlah Satuan Kredit Semester (SKS), asal sekolah, pekerjaan orang tua, penghasilan orang tua, UKT, dan jenis tempat tinggal, pekerjaan sambilan atau tidak, jenis tempat tinggal, kehadiran, dan jurusan. Salah satu metode yang saat ini banyak digunakan untuk tugas klasifikasi dan prediksi adalah *K-Nearest Neighbor (KNN)*, sebuah algoritma

sederhana namun kuat yang dapat digunakan untuk menganalisis dan memprediksi berbagai karakteristik mahasiswa yang relevan (A H Lubis et al, 2020). Untuk meningkatkan akurasi model prediksi, penelitian ini menerapkan dua teknik utama: *Principal Component Analysis* (PCA) dan *K-Nearest Neighbor* (KNN). PCA digunakan untuk reduksi dimensi dengan mengekstraksi fitur utama dari dataset yang memiliki dimensi tinggi. Algoritma PCA berguna dalam mengurangi variabilitas yang tidak perlu tanpa kehilangan informasi penting, sehingga memudahkan klasifikasi data (Yang Yinghua et al, 2018). Setelah itu, algoritma KNN diterapkan untuk mengklasifikasikan mahasiswa berdasarkan jarak terdekat dengan sampel lainnya. KNN adalah algoritma yang efektif untuk klasifikasi Ketika data memiliki karakteristik non-linearitas kompleks (Micheal Arowolo et al, 2020).

PCA merupakan metode reduksi dimensi yang efektif dalam menangani data yang memiliki banyak variabel berkorelasi, seperti data mahasiswa. Dengan PCA, informasi utama dalam data dapat dipertahankan meskipun dimensi data dikurangi, yang menghasilkan model prediksi yang lebih efisien dan akurat (Yang Yinghua, 2020). Di sisi lain, KNN adalah algoritma yang mudah diterapkan dan efektif dalam mengklasifikasian data berdasarkan jarak terdekat antara sampel. Yang menjadikannya cocok untuk dataset dengan struktur yang kompleks (Emny Yossy, 2020)

Penelitian ini menghipotesiskan bahwa penerapan PCA dan KNN akan menghasilkan akurasi yang lebih baik dalam mengidentifikasi mahasiswa dengan potensi drop-out dibandingkan dengan klasifikasi konvensional tanpa reduksi dimensi. Fokus penelitian ini adalah pada klasifikasi lima kategori potensi drop-out: merah (potensi drop-out tinggi), oranye (potensi sedang-tinggi), kuning (potensi sedang), hijau (potensi rendah), dan biru (potensi sangat rendah atau tidak berpotensi drop out). Warna-warna ini digunakan untuk memberikan representasi visual yang mudah dipahami oleh pengambil Keputusan di institusi. (Emny Yossy et al, 2020).

KNN bekerja dengan membandingkan karakteristik data baru (dalam hal ini, data mahasiswa) dengan data yang telah ada, berdasarkan kedekatan antar data dalam ruang dimensi. Kunci keberhasilan KNN dalam memprediksi potensi drop-out mahasiswa terletak pada kemampuannya mengelompokkan mahasiswa kedalam kategori yang relevan dengan Tingkat risiko drop-out yang berbeda-beda. Namun, dengan semakin kompleksnya data pendidikan, yang sering kali memiliki banyak

dimensi atau atribut, penggunaan KNN dapat menjadi tidak efisien. Oleh karena itu, diterapkan metode *Principal Component Analysis (PCA)* untuk mengurangi dimensi data dan meningkatkan efisiensi algoritma (A H Lubis et al, 2020).

Principal Component Analysis (PCA) adalah metode pengurangan dimensi yang digunakan untuk menyederhanakan data dengan cara mengeliminasi fitur atau atribt yang kurang relevan. Dalam konteks data pendidikan, PCA memungkinkan kita untuk mengurangi jumlah variabel yang dianalisis tanpa kehlangan informasi penting yang berkaitan dengan potensi drop-out mahasiswa. Penggunaan PCA Bersama KNN tidak hanya membantu mempercepat proses analisis data, tetapi juga dapat meningkatkan akurasi prediksi dengan mengurangi kebisingan (noise) pada data (A H Lubis et al, 2020).

Tahap penting sebelum melakukan prediksi adalah pengumpulan dan preprocessing data. Data yang digunakan dalam penelitian ini mencakup berbagai faktor, seperti nilai akademik, jumlah kehadiran, status ekonomi, dan informasi demografis lainnya. Pengumpulan data ini harus dilakukan dengan hati-hati agar mencakup semua varibel yang berpotensi mempengaruhi Tingkat drop-out mahasiswa. Setelah data dikumpulkan, dilakukan preprocessing untuk membersihkan data dari data yang hilang atau anomaly, serta menormalisasi data agar siap untuk digunakan dalam algoritma KNN (Ihsan Amra & Ashraf Maghari, 2017).

Setelah data diproses, dilakukan klasifikasi menggunakan algoritma KNN yang telah ditingkatkan dengan teknik PCA untuk reduksi dimensi. Proses klasifikasi ini memungkinkan pengelompokan mahasiswa berdasarkan beberapa tingkat risiko drop-out. Proses klasifikasi ini menghasilkan pengelompokan mahasiswa berdasarkan lima tingkat risiko drop-out, yang ditandai dengan warna. Warna merah menunjukkan potensi drop-out yang tinggi, oranye untuk potensi drop-out sedang-tinggi, kuning untuk potensi drop-out sedang, hijau untuk potensi drop-out rendah, dan biru untuk mahasiswa dengan potensi drop-out sangat rendah atau tidak berpotensi drop-out sama sekali. Penggunaan algoritma PCA-KNN membantu mengidentifikasi mahasiswa dengan risiko drop-out lebih akurat dan efisien, dengan mengelompokkan data ke dalam beberapa kategori yang dapat dimanfaatkan oleh pihak universitas untuk melakukan intervensi yang lebih dini. Dengan pendekatan ini, institusi dapat mengurangi waktu pemrosesan data yang kompleks dan meningkatkan keakuratan prediksi melalui pengurangan atribut yang tidak relevan

(Yang Yinghua et al, 2018)

Evaluasi model klasifikasi dilakukan untuk memastikan bahwa sstem ini bekerja dengan baik dan memberikan prediksi yang akurat. Evaluasi dilakukan dengan menggunakan metrik seperti akurasi, sensitivitas, dan *false positive rate.* Dari hasil evaluasi, diketahui bahwa penggabungan KNN dan PCA dapat meningkatkan akurasi prediksi secara signifikan dibandingkan dengan metode klasifikasi konvensional. PCA membantu mengurangi jumlah atribut yang diproses, yang pada gilirannya meningkatkan efisiensi dan kinerja algoritma KNN (Ihsan Amra & Ashraf Maghari, 2017).

Dengan adanya model prediksi ini, diharapkan perguruan tinggi dapat lebih proaktif dalam mencegah terjadi drop-out. Mahasiswa yang terindentifikasi memiliki risiko drop-out tinggi dapat segera mendapatkan perhatian dan bimbingan tambahan untuk meningkatkan performa akademik mereka. Dengan demikian, model ini dapat menjadi alat yang efektif untuk meningkatkan tingkat kelulusan dan mengurangi angka drop-out di kalangan mahasiswa (A H Lubis et al, 2020).

Penilaian kinerja model dilakukan dengan menggunakan beberapa metrik seperti akurasi, precision, recall, dan F1-score. Model PCA-KNN dievaluasi terhadap data validasi menggunakan cross-validation, dengan hasil yang menunjukkan bahwa integrasi antara kedua algoritma tersebut memberikan hasil yang signifikan dalam mengklasifikasikan potensi drop-out mahasiswa (Yang Yinghua et al, 2018)

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang yang telah dijelaskan, terbentuk beberapa rumusan masalah sebagai berikut :

1. Bagaimana melakukan klasifikasi performa mahasiswa yang berpotensi drop out menggunakan kombinasi Principal Component Analysis (PCA) untuk reduksi dimensi dan optimasi algoritma K-Nearest Neighbors (KNN) guna meningkatkan akurasi klasifikasi?

1.3. Batasan Masalah

Penelitian ini memiliki beberapa batasan masalah yang didefinisikan oleh penulis untuk penelitian ini sebagai berikut:

1. Penelitian ini terbatas pada klasifikasi data karakteristik mahasiswa di salah satu kampus di Jawa Timur, yang dibagi 5 warna kelas, yaitu warna merah (potensi drop-out tinggi), warna orange (potensi drop-out sedang-tinggi), warna kuning (potensi drop-out sedang), warna hijau (potensi drop-out rendah), dan warna biru (tidak berpotensi drop-out atau potensi sangat rendah).

2. Dataset yang digunakan terdiri dari Kumpulan data mahasiswa dan didapatkan melalui Biro Akademik Kemahasiswaan Perencaan dan Kerjasama (BAKPK) UPN "Veteran" Jawa Timur.
3. Pengimplementasian algoritma dilakukan menggunakan Bahasa pemrograman Python.
4. Output yang dihasilkan dari penelitian ini akan diimplementasikan dalam bentuk API (*Application Programming Interface*).
5. Algoritma klasifikasi yang digunakan adalah K-Nearest Neighbors (KNN) yang telah ditingkatkan dengan metode Principal Component Analysis (PCA) untuk mengurangi dimensi data dan meningkatkan efisiensi data.

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disampaikan, tujuan dari penelitian ini adalah sebagai berikut.

1. Mengimplementasikan algoritma K-Nearest Neighbor (KNN) yang ditingkatkan dengan PCA (Principal Component Analysis) untuk mengklasifikasikan data karakteristik mahasiswa yang berpotensi drop-out di salah satu kampus di Jawa Timur.
2. Merancang model algoritma K-Nearest Neighbor (KNN) yang ditingkatkan dengan PCA (Principal Component Analysis) untuk mengklasifikasikan data karakteristik mahasiswa yang berpotensi drop-out di salah satu kampus di Jawa Timur.
3. Mengetahui Tingkat akurasi model algoritma K-Nearest Neighbor (KNN) yang ditingkatkan dengan PCA (Principal Component Analysis) untuk mengklasifikasikan data karakteristik mahasiswa yang berpotensi drop-out di salah satu kampus di Jawa Timur.
4. Memahami hasil klasifikasi yang dihasilkan oleh model algoritma K-Nearest Neighbor (KNN) yang ditingkatkan dengan PCA (Principal Component Analysis) untuk mengklasifikasikan data karakteristik mahasiswa yang berpotensi drop-out di salah satu kampus di Jawa Timur.
5. Merancang model algoritma prediksi yang diimplementasikan dalam bentuk API (Application Programming Interface) berbasis Python, sehingga dapat diakses dan digunakan oleh pengguna untuk memprediksi potensi drop-out mahasiswa secara praktis.

1.5. Manfaat Penelitian

Penelitian ini memberikan beberapa manfaat signifikan. Pertama, dengan mengevaluasi Tingkat akurasi algoritma K-Nearest Neighbor (KNN) yang ditingkatkan dengan Principal Component Analysis (PCA), penelitian ini dapat meningkatkan efektivitas dalam mengidentifikasi mahasiswa yang berpotensi drop-out. Hal ini membantu pihak kampus dalam membuat Keputusan yang lebih tepat dan efisien terkait Upaya pencegahan drop-out. Manfaat lain dari penelitian ini adalah pengembangan Solusi yang lebih efektif dan efisien untuk memprediksi potensi drop-out mahasiswa, yang dapat mempercepat proses identifikasi dan intervensi, sehingga mendukung institusi perguruan tinggi dalam memberikan responds yang lebih cepat dan tepat sasaran terhadap mahasiswa yang beresiko dropout. Dengan manfaat

ini, penelitian ini berpotensi memberikan dampak positif bagi pengelolaan dan intervensi dini terhadap mahasiswa yang berpotensi drop-out dari kampus**.**

BAB II

TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Setiap penelitian harus mempertimbangkan penelitian yang telah dilakukan sebelumnya. Hasil penelitian terdahulu tidak hanya berfungsi sebagai alat perbandingan dan sumber inspirasi, namun juga membantu peneliti dalam mengidentifikasi kelebihan dan kekurangan dari metode yang telah diaplikasikan. Pengetahuian ini sangat penting dan berguna karena menjadi dasar untu memilih dan meneraplam teknik yang paling efektif pada penelitian yang sedang dilakukan. Dibawah ini adalah beberapa penelitian mengenai topik yang diteliti dalam judul "Klasifikasi Karakteristik Mahasiswa Berpotensi Drop-Out Menggunakan Algoritma K-Nearest Neighbors yang Ditingkatkan dengan Principal Component Analysis" :

Pada tahun 2023, Michael Olaolu Arowolo dan rekan rekannya melakukan studi yang berjudul "PCA Model For RNA-Seq Malaria Vector Data Classification Using KNN And Decision Tree Algorithm". Penelitian ini menggunakan data RNA-Seq dari nyamuk Anopheles gambiae yang terdiri dari 2457 sampel gen yang dikumpulkan dari wilayah Kenya Barat antara tahun 2010 hingga 2012. Data ini kemudian diproses menggunakan metode Principal Component Analysis (PCA) reduksi dimensi, menghasilkan 45 komponen laten yang signifikan. Proses klasifikasi dilakukan menggunakan algoritma K-Nearest Neighbor (KNN) dan Dcision Tree, dengan metode validasi 10-fold cross-validation untuk mengevaluasi performa. Hasil penelitian menunjukkan bahwa algoritma KNN memberikan akurasi terbaik sebesar 86,7%, dengan presisi 86,1%, recall 94,9%, dan F-score 90,3%, sementara algoritma Decision Tree memberikan akurasi sebesar 83,3%, dengan presisi 85,4%, recall 89,7%, dan F-score 87,5%. Studi ini mengungkapkan bahwa kombinasi metode PCA dan KNN dapat meningkatkan akurasi dalam klasifikasi data genetik dan menunjukkan efektivitas metode ini dalam menangani dataset berdimensi tinggi. Penelitian ini merekomendasikan penggunaan metode PCA sebagai langkah awal untuk reduksi dimensi data pada dataset berdimensi tinggi, khususnya dalam kasus data RNA-Seq. Dengan menggunakan PCA, fitur-fitur yang tidak relevan dapat dihilangkan sehingga algoritma KNN dan Decision Tree dapat bekerja lebih efisien dan akurat. Hasil penelitian menunjukkan bahwa kombinasi PCA dan KNN menghasilkan akurasi yang lebih tinggi dibanding dengan Decision Tree, oleh karena itu, pendekatan ini dinilai lebih efektif untuk tugas klasifikasi (Arowolo et al, 2023).

Pada tahun 2018, Yang Yinghua dan rekan-rekannya melakukan penelitian berjudul "Fault Monitoring and Classification of Rotating Machine Based on PCA and KNN". Penelitian ini bertujuan untuk mengklasifikasikan berbagai jenis kerusakan pada mesin berputar menggunakan metode Principal Component Analysis (PCA) untuk reduksi dimensi dan K-Nearest Neighbors (KNN) untuk klasifikasi. Data yang digunakan meliputi sinyal getaran, kecepatan, tegangan, dan parameter lain dari perlatan historis yang dikumpulkan dari mesin bertekanan tinggi dengan dua jenis kerusakan, yaitu "light imbalance" dan "heavy imbalance". PCA diterapkan untuk mengurangi dimensi data, sementara KNN digunakan untuk mengklasifikasikan data kerusakan setelah proses reduksi dimensi. Setelah melalui berbagai eksperimen, nilai K terbaik yang dipilih adalah 7, dengan akurasi mencapai 98,75% untuk data kerusakan tipe pertama dan 99,75% untuk tipe kedua. Peneliti ini merekomendasikan penggunaan model PCA dan KNN ini dalam sistem deteksi secara real-time untuk memantau dan mengidentifikasikan kerusakan pada mesin berputar, mengingat efektivitasnya dalam pengklasifikasian berbagai jenis kerusakan.

Pada penelitian tahun 2020, Emny Harna Yossy dan rekan-rekannya melakukan studi yang berjudul "Comparison of Data Mining Classification Algorithms for Student Performance". Penelitian ini menggunakan dataset performa mahasiswa yang diambil dari UCI Machine Learning Repository. Dataset tersebut berisi data performa mahasiswa dalam matematika dengan total 394 siswa, yang mencakup atribut demografis, sosial, dan akademis. Penelitian ini menguji tujuh (7) metode klasifikasi, yaitu K-Nearest Neighbor, Decision Tree, Naive Bayes, AdaBoost, ExtraTree, Bernoulli Naive Bayes, dan Random Forest, dengan Python sebagai platform pengembangannya. Hasil penelitian menunjukkan bahwa algoritma Random Forest G memberikan hasil terbaik dengan akurasi 89,78%, diikuti oleh AdaBoost dengan akurasi 88,04%, dan K-Nearest Neighbor dengan akurasi 86,52%. Hasil ini memperlihatkan bahwa algoritma Random Forest G merupakan metode klasifikasi terbaik dalam memprediksi performa akademis mahasiswa. ada dataset performa mahasiswa, KNN hanya mencapai akurasi 86,52%, yang lebih rendah dibandingkan algoritma lain seperti Random Forest yang mencapai 89,78% dan AdaBoost dengan 88,04%. Hal ini menunjukkan bahwa KNN kurang optimal untuk menangkap pola kompleks dalam dataset ini.

Pada Februari 2019, Nafisa Tasnim dan rekan-rekannya melakukan penelitian berjudul "Identification of Drop Out Students Using Educational Data Mining". Penelitian ini menggunakan adataset "Student Performance Analysis" dari UCI Machine Learning Repository, yang terdiri dari dua set data dengan 395 dan 649 entri, masing-masing berisi 33 atribut. Untuk indentifikasi mahasiswa berpotensi drop-out, penelitian ini menggunakan metode perhitungan nilai ambang (threshold) berdasarkan atribut dengan informasi gain tertinggi, yaitu atribut yang paling mempengaruhi keputusan drop-out. Algoritma yang diterapkan juga mencakup deteksi outlier dengan menggunakan Cooks's Distance untuk mengurangi kesalahan klasifikasi yang dapat disebabkan oleh data yang menyimpang. Dataset pertama (Dataset A) memiliki 395 entri dengan 293 kelas positif (siswa yang

melanjutkan) dan 102 kelas negatif (siswa yang drop-out), menghasilkan rasio ketidakseimbangan 2,87 yang berari jumlah siswa yang melanjutkan hampir 3 kali lebih banyak daripada yang drop-out. Rasio ketidakseimbangan (Imbalance Ratio, IR) adalah klasifikasi yang menggambarkan perbandingan antara jumlah data dalam kelas mayoritas (kelas dengan jumlah entri terbanyak) dan kelas minoritas (kelas dengan jumlah entri lebih sedikit). Ketidakseimbangan data sering muncul ketika salah satu kelas dalam dataset jauh lebih dominan jumlahnya dibandingkan kelas lainnya, seperti kasus "siswa melanjutkan" dan "siswa drop out". Dataset kedua (Dataset B) berisi 649 entri dengan 584 kelas positif dan 65 kelas negatif, dengan rasio ketidakseimbangan 8,98 yang berarti jumlah siswa yang melanjutkan sekitar 9 kali lebih banyak dibandingkan yang drop-out. Dalam eksperimen ini, model diuji menggunakan validasi 5-fold, dan algoritma dijalankan dalam MATLAB R20017a. Hasilnya menunjukkan bahwa pendekatan berbasis threshold memiliki performa yang lebih baik dibandingkan algoritma lain seperti Logistic Regression, Naive Bayes, dan Support Vector Machine. Nilai akurasi, presisi, recall, dan F1-score untuk pendekatan ini sangat baik, bahkan pada data yang telah mengalami mengalami deteksi outlier, meunjukkan bahwa metode threshold ini dapat menangani data yang menyimpang dengan baik. Penelitian ini merekomendasikan penggunaan metode berbasis threshold dengan atribut berinformasi gain tinggi untuk klasifikasi mahasiswa berpotensi drop-out di masa depan.

Pada tahun 2021, Fandy Indra Pratama dan Avira Budianita melakukan penelitian berjudul "Optimization of K-NN Classification in Human Gait Recognition". Penelitian ini bertujuan untuk meningkatkan akurasi pengenalan dengan meningkatkan fitur ekstraksi melalui metode Gabor Wavelet dan Principal Component Analysis (PCA). Dataset yang digunakan berasal dari fitur segmentasi GEI pada Casia B Dataset, yang merupakan citra 2D skala abu-abu dari orang-orang yang berjalan dengan berbagai sudut pandang. Penelitian ini menunjukkan bahwa dengan menggunakan metode Gabor Wavelet untuk ekstraksi fitur dan PCA untuk reduksi dimensi, akurasi rata-rata dari sudut pandang 0º hingga 180º mencapai 98,50%. Nilai K yang digunakan pada metode K-Nearest Neighbors (KNN) adalah 1, sesuai rekomendasi penelitian sebelumnya, karena nilai K yang lebih tinggi menyebabkan anomali yang berpotensi menghasilkan klasifikasi yang kurang akurat. Dalam penelitian yang menggunakan optimasi fitur Gabor Wavelet dan PCA untuk pengenalan gait, akurasi rata-rata pengenalan menggunakan algoritma KNN meningkat menjadi 98,50% dari nilai akurasi sebelumnya yang tertinggi yaitu 97,39% yang dicapai oleh metode Group Lasso dan CDA dengan KNN. Ini menunjukkan peningkatan sebesar 1,11% setelah menggunakan optimasi PCA dalam kombinasi dengan fitur Gabor Wavelet. Penelitian ini merekomendasikan bahwa algoritma K-Nearest Neighbors (KNN) dapat ditingkatkan akurasinya dengan mengoptimalkan fitur yang digunakan. Dicapai melalui ekstraksi fitur dengan Gabor Wavelet dan reduksi dimensi menggunakan Principal Component Analysis (PCA).

2.2. Drop-out pada Perguruan Tinggi

Drop-out atau putus sekolah merupakan isu yang sering terjadi di perguruan tinggi dan memiliki dampak serius pada institusi pendidikan. Tingginya angka mahasiswa yang berhenti studi dapat menurunkan reputasi universitas serta memengaruhi akreditasi dan kualitas lulusan yang dihasilkan (Utari et al, 2020). Berdasarkan data dari Buku Statistik Pendidikan Tinggi 2022, angka putus kuliah pada perguruan tinggi di Indonesia menunjukkan variasi yang signifikan antar provinsi. Beberapa provinsi dengan tingkat drop out yang tinggi mencakup Jawa Timur sebesar 14,84%, Jawa Barat sebesar 10,57%, dan DKI Jakarta sebesar 12,82%. Di sisi lain, beberapa provinsi dengan angka putus kuliah lebih rendah mencakup Bangka Belitung dengan 0,14% dan Kalimantan Utara sebesar 0,13%. Secara nasional, angka putus kuliah di tingkat perguruan tinggi di Indonesia mencerminkan tantangan besar dalam mempertahankan kelulusan mahasiswa. Data ini dapat menunjukkan berbagai faktor yang mempengaruhi angka putus kuliah, termasuk akses pendidikan, ekonomi lokal, dan kualitas serta manajemen institusi pendidikan di masing-masing provinsi.

Di Jawa Timur, angka drop out atau putus kuliah tertinggi di Indonesia, yaitu 14,84%, dapat dikaitkan dengan berbagai faktor. Beberapa penyebab potensial mencakup kondisi ekonomi mahasiswa, keterbatasan akses dukungan finansial, dan mungkin adanya tekanan akademik yang tinggi. Selain itu, faktor sosial seperti kurangnya dukungan lingkungan atau keterbatasan layanan pendukung di kampus juga dapat berperan dalam meningkatkan angka putus kuliah di wilayah ini. Beberapa faktor umum yang menyebabkan drop out di antaranya adalah kondisi ekonomi, tekanan akademik, serta kurangnya dukungan dari keluarga dan lingkungan kampus. Faktor ekonomi menjadi salah satu pemicu utama di mana mahasiswa dari keluarga berpenghasilan rendah memiliki kemungkinan lebih tinggi untuk tidak menyelesaikan studi mereka karena ketidakmampuan dalam membiayai pendidikan (Astin, 1984).

Drop out atau putus kuliah adalah fenomena ketika seorang mahasiswa berhenti mengikuti proses pendidikan di perguruan tinggi sebelum menyelesaikan masa studinya. Kondisi ini dapat disebabkan oleh berbagai faktor, baik dari aspek internah mahasiswa maupun lingkungan eksternal, seperti kondisi ekonomi, dukungan sosial, dan tuntutan akademik. Menurut Astin (1984), kondisi sosial-sekonomi yang rendah dan keterbatasan akses ke fasilitas pendukung juga sering kali menjadi hambatan bagi kelangsungan studi mahasiswa. Strategi pencegahan untuk mengurangi tingkat drop out ini mencakup peningkatan kualitas layanan konseling dan bimbingan, program orientasi yang membantu mahasiswa beradaptasi dengan kehidupan kampus, serta program bantuan keuangan untuk mahasiswa dari latar belakang ekonomi rendah. Tindakan ini dapat membantu mahasiswa mengatasi hambatan yang mereka hadapi selama proses perkuliahan dan mendorong kelanjutan studi hingga lulus

(Kementerian Pendidikan dan Kebudayaan RI, 2019).

Melalui analisis data akademik, seperti IPK Semester, jumlah SKS yang diambil, dan prestasi akademik, algoritma dapat membantu mengidentifikasi pola yang berkaita dengan risiko putus kuliah di kalangan mahasiswa. Langkah pencegahan untuk mengurangi angka drop out melibatkan dukungan akademik yang lebih baik, program konseling, dan bantuan keuangan. Prediksi dini terhadap mahasiswa yang berisiko tinggi juga sangat penting agar pihak universitas dapat merancang kebijakan strategis yang tepat, baik dalam menyediakan dukungan finansial maupun layanan bimbingan akademik untuk mendukung keberhasilan studi mahasiswa secara optimal (Utari et al., 2020).

2.3. Machine Learning

Machine learning adalah subbidang dari kecerdasan buatan yang berfokus pada pengembangan algoritma yang memungkinkan komputer untuk belajar dan membuat prediksi berdasarkan data yang tersedia. Menurut Thomas dan Gupta (2020), machine learning membantu dalam menganalisis data dan membangun model prediktif yang dapat digunakan dalam pengambilan keputusan. Proses ini sering kali disebut sebagai "belajar dari data," di mana data menjadi materi pembelajaran utama. Metode pembelajaran dalam machine learning umumnya dibagi menjadi 3 kategori utama :

1. **Supervised Learning**

   Supervised learning adalah metode pembelajaran yang terawasi, di mana program diberikan contoh data yang sudah diberi label sebagai materi pelatihan. Dalam pendekatan ini, model belajar untuk memetakan input ke output berdasarkan contoh yang diberikan. Misalnya, jika model dilatih dengan gambar bangku dan meja yang sudah dilabeli, maka setelah proses pelatihan, model diharapkan mampu mengklasifikasikan objek baru ke dalam kategori yang tepat (Thomas & Gupta, 2020). Metode ini sering menggunakan training set dan test set untuk mengukur performa program.

2. **Unsupervised Learning**

   Unsupervised learning adalah metode pembelajaran yang tidak terawasi, di mana program mencoba menemukan pola dari data tanpa adanya label klasifikasi. Pendekatan ini bertujuan untuk menemukan struktur dalam data, seperti pengelompokan atau clustering. Berbeda dengan supervised learning, metode ini tidak memiliki target output yang telah ditentukan sebelumnya (Thomas & Gupta, 2020).

3. **Reinforcement Learning**

   Reinforcement learning adalah metode pembelajaran di mana program belajar melalui interaksi dengan lingkungan yang dinamis.

Dalam pendekatan ini, program mendapatkan ganjaran (reward) ketika membuat keputusan yang benar dan hukuman (punishment) ketika membuat kesalahan. Pengalaman dari interaksi tersebut akhirnya membentuk pemahaman yang lebih baik, dan program dapat membuat keputusan yang lebih cerdas di masa depan berdasarkan pola-pola yang telah dipelajarinya

2.4. Data Mining

Big Data mengubah jenis data yang karena terdiri dari data yang melampaui data terstruktur yang direpresentasikan dalam tabel relasional. Seperti yang disebutkan sebelumnya, Big Data mencakup data dari semua jenis termasuk teks, audio, video, aliran klik, berkas log, dan banyak lagi. Big Data mungkin sensitif terhadap waktu dan oleh karena itu mungkin harus digunakan saat mengalir ke perusahaan untuk memaksimalkan nilainya bagi bisnis. Teknologi penambangan data memungkinkan kita untuk melihat dan mengukur berbagai hal seperti yang belum pernah terjadi sebelumnya; seperti mikroskop yang memungkinkan para ilmuwan untuk memeriksa misteri kehidupan di tingkat seluler. Beberapa orang mengklaim bahwa Big Data akan membuka pintu untuk membuat keputusan yang lebih cerdas di hampir setiap bidang. Ini akan membuat masyarakat modern menjadi masyarakat yang digerakkan oleh data. Penambangan data atau penemuan pengetahuan dalam basis data (KDD) adalah kumpulan teknik eksplorasi yang didasarkan pada metode dan alat analitis tingkat lanjut untuk menangani sejumlah besar informasi. Teknik-teknik tersebut dapat menemukan pola-pola baru yang dapat membantu perusahaan dalam memahami bisnis dengan lebih baik dan dalam peramalan. Banyak teknik penambangan data yang terkait erat dengan beberapa teknik pembelajaran mesin yang telah dikembangkan selama 50 tahun terakhir. Teknik-teknik lainnya terkait dengan teknik-teknik yang telah dikembangkan dalam statistik, terkadang disebut analisis data eksploratori. Teknik-teknik ini dikembangkan beberapa waktu lalu dan dirancang untuk menangani sejumlah kecil data. Teknik-teknik tersebut kini telah dimodifikasi untuk menangani sejumlah besar data. Namun, teknik-teknik lainnya tergolong relatif baru, misalnya penambangan data web.

Data Mining adalah kumpulan teknik untuk penemuan otomatis yang efisien dari pola-pola yang sebelumnya tidak diketahui, valid, baru, berguna, dan dapat dipahami dalam basis data besar. Pola-pola tersebut harus dapat ditindaklanjuti sehingga dapat digunakan dalam pengambilan keputusan institusi. Diasumsikan sedang erurusan dengan sejumlah besar data. Teknik data mining dapat digunakan untuk data dalam jumlah yang lebih kecil, tetapi semakin besar datanya, semakin besar peluang untuk menemukan sesuatu yang baru dan menarik. Selain itu, jika sesuatu yang menarik ditemukan, sejumlah besar data memberikan keyakinan dalam penemuan tersebut. Dengan kata lain, memastikan bahwa apa yang telah ditemukan memang merupakan sesuatu yang menarik tentang proses yang mendasari data yang sedang dianalisis dan bukan sekadar hasil dari

beberapa fluktuasi acak. Kami ingin memastikan bahwa pola baru yang telah kami temukan sebenarnya konsisten dan dapat direplikasi.

Data Mining dapat digunakan untuk membuat deskripsi dan prediksi. Penting bagi institusi untuk dapat memprediksi beberapa aspek bisnis guna membantu perencanaan masa depan. Bahkan di universitas, prediksi atau peramalan jumlah mahasiswa untuk tahun berikutnya memegang peranan penting dalam perencanaan. Data mining dalam institusi pendidikan sangatlah penting karena beberapa peneliti telah mengembangkan model prediksi untuk tingkat drop-out siswa karena tingkat drop-out merupakan perhatian penting institusi pendidikan. Sebuah studi yang dilakukan di Belanda menunjukkan bahwa pengklasifikasi yang cukup sederhana memberikan hasil yang berguna dengan akurasi antara 75 dan 80% yang lebih baik daripada model lain yang lebih canggih. Studi tersebut menemukan bahwa prediktor keberhasilan yang terkuat adalah nilai untuk mata kuliah Aljabar Linear, yang secara umum tidak dianggap sebagai mata kuliah yang menentukan. Prediktor kuat lainnya adalah nilai untuk Kalkulus dan Jaringan.

Studi data mining di bidang pendidikan lainnya menemukan bahwa eksplorasi data yang difokuskan pada jumlah latihan yang dicoba dikombinasikan dengan klasifikasi mampu mengidentifikasi siswa yang berisiko. Pengelompokan dan visualisasi klaster mampu mengidentifikasi perilaku tertentu di antara siswa yang gagal. Peringatan yang tepat waktu dan tepat kepada siswa yang berisiko mampu membantu banyak siswa lulus dalam ujian akhir (G.K. Gupta, 2014)

2.5. Klasifikasi

Klasifikasi merupakan metode utama dalam machine learning yang bertujuan untuk mengkategorikan data ke dalam kelas tertentu berdasarkan parameter yang sudah diketahui sebelumnya. Salah satu metode yang sering digunakan untuk klasifikasi adalah algoritma *K-Nearest Neighbors* (KNN). KNN bekerja dengan cara membandingkan data yang baru dengan data yang sudah ada berdasarkan kedekatan jarak di antara mereka. Setiap data baru akan dikelompokkan ke dalam kelas yang paling banyak muncul di antara tetangganya (Jayasri & Aruna, 2022).

Pada dasarnya, algoritma KNN adalah bagian dari *supervised learning*, di mana sistem memerlukan dataset berlabel untuk melatih model. Data berlabel ini berfungsi sebagai contoh untuk menentukan kelas dari data yang belum diketahui. Proses klasifikasi dalam KNN akan menghitung jarak antara titik data baru dengan setiap data yang sudah ada dalam dataset, biasanya menggunakan metrik seperti jarak Euclidean atau Manhattan. Dengan metode ini, KNN efektif untuk masalah klasifikasi sederhana namun menjadi lambat ketika dataset berukuran besar

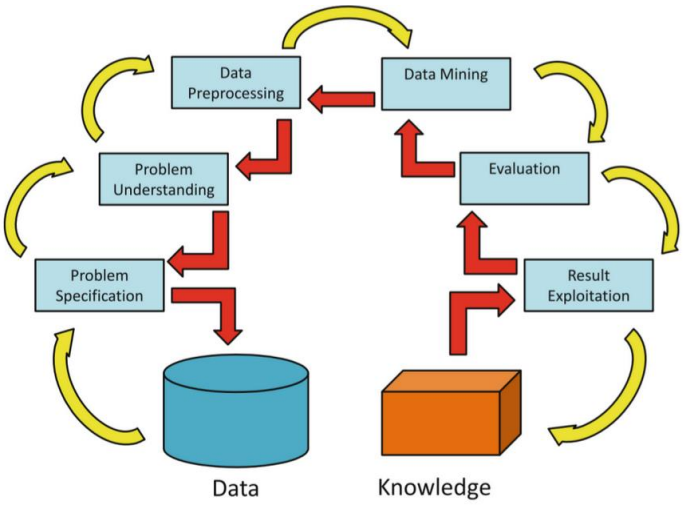
Dalam konteks yang lebih luas, teknik KNN sering digabungkan dengan metode lainnya, seperti analisis komponen utama (PCA) untuk mengurangi dimensi data, atau algoritma *outlier detection* untuk

meningkatkan akurasi. Gabungan ini memungkinkan KNN bekerja lebih efisien, terutama pada data berskala besar yang sering ditemui dalam analisis *big data*. Misalnya, penelitian oleh Jayasri dan Aruna (2022) menggabungkan KNN dengan framework MapReduce untuk memproses data dalam jumlah besar, khususnya di bidang kesehatan, di mana data pasien diabetes dapat diklasifikasikan lebih akurat dengan pendekatan berbasis aturan asosiasi dan deteksi anomali (outlier) yang dikembangkan lebih lanjut

Metode klasifikasi KNN telah digunakan dalam berbagai aplikasi untuk prediksi dan optimisasi, termasuk untuk deteksi penyakit, pengelompokan pelanggan, dan rekomendasi produk. Kombinasi KNN dengan teknik modern seperti jaringan perhatian hirarkis (hierarchical attention network) memberikan fleksibilitas dalam mengolah data yang lebih kompleks, seperti yang terlihat pada penelitian big data di bidang kesehatan (Jayasri & Aruna, 2022).

2.6. Pre-processing Data

Dalam buku berjudul “Data Preprocessing in Data Mining” yang di tulis oleh Salvador Garcia dkk, disebutkan bahwa metode data mining ada beberapa langkah, yaitu :



Proses di atas di sebut juga dengan KDD Process atau Knowledge Discovery in Database. KDD Process adalah proses untuk mengidentifikasi pola yang valid, baru, berpotensi berguna, dan pada akhirnya dapat dipahami dalam data. Aspek kunci yang menjadi ciri khas proses KDD adalah cara membaginya kedalam beberapa tahap. Ada 6 langkah yang di sebutkan dalam buku ini, salah satunya adalah **Data Preprocessing.**

Data Preprocessing mencakup operasi untuk pembersihan data (seperti menangani penghapusan noise dan data tidak konsisten), integrasi data (dimana beberapa sumber data dapat digabungkan menjadi satu), transformasi data (dimana data diubah dan dikonsolidasikan kedalam bentuk yang sesuai untuk tugas Data Mining tertentu atau operasi agregasi) dan reduksi data, termasuk pemilihan dan ekstraksi fitur.

Data masukan harus di sediakan dalam jumlah, struktur, dan format yang sesuai dengan setiap tugas Data Mining secara sempurna. Sayangnya, basis data di dunia nyata sangatlah dipengaruhi oleh faktor-faktor negatif seperti adanya noise, missing values, data yang tidak konsisten, dan berlebihan, serta ukuran yang sangat besar dimensinya. Dengan demikian, data berkualitas rendah akan menghasilkan konerja Data Mining berkualitas rendah.

2.7. Principal Component Analysis (PCA)

Dalam buku berjudul "Machine Learning and Data Mining in Pattern Recognition" yang dipublish oleh Springer Berlin Heidelberg, di sebutkan bahwa PCA adalah metode analisis statistik multivariat yang sangat umum digunakan untuk menemukan atribut yang sangat berkorelasi dan mengurangi dimensionalitas. Idenya adalah mengidentifikasi K komponen utama untuk data set D-Dimensi (K << D) yang menjelaskan sebagian besar varians dataset, misalnya lebih dari 80%, yang memungkinkan pengurangan dimensionalitas dataset dari dimensi D ke dimensi K tanpa banyak kehilangan informasi. PCA banyak digunakan dalam Data Mining dan beberapa metode pengelompokan data berbasis PCA telah di kembangkan di masa lalu. Dalam buku ini juga di sebutkan bahwa Principal Component Analysis (PCA) merupakan salah satu metode linier yang paling diterima untuk mengekstrak informasi relevan dari kumpulan data berdimensi tinggi. Metode ini mengurangi kompleksitas kumpulan data ke dimensi yang lebih rendah untuk mengungkap struktur tersembunyi dan sederhana yang sering kali mendasarinya. PCA memproyeksikan data dalam dimensi yang lebih rendah di sepanjang vektor arah yang relevan.

2.8. K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (KNN) merupakan salah satu algoritma yang paling banyak digunakan dalam klasifikasi data karena pendekatannya yang non-parametrik dan fleksibel (Zhang, 2021). Algoritma ini bekerja dengan cara menentukan kelas suatu data uji berdasarkan mayoritas kelas dari tetangga terdekatnya dalam data latih. Setiap objek dalam data uji diprediksi kelasnya berdasarkan K tetangga terdekatnya menggunakan aturan mayoritas atau aturan pembobotan jarak. KNN sangat efektif untuk data dengan distribusi yang bervariasi dan sering digunakan dalam klasifikasi yang tidak memerlukan pembentukan model (Zhang, 2021).

Proses KNN terbagi menjadi dua fase utama, yaitu fase pelatihan dan prediksi. Pada fase pelatihan, nilai K yang optimal ditentukan menggunakan teknik validasi silang seperti k-fold cross-validation untuk meningkatkan akurasi prediksi. Selanjutnya, pada fase prediksi, KNN mencari K tetangga terdekat dalam data latih untuk setiap data uji, dengan menggunakan aturan mayoritas untuk menentukan kelas data tersebut (Zhang et al., 2017).

Tantangan utama dalam KNN meliputi penentuan nilai K yang

tepat, pemilihan tetangga terdekat yang optimal, serta pemilihan fungsi jarak yang sesuai. Nilai K yang optimal tidak selalu seragam untuk semua data uji karena distribusi data yang bervariasi. Beberapa penelitian mengusulkan untuk menetapkan nilai K yang berbeda pada sub-ruang sampel atau pada setiap data uji individu untuk meningkatkan akurasi prediksi (Zhang et al., 2017). Selain itu, pemilihan tetangga terdekat memerlukan pengukuran jarak yang efektif agar dapat mengatasi ketidakseimbangan kelas pada data latih (Song et al., 2007). Algoritma KNN juga dikategorikan sebagai metode "lazy learning" karena hanya membutuhkan perhitungan jarak pada saat prediksi, bukan pada saat pelatihan, sehingga lebih fleksibel dalam menangani berbagai jenis data.

2.9. Metrik Performa (Evaluasi Konerja Model dan Sistem API)

a. Akurasi (Accuracy)

Akurasi mengukur persentase prediksi yang benar dibandingkan dengan semua data uji.

$$\text{Akurasi} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Data (TP + TN + FP + FN)}}$$

TP (True Positive): Prediksi benar untuk kelas positif

TN (True Negative): Prediksi benar untuk kelas negatif

FP (False Positive): Prediksi salah untuk kelas positif

FN (False Negative): Prediksi salah untuk kelas negatif

b. Presisi (Precision)

Presisi menunjukkan keakuratan prediksi positif yang dibuat oleh model.

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

c. Recall (Sensitivitas atau True Positive Rate)

Recall mengukur proporsi total kasus positif yang berhasil dikenali dengan benar oleh model.

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

d. F1 Score

F1 Score merupakan rata-rata harmonis antara presisi dan recall, bermanfaat dalam situasi di mana keseimbangan antara keduanya diperlukan.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

e. Mean Absolute Error (MAE)

MAE adalah rata-rata dari nilai absolut perbedaan antara prediksi dan nilai aktual, sering digunakan dalam model regresi.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

yi: Nilai aktual

y^i: Nilai prediksi

n: Jumlah data

f. Mean Squared Error (MSE)

MSE adalah rata-rata dari kuadrat perbedaan antara nilai prediksi dan nilai aktual.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

g. Root Mean Squared Error (RMSE)

RMSE adalah akar dari MSE, mengukur jarak antara prediksi dan nilai aktual dalam satuan yang sama.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

h. R-squared (R^2) atau Koefisien Determinasi

R^2 mengukur seberapa besar variabilitas data yang dapat dijelaskan oleh

model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$\bar{y}$: Nilai rata-rata dari yi

Metrik-metrik ini membantu dalam menilai dan memilih model yang sesuai dengan kebutuhan dataset tertentu (Rafael et al, 2020)

2.10. Uji Coba Model

Uji coba model adalah proses evaluasi performa suatu model pembelajaran mesin dengan menggunakan data tertentu untuk menilai efektivitas dan akurasinya. Proses ini umumnya melibatkan pembagian dataset menjadi data pelatihan dan data pengujian, di mana data pelatihan digunakan untuk melatih model, dan data pengujian untuk mengevaluasi hasil prediksi model tersebut (Karal, 2020). Salah satu metode yang sering digunakan dalam uji coba model adalah k-fold cross-validation, di mana data dibagi menjadi beberapa subset atau "fold" yang bergantian digunakan sebagai data uji dan data latih.

Tujuan utama dari uji coba model adalah memastikan bahwa model memiliki kemampuan generalisasi yang baik ketika diterapkan pada data baru atau tidak dikenal. Proses ini juga berguna untuk mengidentifikasi parameter model yang optimal, terutama pada algoritma berbasis kernel seperti Support Vector Machine (SVM). Dalam studi tertentu, variasi kernel pada SVM diuji dengan nilai "k" yang berbeda, menunjukkan bahwa kernel Gaussian dan Linear menghasilkan akurasi terbaik pada kondisi tertentu (Karal, 2020).

2.10.1. K-Fold Cross Validation

K-Fold Cross Validation adalah metode validasi dalam pembelajaran mesin yang bertujuan untuk mengukur performa model dengan membagi dataset menjadi beberapa bagian atau “folds” (Karal, 2020). Dalam metode ini, data dibagi menjadi sejumlah “k” subset yang sama besar, di mana pada setiap iterasi, satu subset digunakan sebagai data uji, sementara sisanya digunakan untuk melatih model. Proses ini diulang hingga setiap subset menjadi data uji tepat satu kali, dan rata-rata hasilnya dihitung untuk menentukan akurasi keseluruhan model.

K-Fold Cross Validation merupakan teknik yang populer untuk mengurangi kemungkinan overfitting pada model pembelajaran mesin karena memanfaatkan keseluruhan data secara bergantian sebagai data uji (Karal, 2020). Selain itu, pemilihan nilai "k" yang tepat sangat penting, karena nilai "k" yang lebih kecil cenderung menghasilkan varian yang lebih tinggi pada estimasi performa, sedangkan nilai "k" yang lebih besar membutuhkan waktu komputasi lebih lama (Karal, 2020).

Menurut penelitian, variasi kernel pada algoritma Support Vector Machine (SVM) dapat menunjukkan hasil akurasi yang berbeda untuk nilai "k" yang beragam dalam K-Fold Cross Validation. Misalnya, kernel Gaussian dan Linear pada SVM memberikan hasil yang paling akurat dalam beberapa dataset tertentu, terutama saat k bernilai 10 (Karal, 2020).

Berikut merupakan langkah-langkah uji coba model menurut Karal (2020):

1) Memilih metode pembagian data (contohnya, k-fold cross-validation).

2) Melatih model pada data pelatihan.

3) Menguji model pada data uji.

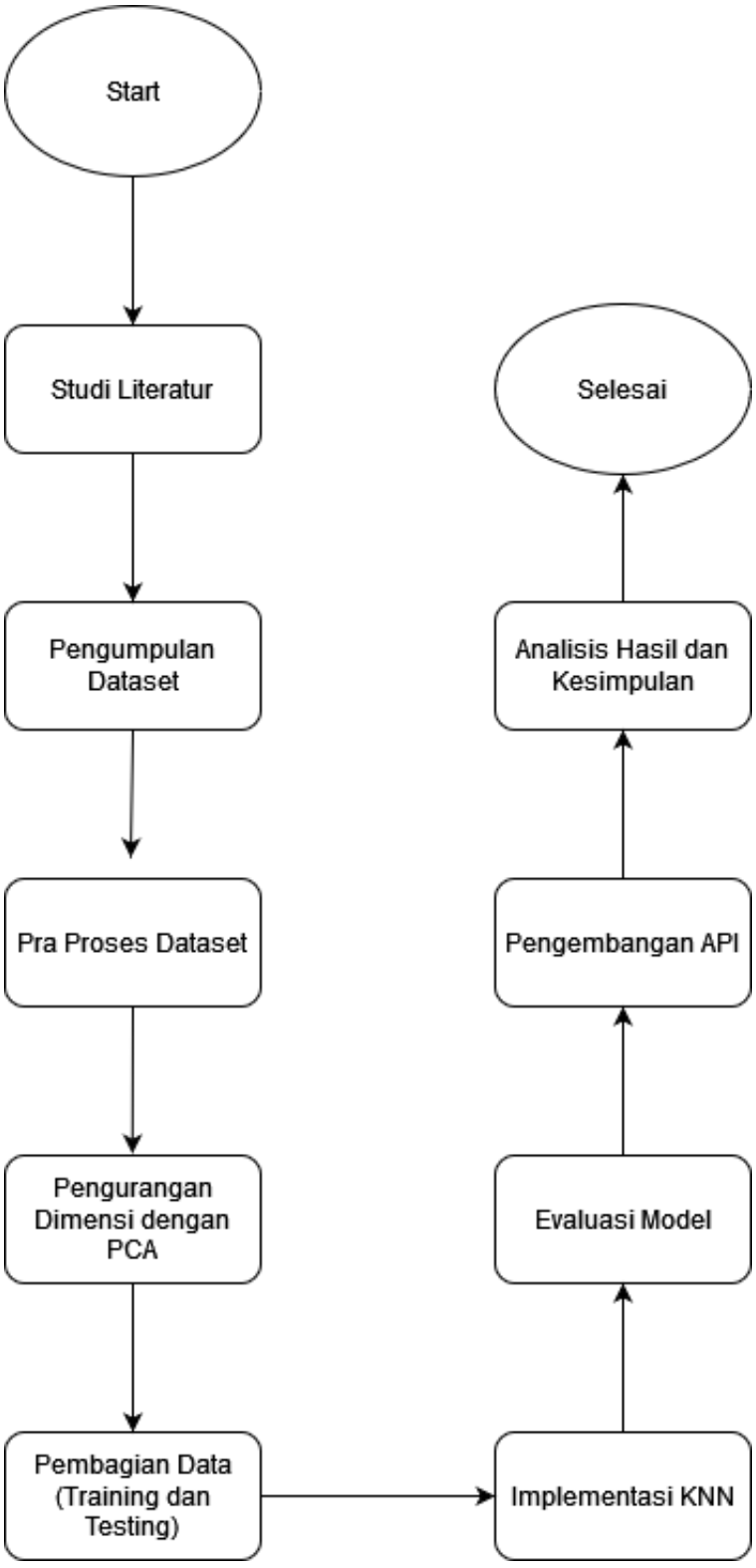
4) Menghitung rata-rata akurasi untuk semua fold.

Uji coba model dengan pendekatan ini penting untuk memastikan bahwa model memiliki performa yang stabil dan andal pada berbagai skenario data.

BAB III
METODOLOGI PENELITIAN

3.1. Tahapan Penelitian

Penelitian ini disusun melalui serangkaian langkah kerja yang sistematis untuk memastikan bahwa setiap tahap penelitian dilakukan dengan baik dan terarah. Langkah-langkah ini dirancang untuk menjawab rumusan masalah dan mencapai tujuan penelitian yang telah ditetapkan pada bab sebelumnya. Tahapan penelitian yang dilakukan meliputi studi literatur, pengumpulan data, eksplorasi data, penerapan metode analisis, hingga pengujian model.


Start
Studi Literatur
Pengumpulan Dataset
Pra Proses Dataset
Pengurangan Dimensi dengan PCA
Pembagian Data (Training dan Testing)
Implementasi KNN
Evaluasi Model
Pengembangan API
Analisis Hasil dan Kesimpulan
Selesai

3.2. Studi Literatur

Literatur yang telah dikaji pada bagian penelitian terkait akan menjadi dasar teori atas studi dan penelitian yang dilakukan agar dapat menjadi landasan teori atas dilaksanakannya penelitian ini. Konsep dan dasar teori yang digunakan dapat meliputi sumber buku, jurnal dari penelitian yang telah dilakukan sebelumnya yang relevan. Referensi dari studi literatur yang telah dipelajari dan digunakan pada penelitian ini dilampirkan pada daftar pustaka di bagian akhir laporan skripsi. Studi literatur berfungsi sebagai landasan teoritis dalam penelitian ini. Sumber yang digunakan mencakup buku, jurnal ilmiah, serta penelitian terdahulu yang relevan. Studi ini bertujuan untuk memahami konsep dan teori yang mendasari algoritma K-Nearest Neighbors (KNN) dan Principal Component Analysis (PCA), serta penerapannya dalam klasifikasi. Referensi yang digunakan dicantumkan dalam daftar pustaka pada akhir laporan.

3.3. Pengumpulan Dataset

Penelitian ini menggunakan data primer yang diperoleh dari Biro Akademik, Kemahasiswaan, Perencanaan, dan Kerjasama di UPN "Veteran" Jawa Timur. Dataset mencakup informasi karakteristik mahasiswa seperti IPK, jumlah SKS, penghasilan orang tua, jenis tempat tinggal, tingkat kehadiran, dan jurusan. Data dikumpulkan dari beberapa fakultas untuk memperoleh representasi yang komprehensif dari mahasiswa angkatan 2021 dan 2022. Fakultas yang terlibat dalam penelitian ini mencakup :

- Fakultas Ekonomi dan Bisnis dengan program studi di bidang Ekonomi Pembangunan, Akuntansi, Manajemen, dan Kewirausahaan.
- Fakultas Pertanian yang meliputi program studi Agroteknologi dan Agribisnis
- Fakultas Teknik dan Sains yang terdiri dari program studi Teknik Kimia, Teknik Industri, Teknik Lingkungan, Teknologi Pangan, Teknik Mesin, dan Fisika.
- Fakultas Ilmu Komputer yang mencakup Teknik Informatika, Sistem Informasi, Data Sains, dan Bisnis Digital.
- Fakultas Ilmu Sosial dan Ilmu Politik dengan program studi Administrasi Negara, Administrasi Bisnis, Ilmu Komunikasi, Hubungan Internasional, Pariwisata, dan Lunguistik Indonesia.
- Fakultas Arsitektur dan Desain yang terdiri dari program studi Arsitektur, Desain Komunikasi Visual, dan Desain Interior.
- Fakultas Hukum dengan program studi Ilmu Hukum.
- Fakultas Kedokteran dengan program studi Kedokteran.

Data ini dikumpulkan untuk memahami karakteristik mahasiswa dari beragam program studi dalam universitas.

Setelah data terkumpul, tahap berikutnya adalah Eksplorasi Data Awal (Exploratory Data Analysis/EDA). EDA bertujuan untuk memahami pola dalam data, mengidentifikasi hubungan antar variabel, serta menemukan karakteristik mahasiswa yang berpotensi drop-out. Visualisasi data seperti distribusi variabel, analisis korelasi, serta perbandingan karakteristik antar fakultas digunakan untuk mendukung analisis awal ini. Hasil EDA memberikan wawasan penting yang menjadi dasar pengembangan model prediksi di tahap selanjutnya.

Pada penelitian ini dilakukan Eksplorasi Data Awal atau Exploratory Data Analysis (EDA) yang akan mempermudah untuk memahami pola dalam data dan mendapatkan wawasan awal mengenai variabel-variabel yang mungkin berpengaruh. Untuk plot distribusi dari variabel utama, terdapat distribusi nilai IPK, distribusi SKS, dan distribusi penghasilan orang tua. Kemudian akan dilakukan analisis korelasi untuk melihat bagaimana hubungan antar variabel, seperti korelasi antara IPK dengan jumlah SKS dan penghasilan orang tua. Dilakukan juga segmentasi berdasarkan fakultas guna membandingkan karakteristik mahasiswa yang berpotensi drop-out di berbagai fakultas, seperti rata-rata IPK dan kehadiran per fakultas. Visualisasi data akan menggunakan grafik seperti boxplot atau heatmap untuk menunjukkan korelasi dan distribusi variabel utama. Dengan adanya EDA, diharapkan dapat memberikan wawasan yang lebih mendalam mengenai faktor-faktor yang berkontribusi terhadap potensi drop-out, yang akan menjadi dasar untuk tahap analisis lanjutan dalam model prediksi.

Reduksi dimensi menggunakan Principal Component Analysis (PCA) digunakan untuk meningkatkan performa algoritma K-Nearest Neighbors, PCA digunakan untuk mengurangi dimensi data sambil mempertahankan informasi penting. PCA akan mengidentifikasi komponen utama yang menjelaskan variasi terbesar dalam data. Variabel yang dihasilkan dari PCA kemudian akan digunakan dalam model Klasifikasi K-Nearest Neighbors untuk memprediksi potensi drop-out.

Setelah model dikembangkan dan diuji, dibuat API yang dapat diakses untuk klasifikasi karakteristik mahasiswa berpotensi drop-out. API ini memungkinkan integrasi hasil analisis dengan sistem informasi kampus, sehingga dapat digunakan untuk membantu dalam mengambil keputusan terkait potensi drop-out mahasiswa.

Data ini dikumpulkan melalui permohonan dari Biro Akademik, Kemahasiswaan, Perencanaan, dan Kerjasama di UPN “Veteran” Jawa Timur. Berikut adalah deskripsi data yang digunakan dalam penelitian ini beserta tujuan dan alasan pemilihan setiap kolomnya.

Kolom Data	**Keterangan**	**Alasan Pemilihan**
Nilai IPK	Nilai Indeks Prestasi	IPK merupakan indikator

	Kumulatif (IPK) mahasiswa pada akhir semester terakhir yang diambil	kinerja akademik yang penting dalam menentukan potensi risiko drop-out mahasiswa.
Jumlah SKS	Jumlah Satuan Kredit Semester (SKS) yang telah ditempuh oleh mahasiswa	Jumlah SKS memberikan gambaran tentang beban akademik yang diambil mahasiswa, yang dapat mempengaruhi performa mereka.
Penghasilan Orang Tua	Pendapatan bulanan orang tua mahasiswa	Status ekonomi dapat berpengaruh pada ketahanan mahasiswa dalam menyelesaikan studinya, terutama dalam menghadapi kesulitan.
Jenis Tempat Tinggal	Status tempat tinggal mahasiswa, misalnya kos, asrama, rumah sendiri	Tempat tinggal dapat mempengaruhi fokus dan konsistensi kehadiran mahasiswa dalam aktivitas akademik.
Kehadiran	Tingkat kehadiran mahasiswa dalam perkuliahan	Kehadiran menunjukkan komitmen mahasiswa terhadap kegiatan akademik dan dapat berpengaruh pada prestasi akademik.
Jurusan	Program studi atau jurusan yang diambil oleh mahasiswa	Jurusan dapat memengaruhi pola studi dan kinerja mahasiswa berdasarkan tuntutan akademik yang berbeda-beda di tiap jurusan.

3.4. Dataset

Dataset yang digunakan sebanyak satu *dataset* yang diperoleh secara publik dalam website Kaggle. *Dataset* tersebut bernama University_Student_Profiles. *Dataset* tersebut berisikan sebanyak xx data mahasiswa dengan informasi lengkap dari ID, tahun masuk, jumlah cuti kuliah, umur, pendapatan, IPK persiapan, jurusan

SMA, IPK, jalur masuk, semester yang sudah dijalani, nilai TPA, status, dan fakultas.

Dataset berformat ".csv" yang memiliki ukuran a x b tabel. Oleh karena itu dilakukan *Cleaning Data* agar tidak ada data yang tidak legkap atau kurang valid dalam penelitian nantinya. Pada Tabel x.x merupakan potongan data dari dataset *data_output_cleaned.csv*

Tabel x.x Isi *dataset data_output_cleaned.csv*

No	ID	TahunMasuk	Prodi	JenisKelamin	JumlahCuti
1	ID00001	2021.0	S-1 TEKNIK ELEKTRO	L	0.0
2	ID00002	2021.0	S.Tr. TEKNOLOGI REKAYASA OTOMASI	L	0.0
3	ID00003	2021.0	S-1 FISIKA	P	0.0
4	ID00004	2020.0	S-1 BIOLOGI	P	0.0
5	ID00005	2020.0	S-1 SISTEM INFORMASI	L	0.0
6	ID00006	2019.0	S-1 TEKNIK FISIKA	P	0.0
7	ID00007	2019.0	S-1 TEKNIK ELEKTRO	L	0.0
8	ID00008	2021.0	S-1 SISTEM INFORMASI	P	0.0
9	ID00009	2019.0	S-1 TEKNIK INFORMATIKA	L	0.0
10	ID00010	2020.0	S-1 TEKNIK INDUSTRI	L	0.0
11	ID00011	2021.0	S.Tr. TEKNOLOGI REKAYASA OTOMASI	L	0.0
12	ID00012	2021.0	S.Tr. TEKNOLOGI REKAYASA KONVERSI ENERGI	L	0.0
13	ID00013	2020.0	S-1 TEKNIK SISTEM PERKAPALAN	L	0.0
14	ID00014	2019.0	S-1 TEKNIK SIPIL	P	0.0
15	ID00015	2021.0	S-1 FISIKA	L	0.0
16	ID00016	2021.0	S-1 TEKNIK BIOMEDIK	L	0.0

17	ID00017	2020.0	S-1 TEKNIK INFORMATIKA	P	0.0
18	ID00018	2021.0	S-1 PERENCANAAN WILAYAH DAN KOTA	P	0.0
19	ID00019	2020.0	S-1 TEKNIK SISTEM PERKAPALAN	L	0.0
20	ID00020	2019.0	S-1 TEKNIK KOMPUTER	L	0.0
21	ID00021	2021.0	S-1 TEKNIK SIPIL	L	0.0
22	ID00022	2020.0	S-1 TEKNOLOGI INFORMASI	L	0.0
23	ID00023	2020.0	S-1 TEKNIK INFORMATIKA	L	0.0
24	ID00024	2020.0	S-1 KIMIA	L	0.0
25	ID00025	2019.0	S-1 TEKNIK ELEKTRO	L	0.0

No	Umur	Pendapatan	IPKPersiapan	JurusanSMA	IPK
1	21.0	Rp. 2.500.001 - Rp. 3.000.000	3,68	SMA/MA IPA	3.55
2	22.0	Rp. 1.500.001 - Rp. 2.000.000	1,42	SMA/MA IPA	1.42
3	22.0		3,15	SMA/MA IPA	3.05
4	23.0	Rp. 1.500.001 - Rp. 2.000.000	3,25	SMA/MA IPA	3.11
5	21.0	Lebih dari Rp. 15.000.000	3,56	SMA/MA IPA	3.67
6	23.0	Rp. 7.500.001 - Rp. 10.000.000	3,5	SMA/MA IPA	3.41
7	22.0	Rp. 5.000.001 - Rp.	3,56	SMA/MA IPA	3.79

		7.500.000			
8	19.0	Rp. 10.000.001 - Rp. 15.000.000	3,5	SMA/MA IPA	3.56
9	22.0	Rp. 500.001 - Rp. 1.000.000	3	SMA/MA IPA	3.24
10	21.0	Rp. 3.000.001 - Rp. 4.000.000	3,63	SMA/MA IPA	3.42
11	22.0	Rp. 2.500.001 - Rp. 3.000.000	3,36	SMA/MA IPA	3.35
12	21.0	Rp. 1.000.001 - Rp. 1.500.000	2,49	SMA/MA IPA	2.38
13	22.0	Rp. 5.000.001 - Rp. 7.500.000	3,26	SMA/MA IPA	3.17
14	22.0	Rp. 5.000.001 - Rp. 7.500.000	2,42	SMA/MA IPA	3.4
15	20.0	Rp. 7.500.001 - Rp. 10.000.000	3,46	SMA/MA IPA	3.31
16	20.0	Rp. 3.000.001 - Rp. 4.000.000	3,4	SMA/MA IPA	3.51
17	20.0	Rp. 10.000.001 - Rp. 15.000.000	3,8	SMA/MA IPA	3.67
18	20.0	Rp. 10.000.001 - Rp. 15.000.000	3,38	SMA/MA IPA	3.39
19	21.0	Rp. 10.000.001 - Rp. 15.000.000	3,47	SMA/MA IPA	3.42
20	22.0	Rp. 4.000.001 -	3,56	SMA/MA IPA	3.62

		Rp. 5.000.000			
21	20.0	Rp. 5.000.001 - Rp. 7.500.000	3,09	SMA/MA IPA	3.45
22	21.0	Rp. 10.000.001 - Rp. 15.000.000	3,88	SMA/MA IPA	3.86
23	21.0	Rp. 1.500.001 - Rp. 2.000.000	3,25	SMK Teknik	3.4
24	22.0	Rp. 4.000.001 - Rp. 5.000.000	3,72	SMK Teknik	3.76
25	22.0	Lebih dari Rp. 15.000.000	3,33	SMA/MA IPA	3.22

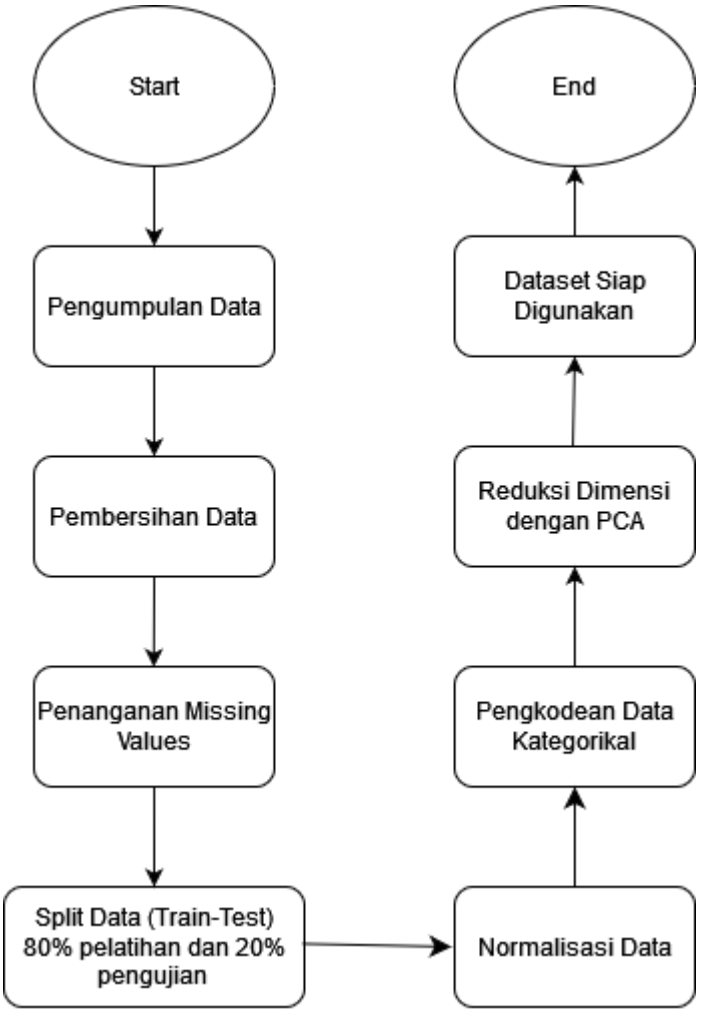
No	JalurMasuk	SemesterYangSudahDijalani	Nilai TPA	Status	Fakultas
1	SBMPTN	4.0	452.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
2	PKM Mandiri	4.0	550.0	Tidak DO	Fakultas Vokasi
3	PKM Mandiri	4.0	389.0	Tidak DO	Fakultas Seni
4	SNBT KIP-K	6.0	406.0	Tidak DO	Fakultas Seni
5	PKM Kemitraan	6.0	612.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
6	PKM Mandiri	8.0	582.0	Tidak DO	Fakultas Teknik Industri
7	SBMPTN	8.0	376.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
8	PKM Mandiri	4.0	479.0	Tidak DO	Fakultas Teknik Elektro dan Komputer

9	SNBP KIP-K	8.0	456.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
10	SBMPTN	6.0	596.0	Tidak DO	Fakultas Teknik Industri
11	SBMPTN	4.0	401.0	Tidak DO	Fakultas Vokasi
12	SNMPTN/SNBP	4.0	567.0	Tidak DO	Fakultas Vokasi
13	SNMPTN/SNBP	6.0	483.0	Tidak DO	Fakultas Teknik Kapal
14	SNMPTN/SNBP	8.0	547.0	Tidak DO	Fakultas Teknik Sipil dan Perencanaan
15	SNMPTN/SNBP	4.0	437.0	Tidak DO	Fakultas Seni
16	SNMPTN/SNBP	4.0	432.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
17	PKM Kemitraan	6.0	418.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
18	SBMPTN	4.0	466.0	Tidak DO	Fakultas Teknik Sipil dan Perencanaan
19	SBMPTN	6.0	364.0	Tidak DO	Fakultas Teknik Kapal
20	SBMPTN	8.0	492.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
21	SBMPTN	4.0	463.0	Tidak DO	Fakultas Teknik Sipil dan Perencanaan
22	PKM Mandiri	6.0	428.0	Tidak DO	Fakultas Teknik Elektro dan Komputer
23	SNBP KIP-K	6.0	579.0	Tidak DO	Fakultas Teknik Elektro dan Komputer

24	SBMPTN	6.0	374.0	Tidak DO	Fakultas Seni
25	SBMPTN	8.0	606.0	Tidak DO	Fakultas Teknik Elektro dan Komputer

3.5. Pra-Proses Data

Dari dataset mahasiswa akan dilakukan praproses data agar data tersebut seragam. Beberapa tahapan praproses yang akan dilakukan terdapat.



Pada tahap preprocessing, data mentah yang dikumpulkan dari sistem akademik diolah untuk memastikan data yang digunakan bersih dan siap untuk dianalisis. Langkah-langkah preprocessing yang dilakukan adalah sebagai berikut:

3.5.1. Pengumpulan Data

Mengumpulkan dataset mahasiswa yang mencakup atribut seperti IPS,

jumlah semester, IPK, SKS total, asal sekolah, pekerjaan orang tua, penghasilan orang tua, UKT, jenis tempat tinggal, ada pekerjaan sambilan apa tidak, kehadiran, dan jurusan.

3.5.2. Pembersihan Data

Menghapus data yang tidak valid, duplikat, atau yang memiliki nilai hilang (missing values) dan menangani nilai-nilai outliers yang dapat mempengaruhi model.

3.5.3. Penanganan Missing Value

Langkah pertama adalah memeriksa data untuk menemukan nilai yang kosong atau hilang (null values) pada kolom-kolom penting, seperti IPK, jumlah SKS, penghasilan orang tua, jenis tempat tinggal, kehadiran, dan jurusan. Jika terdapat nilai yang hilang pada dataset, lakukan pengisian nilai (imputation) atau metode lain yang sesuai menggunakan rata-rata atau median. Baris yang memiliki null values pada kolom-kolom ini akan dihapus untuk memastikan kualitas data dan menghindari bias atau ketidakakuratan saat pemodelan.

3.5.4. Membagi Data untuk Pelatihan dan Pengujian (Train-Test Split)

Setelah data bersih dari nilai-nilai kosong, data akan dibagi menjadi dua set: data pelatihan dan data pengujian. Perbandingan yang digunakan adalah 80% data untuk pelatihan dan 20% untuk pengujian. Pembagian data ini bertujuan untuk memvalidasi model klasifikasi agar dapat dievaluasi performanya pada data yang belum pernah dilihat oleh model.

3.5.5. Normalisasi Data atau Standarisasi Data Numerik (Scalarization)

Melakukan standardisasi data numerik agar semua fitur berada pada skala yang sama. Untuk data numerik seperti nilai IPK, jumlah SKS, dan penghasilan orang tua, dilakukan standarisasi (scaling) agar semua fitur berada pada skala yang sama. Melakukan normalisasi atau standardisasi data untuk memastikan semua fitur berada pada skala yang sama, sehingga lebih optimal dalam proses klasifikasi. Standarisasi ini dilakukan untuk menghindari bias model akibat perbedaan skala antar fitur, yang dapat mempengaruhi kinerja algoritma K-Nearest Neighbors (K-NN) karena algoritma ini sensitif terhadap skala jarak. Skala yang digunakan adalah standard scaler, di mana data dinormalisasi menjadi distribusi dengan rata-rata 0 dan standar deviasi 1.

3.5.6. Pengkodean Data Kategorikal (Jika Ada)

Mengonversi data kategorikal menjadi format numerik yang dapat digunakan dalam model. Misalnya, menggunakan Label Encoding.

3.5.7. Reduksi Dimensi dengan PCA

Menerapkan Principal Component Analysis (PCA) untuk mengurangi dimensi dataset, mempertahankan fitur yang paling signifikan, dan

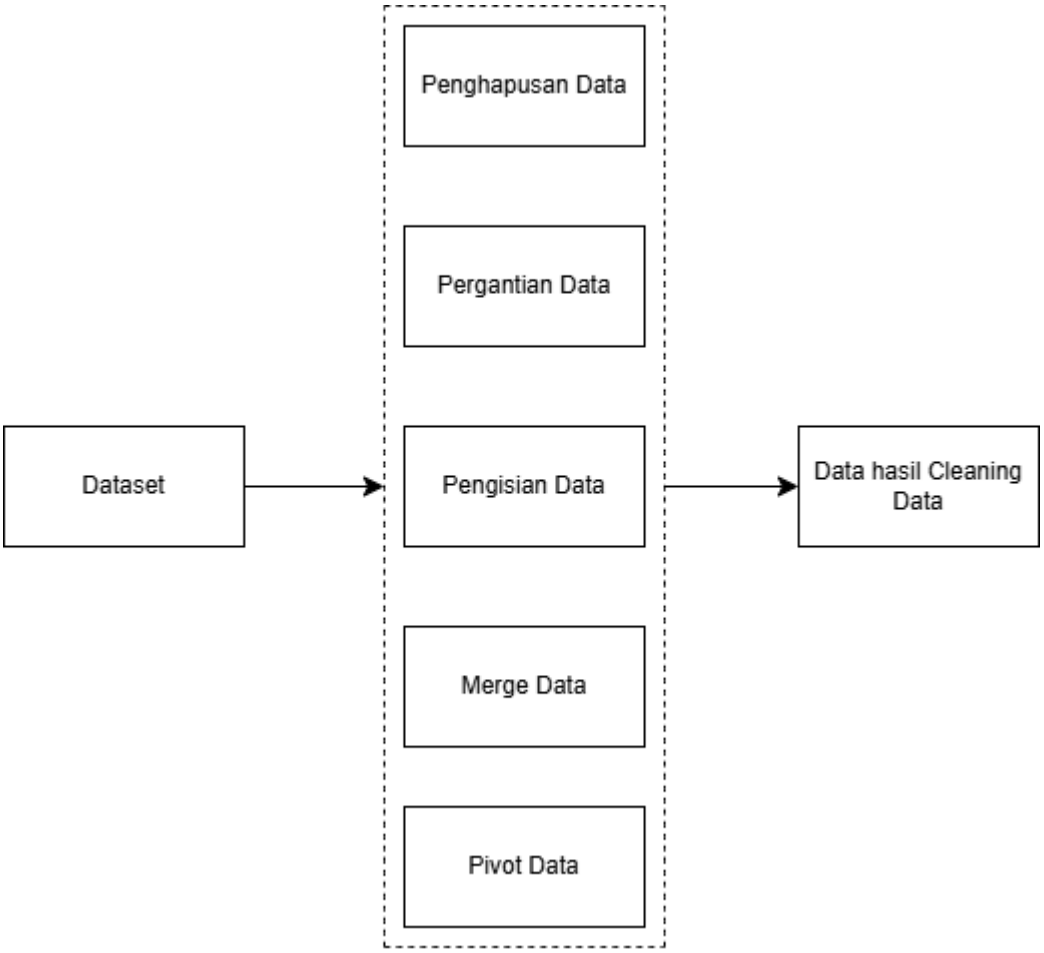
menghilangkan fitur yang kurang penting.

3.5.8. Dataset Siap Digunakan

Dataset hasil pra-proses siap digunakan untuk pelatihan dan pengujian model klasifikasi.

3.6. Cleaning Data

Kedua *dataset* akan dilakukan *Cleaning Data* yang beragam tergantung pada kebutuhan data pada proses selanjutnya. Terdapat 5 proses *Cleaning Data* pada Gambar x.x yang akan dilakukan.



Gambar x.x Proses *Cleaning Data*

Pada Gambar x.x dijelaskan alur dari proses *Cleaning Data* yang dilakukan pada *dataset* yang disediakn dan disesuaikan pada penggunaannya pada proses selanjutnya. Karena *dataset* yang digunakan merupakan data tabel maka fungsi *cleaning* sangat mirip seperti fungsi-fungsi pada pengolahan *database.*

3.6.1. Penghapusan Data

Penghapusan data dilakukan pada data-data yang tidak perlu digunakan atau data

uang sudah tidak memenuhi syarat untuk dipakai pada penelitian ini. Penghapusan data ini bertujuan untuk menghindari *error* data yang ditimbulkan oleh data yang tidak cocok dengan kriteria yang ada. Data yang dihapus berupa data yang tidak mempunyai salah satu kolom penting dalam *dataset.* Data yang dihapus bisa jadi hanya satu data saja atau bahkan bisa satu baris maupun satu kolom yang dihapus. Data tersebut akan dihapus dan data sesudahnya akan dinaikkan agar dapat mengisi kekosongan dari data sebelumnya. Pada Tabel x.x merupakan contoh dari metode penghapusan data.

Gambar x.x Contoh untuk penghapusan data

```
# 3. Menampilkan 5 data teratas sebelum pembersihan dalam bentuk DataFrame
print("\nData sebelum pembersihan:")
print(pd.DataFrame(data.head()))
```

```
Data sebelum pembersihan:
                                                ID;TahunMasuk;Prodi;JenisKelamin;J
ID00001;2021.0;S-1 TEKNIK ELEKTRO ;L;0.0;21.0;R...  68; SMA/MA IPA;3.55;SBMPTN;4.0;45
ID00002;2021.0;S.Tr. TEKNOLOGI REKAYASA OTOMASI...  42; SMA/MA IPA;1.42;PKM Mandiri;4
ID00003;2021.0;S-1 FISIKA ;P;0.0;22.0;;3            15; SMA/MA IPA;3.05;PKM Mandiri;4
ID00004;2020.0;S-1 BIOLOGI ;P;0.0;23.0;Rp. 1.50...  25; SMA/MA IPA;3.11;SNBT KIP-K;6.
ID00005;2020.0;S-1 SISTEM INFORMASI ;L;0.0;21.0...  56; SMA/MA IPA;3.67;PKM Kemitraar
```

Pada Gambar x.x diperlihatkan beberapa baris dan kolom sebuah tabel. Jika data tersebut tidak lengkap dan tidak bisa digunakan maka dibutuhkan fungsi *dropna* dengan contoh pada Gambar x.x

Gambar x.x Menampilkan kolom yang memiliki nilai null

```
# 4. Menampilkan kolom yang memiliki nilai null
print("\nKolom dengan nilai null dan jumlahnya:")
print(data.isnull().sum()[data.isnull().sum() > 0])
```

```
Kolom dengan nilai null dan jumlahnya:
ID;TahunMasuk;Prodi;JenisKelamin;JumlahCuti;Umur;Pendapatan;IPKPersiapan;JurusanSMA;I
dtype: int64
```

```
print("Kolom yang tersedia dalam dataset:", data.columns.tolist())
```

```
Kolom yang tersedia dalam dataset: ['ID;TahunMasuk;Prodi;JenisKelamin;JumlahCuti;Umur
```

Pada Gambar x.x diperlihatkan bahwa tidak ada data dalam tabel yang memiliki nilai null.

Gambar x.x Contoh penghapusan data

```
# 6. Menghapus baris yang memiliki nilai kosong di kolom penting
kolom_ada = [kolom for kolom in kunci_kolom if kolom in data.columns]
cleaned_data = data.dropna(subset=kolom_ada)


# 7. Menampilkan 5 data teratas setelah pembersihan dalam bentuk DataFrame
print("\nData setelah pembersihan:")
print(cleaned_data.head())
```

```
Data setelah pembersihan:
                                                  ID;TahunMasuk;Prodi;JenisKelamin;J
ID00001;2021.0;S-1 TEKNIK ELEKTRO ;L;0.0;21.0;R...  68; SMA/MA IPA;3.55;SBMPTN;4.0;45
ID00002;2021.0;S.Tr. TEKNOLOGI REKAYASA OTOMASI...  42; SMA/MA IPA;1.42;PKM Mandiri;4
ID00003;2021.0;S-1 FISIKA ;P;0.0;22.0;;3            15; SMA/MA IPA;3.05;PKM Mandiri;4
ID00004;2020.0;S-1 BIOLOGI ;P;0.0;23.0;Rp. 1.50...  25; SMA/MA IPA;3.11;SNBT KIP-K;6.
ID00005;2020.0;S-1 SISTEM INFORMASI ;L;0.0;21.0...  56; SMA/MA IPA;3.67;PKM Kemitraar
```

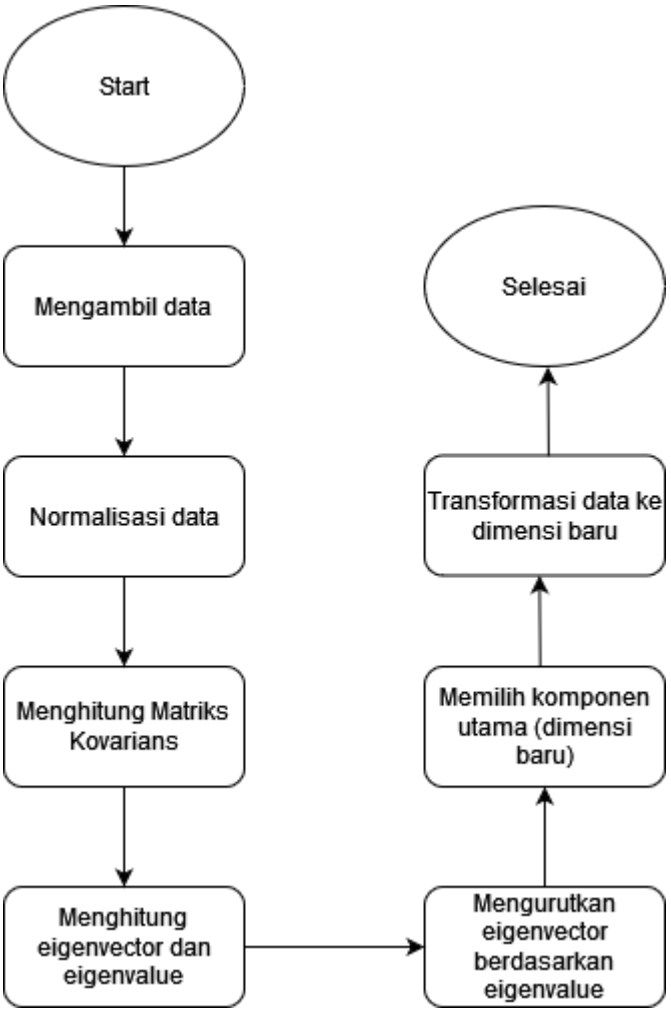
Setelah dilakukan penghapusan data yang dilaksanakan dengan fungsi *dropna* maka menyisakan data yang isinya sudah sesuai dalam artian tidak ada data yang kosong dalam baris data tersebut.

3.7. Principal Component Analysis (PCA)

Proses dimulai dengan menyiapkan dataset mahasiswa yang mencakup berbagai karakteristik, seperti nilai akademik, kehadiran, status sosial-ekonomi, dan data demografis. Setelah data dikumpulkan, dataset disiapkan dengan membersihkan data yang tidak valid, mengisi nilai yang hilang jika diperlukan, serta menormalisasi data agar setiap fitur berada dalam skala yang seragam.

Langkah berikutnya adalah menerapkan Principal Component Analysis (PCA) untuk mengurangi dimensi dataset. Pada tahap pertama, rata-rata dari setiap fitur dihitung, dan kemudian setiap nilai fitur dikurangi dengan rata-rata tersebut untuk mendapatkan data yang terstandarisasi. Setelah data distandardisasi, dihitung matriks kovarians untuk memahami korelasi antar-fitur. PCA diterapkan untuk mereduksi dimensi dataset tanpa kehilangan informasi penting. Proses ini melibatkan perhitungan eigenvectors dan eigenvalues dari matriks kovarians data, untuk menentukan komponen utama yang menjelaskan variasi terbesar dalam dataset. Hasil PCA digunakan sebagai input pada algoritma KNN untuk meningkatkan performa prediksi.

Selanjutnya, dilakukan perhitungan eigenvectors dan eigenvalues dari matriks kovarians tersebut. Eigenvalues menunjukkan ukuran variansi pada setiap arah komponen, sementara eigenvectors menunjukkan arah komponen tersebut. Berdasarkan nilai eigenvalues yang tertinggi, komponen utama dipilih untuk mempertahankan persentase variansi yang diinginkan. Data asli kemudian ditransformasikan ke ruang fitur baru yang ditentukan oleh komponen utama tersebut, menghasilkan dataset berdimensi lebih rendah.



Dalam konteks dataset nilai mahasiswa, eigenvalues dan eigenvectors dapat membantu mengidentifikasi pola utama dalam data, serta mengurangi dimensi data sambil mempertahankan informasi yang paling penting. Misalkan dataset nilai mahasiswa memiliki berbagai fitur seperti nilai mata kuliah, tingkat kehadiran, umur, semester, dan status sosial-ekonomi. Berikut adalah cara eigenvalues dan eigenvectors diterapkan dalam konteks ini:

3.7.1. Eigenvectors (Vektor Eigen)

Eigenvectors adalah vektor arah yang mewakili pola umum di data mahasiswa. Setiap eigenvector menunjukkan kombinasi fitur yang memiliki variasi data yang maksimal. Misalnya, sebuah eigenvector dalam dataset ini mungkin menunjukkan pola bahwa nilai akademik dan kehadiran memiliki hubungan erat. Artinya, mahasiswa dengan kehadiran tinggi cenderung memiliki nilai lebih tinggi, atau sebaliknya. Dalam PCA, eigenvectors menunjukkan arah utama di mana variasi dalam nilai mahasiswa terjadi. Jadi, eigenvectors yang didapatkan mungkin menunjukkan bahwa variasi terbesar dalam dataset nilai mahasiswa adalah dalam nilai mata kuliah utama, sedangkan variasi lebih kecil terdapat pada fitur lain, seperti umur atau semester.

3.7.2. Eigenvalues (Nilai Eigen)

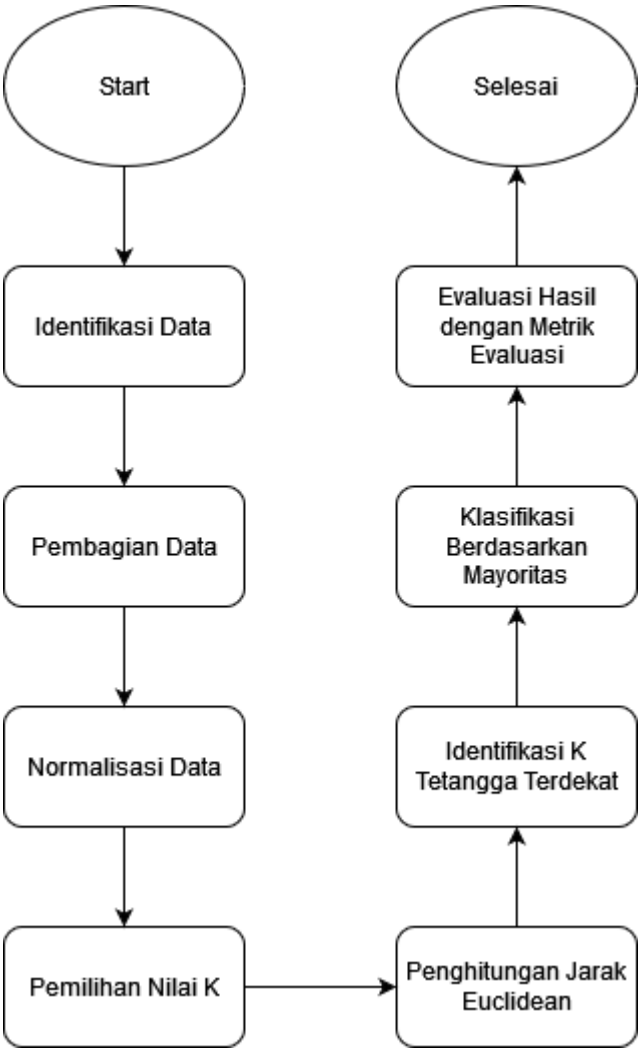
Eigenvalues mewakili ukuran atau kekuatan dari variasi data yang direpresentasikan oleh setiap eigenvector. Misalnya, jika satu eigenvalue memiliki nilai besar untuk vektor yang mewakili kombinasi nilai akademik dan kehadiran, ini berarti bahwa kombinasi tersebut menjelaskan banyak variansi dalam dataset mahasiswa. Semakin besar eigenvalue, semakin besar peran atau pengaruh kombinasi fitur tersebut dalam keseluruhan dataset. Dalam PCA, eigenvalues digunakan untuk menentukan seberapa signifikan setiap eigenvector. Jika sebuah eigenvalue kecil, komponen utama (eigenvector) yang terkait dengan nilai tersebut kemungkinan tidak memiliki banyak informasi dan bisa diabaikan dalam analisis, sehingga mengurangi dimensi dataset tanpa kehilangan banyak informasi.

Setelah transformasi PCA selesai, dataset hasil reduksi dimensi ini kemudian dibagi menjadi data latih dan data uji. Langkah berikutnya adalah klasifikasi menggunakan algoritma K-Nearest Neighbors (K-NN). Dalam proses K-NN, data baru yang akan diklasifikasikan dihitung jaraknya dengan data dalam data latih. Berdasarkan jarak terdekat ke sejumlah data latih, kelas data baru tersebut ditentukan.

Tahap akhir adalah evaluasi performa model dengan metrik seperti akurasi, presisi, dan recall untuk memastikan kemampuan model dalam mengidentifikasi mahasiswa yang berpotensi drop-out dengan tepat. Proses ini berakhir setelah evaluasi selesai dilakukan.

3.8. Perancangan Model KNN

Model KNN dirancang untuk mengklasifikasikan mahasiswa berdasarkan karakteristiknya. Setelah dataset direduksi dimensinya oleh PCA, algoritma KNN digunakan untuk mengklasifikasikan data dengan menghitung jarak terdekat antar sampel. Model ini dilatih dan diuji menggunakan data latih dan data uji yang telah diproses.



Literatur yang telah dikaji pada bagian penelitian terkait akan menjadi dasar teori atas studi dan penelitian yang dilakukan agar dapat menjadi landasan teori atas dilaksanakannya penelitian ini. Konsep dan dasar teori yang digunakan dapat meliputi sumber buku, jurnal dari penelitian yang telah dilakukan sebelumnya yang relevan. Referensi dari studi literatur yang telah dipelajari dan digunakan pada penelitian ini dilampirkan pada daftar pustaka di bagian akhir laporan skripsi.

3.9. Skenario Pengujian

Skenario pengujian dilakukan dengan memvariasikan data split dan nilai K pada algoritma KNN yang telah ditingkatkan dengan PCA. Pengujian dilakukan dengan memvariasikan parameter seperti proporsi data latih-uji dan nilai K dalam KNN. Beberapa skenario dirancang untuk mengevaluasi performa model, dengan metrik seperti akurasi, presisi, dan recall. Berikut adalah skenario pengujia yang akan diterapkan dalam penelitian ini :

a. Metode : Menggunakan algoritma K-Nearest Neighbors (KNN) dengan

peningkatan menggunakan Principal Component Analysis (PCA) untuk reduksi dimensi.

b. Latih : Melakukan pelatihan pada data yang telah di proses menggunakan PCA.

c. Data Split : Menggunakan beberapa variasi data split untuk membagi data latih dan data uji (contoh : 79:30, 80:20, dan 90:10).

d. Uji : Pengujian akurasi dilakukan pada data uji untuk melihat performa model.

e. Parameter K : Menggunakan beberapa nilai K berbeda pada algoritma KNN untuk membandingkan hasil akurasi.

Deskripsi Skenario :

a. Skenario 1 dan 2 : Menggunakan data split 70:30 dan menguji dengan dua nilai K (K=3 dan K=5).

b. Skenario 3 dan 4 : Menggunakan data split 80:20 dan menguji dengan dua nilai K (K=3 dan K=5)

c. Skenario 5 dan 6 : Menggunakan data split 90:10 dan menguji dengan dua nilai K (K=3 dan K=5)

Berikut ini adalah tabel skenario pengujian yang dapat digunakan:

Metode	Latih	Data Split	Uji	Parameter K	Skenario
KNN + PCA	Data PCA	70:30	Data Uji	K = 3	Skenario 1
KNN + PCA	Data PCA	70:30	Data Uji	K = 5	Skenario 2
KNN + PCA	Data PCA	80:20	Data Uji	K = 3	Skenario 3
KNN + PCA	Data PCA	80:20	Data Uji	K = 5	Skenario 4
KNN + PCA	Data PCA	90:10	Data Uji	K = 3	Skenario 5
KNN + PCA	Data PCA	90:10	Data Uji	K = 5	Skenario 6

Hasil dari setiap skenario dianalisis untuk menentukan konfigurasi terbaik

dalam memprediksi potensi drop-out mahasiswa.

DAFTAR PUSTAKA

1. Bila referensi berupa *Buku*

Dick, H.W. 1990. **Industri Pelayaran Indonesia : Kompetisi dan Regulasi**. Diterjemahkan oleh Burhanuddin A. Jakarta: LP3ES.

Franklin, J.H. 1985. **Fundamentals of Mathematics**. Chicago: University of Chicago Press.

Kernighan, B.W., dan Dennis M. R. 1987. **The C Programming Language**. Englewood Cliffs, N.J.: Prentice Hall.

Kuo S.M. dan Morgan D.R. 1996. **Active Noise Control Systems: Algorithms and DSP Implementation**. John Wiley & Sons, Inc.

Whaley, W. G., Osmond P. B., dan Henry S.L. 1983. **Logic and Boolean Logic**. London: John Murray.

2. Bila referensi merupakan bagian dari buku yang ditulis oleh banyak penulis

Abraham, G.H. 1989. **Differential and Integral** in Franklin, J.H. (Ed). **Fundamentals of Mathematics**. Chicago: University of Chicago Press.

3. Bila referensi berupa *Prosiding*

Akazana, S. 1983. "The Scope Of The Japanese Information Industry In The 1980s". **Proceeding Of The Forty First FID Congress**. Hongkong, 13-16 September. Diedit oleh K.R. Brown. New York : North Holland Publishing Company.

Cavalieri, S., Di Stefano, A., dan Mirabella, O., 1991. “Assessment of the Priority Mechanism in the Fieldbus Data Link Layer”. **Proceeding Industrial Electronics, Control and Instrumentation.** IECON ‘91.

Henry, R.R., 1990. “Performance of IEEE 802 Local Area Networks”. **IEEE Proceeding Southeastcon.** Session 5D4:414-419.

Simar, Ray Jr. 1986. “Floating-Point Arithmatic with the TMS322010”, **Digital Signal Processing Applications with the TMS320 Family.** Texas Instruments.

4. Bila referensi berupa *artikel dalam Jurnal*

Bondavalli, A., Conti, M., Gregori, E., Lenzini, L., and Strigini, L., Feb. 1990. “MAC protocols for High-speed MANs: Performance Comparasions for a Family of Fasnet-based Protocols”. **Computer Networks and ISDN Systems** 18, 2:97-113.

Conti, M., Gregori, E., and Lenzini, L., March 1994. “E-DCP An Extension of the Distributed-control Polling MAC Protocol (DCP) for Integrated Services”. **Computer Networks and ISDN Systems** 26, 6-8:711-719.

Jackson, R. 1979. "Running Down The Up Escalator : Regional Inequality In Papua New Guinea". **Australian Geographer** 14 (May) : 175□18 4.

Koubias, S.A. and Papadopoulos, G.D., Aug. 1995. “Modern Fieldbus Communication Architectures for Real-time Industrial Applications”. **Computer in Industry** 26, 3:243-252.

Linge, N., Ball, E., Tasker, R., dan Kummer, P., 1987. “A Bridge Protocol for Creating a Spanning Tree Topology within an IEEE 802 Extended LAN Environment”. **Computer Networks and ISDN Systems** 13, 4&5:323-332.

Shin, K.G., dan Chou, C.C., June 1996. “Design and Evaluation of Real-time Communication for Fieldbus Based Manufacturing Systems”, **IEEE Transactions on Robotic and Automation** 12, 3:357-367.

5. Bila referensi berupa *artikel dalam Majalah*

Santori, M. dan Zech, K., Maret 1996. “Fieldbus brings Protocol to Process Control”. **IEEE Spectrum** 33, 3:60-64.

Weber, B. 1985. "The Myth Maker : The Creative Mind". **New York Times Magazines**, 20 October, 42.

6. Bila referensi berupa *artikel dalam Surat Kabar*

Kompas (Jakarta). 1992. 4 Januari.

Jawa Pos (Surabaya). 1993. 21 April.

Rahayu, S. 1992. "Hendak Kemana Arsitektur Rumah Susun Indonesia?". **Kompas** (Jakarta), 5 Maret.

Sjahrir, A. 1993. "Prospek Ekonomi Indonesia". **Jawa Pos** (Surabaya), 22 Maret.

7. Bila referensi berupa *artikel dari Internet*

Coutinho, J., Martin, S., Samata, G., Tapley, S. dan Wilkin, D., 1995. **Fieldbus Tutorial,** <URL:http://kernow.curtin.edu.au/www/fieldbus/fieldbus.htm>.

Pinto, J.J., Feb. 1997. **Fieldbus: A Neutral Instrumentation Vendor's Perspective Communicatio,** <URL:http://www.actionio.com/jimpinto/fbarticl.html>.

8. Referensi lainnya (Manual, Brosur, dan sejenisnya)

Reliable Supply in Reliable Quality. **Brosur PT. Dharma Sarana Perdana.** Pulo Gadung Industrial Estate**,** Jakarta.

Engineering Education and Training. **Catalogue Plint Engineering**. Oakland Park, Wokingham.

Monograf Kelurahan Wonorejo, Rungkut, Surabaya, 2006.

LAMPIRAN PERSYARATAN FISIK DAN TATA LETAK

1. Kertas

Kertas yang digunakan adalah HVS 70 mg berukuran A4. Apabila terdapat gambar-gambar yang menggunakan kertas berukuran lebih besar dari A4, hendaknya dilipat sesuai dengan aturan yang berlaku. Pengetikan hanya dilakukan pada satu muka kertas, tidak bolak balik.

2. Margin

Batas pengetikan naskah adalah sebagai berikut :

- Margin kiri: 4 cm
- Margin atas: 3 cm
- Margin kanan: 3 cm
- Margin bawah: 3 cm

3. Jenis dan Ukuran Huruf

Jenis huruf yang dipakai dalam skripsi adalah Times New Roman dengan ketentuan sebagai berikut:

- Judul bab pada level 1 berukuran 14 pt
- Judul subbab pada level 2 berukuran 12 pt
- Judul subbab pada level 3 berukuran 12 pt
- Judul subbab pada level 4 berukuran 12 pt
- Badan teks berukuran 12 pt

Penggunaan jenis dan ukuran ini harus konsisten. Untuk memudahkan memelihara konsistensi sekaligus penyusunan struktur skripsi, fasilitas seperti *styles* dan *multilevel list* dalam program pengolah kata dapat digunakan. Sebuah *template* untuk skripsi ini telah disediakan untuk membantu mahasiswa. *Styles* dan *multilevel list* dalam template tersebut sudah dirancang untuk jenis dan ukuran huruf yang disyaratkan.

4. Spasi

Jarak standar antar baris dalam badan teks adalah satu setengah (1,5) spasi. Jarak antar paragraf, antara judul bab dan judul subbab, antara judul subbab dan badan teks, dan seterusnya, dapat dilihat pada masing-masing *style* yang digunakan dan tersedia dalam *template* untuk skripsi ini.

5. Kepala Bab dan Subbab

Kepala bab terdiri dari kata "BAB" yang diikuti dengan nomor bab dengan angka Romawi besar. Untuk judul dari bab tersebut dituliskan satu baris dibawah kepala bab, misalnya :

BAB I

PENDAHULUAN

Kepala subbab diawali dengan nomor sesuai tingkat hirarkinya dan diikuti dengan judul subbab, misalnya "1.1 Latar Belakang". Penomoran subbab disarankan tidak lebih dari 4 level (maksimal subbab X.X.X.X). Kepala bab dan subbab tidak boleh mengandung *widow* atau *orphan* sehingga nampak menggantung atau terputus di bagian awal atau akhir sebuah halaman. *Widow* adalah sebuah paragraf dengan hanya satu baris pertama pada akhir halaman sedangkan sisanya berada pada halaman berikutnya. *Orphan* adalah baris terakhir dari satu paragraf yang tertulis pada awal suatu halaman sedangkan baris lainnya dari paragraf tersebut berada pada halaman sebelumnya.

6. Nomor Halaman

Bagian awal skripsi menggunakan nomor halaman berupa angka Romawi kecil (i, ii, iii, iv, dan seterusnya) yang dimulai dari sampul dalam. Sedangkan bagian utama dan bagian akhir skripsi menggunakan nomor halaman berupa angka Arab (1, 2, 3, dan seterusnya) yang dimulai dari bab I. Semua nomor halaman diletakkan di tengah bawah.

LAMPIRAN PENGGUNAAN BAHASA

Bahasa yang dipakai dalam skripsi adalah bahasa Bahasa Indonesia yang baku. Setiap kalimat harus memiliki subjek dan predikat, dan umumnya dilengkapi juga dengan objek, pelengkap, atau keterangan. Setiap paragraf biasanya terdiri dari beberapa kalimat. Penuturan isi dalam kalimat, paragraf, maupun antar paragraf harus menggunakan bahasa yang tepat dan menggambarkan alur logika yang runtut.

Penulisan bahasa asing yang sudah diserap dalam Bahasa Indonesia disesuaikan dengan kaidah Bahasa Indonesia. Sedapat mungkin dihindari penggunaan bahasa asing jika istilah dalam bahasa Indonesia sudah ada. Jika terpaksa menggunakan istilah dalam bahasa asing, maka penulisannya harus sesuai ejaan aslinya dan dicetak miring (*italic*), kecuali jika istilah tersebut adalah nama.

Sebagai referensi untuk penulisan Bahasa Indonesia yang baku, dokumen berikut dapat digunakan:

- Kamus Bahasa Indonesia, Tim Penyusun, Pusat Bahasa Departemen Pendidikan Nasional, Jakarta 2008
- Peraturan Menteri Pendidikan Nasional Republik Indonesia nomor 46 tahun 2009 tentang Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan
- Kamus Besar Bahasa Indonesia dalam jaringan (KBBI daring): http://bahasa.kemdiknas.go.id/kbbi/index.php