BAB II
TINJAUAN PUSTAKA

2.1 Penelittian terkait

Penelitian ini didasarkan pada penelitian yang telah dilakukan atau penelitian sebelumnya yaitu penelitian yang dilakukan oleh Sri Widaningsih, [4] "Perbandingan Metode Data Mining Untuk Memprediksi Nilai dan Waktu Kelulusan Mahasiswa Program Studi Teknik Informatika Menggunakan C4.5, Naïve Bayes Algoritma KNn, dan SVM. Penelitian ini menggunakan variabel kelulusan, IPK, IPS, dan jenis kelamin. Hasil akhir dari keempat algoritma menunjukkan bahwa algoritma Naïve Bayes merupakan algoritma terbaik dalam memprediksi kelulusan mahasiswa tepat waktu dan IPK ≥ 3 dengan akurasi (76,79%), error (23,17%), dan AUC (0,850).

Penelitian lain dilakukan oleh Oktaviana Bangun dkk[5], dengan judul "Metode Algoritma Linear Support Vector Machine (SVM) dalam Memprediksi Kelulusan Siswa". Variabel yang digunakan dalam penelitian ini adalah Nama, IPS 1 hingga IPS 4 dan keterangan Ya atau Tidak. Hasil pengujian dengan 70% data latih dan 30% data uji menunjukkan bahwa algoritma linear SVM memberikan nilai akurasi sebesar 90%.

Penelitian selanjutnya dilakukan oleh Suhardjono dkk,[6], dengan judul "Prediksi Waktu Kelulusan Mahasiswa Menggunakan SVM Berbasis PSO". Variabel yang digunakan dalam penelitian ini adalah Jurusan SMA, STLA Asal dan IPK1-IPK6. Dari hasil yang diperoleh penggunaan model support vector machine berbasis Particle Swarm Optimizer dapat ditingkatka

akurasi prediksi dari 85,81% menjadi 86,43%. dengan peningkatan sebesar 00,62%. Sehingga prediksi kelulusan mahasiswa dapat akurat dan optimal dalam mengukur parameter yang dibutuhkan.

Penelitian selanjutnya "Klasifikasi Prediksi Kelulusan Mahasiswa Menggunakan Metode Support Vector Machine (SVM)" dilakukan oleh M Riski Qisthiano [7]. Penelitian ini menggunakan beberapa variabel antara lain nama, jurusan, fakultas, tahun masuk perguruan tinggi dan tahun kelulusan mahasiswa. Hasil dari penelitian ini adalah akurasi hasil klasifikasi untuk prediksi yang diperoleh dari alat Rapidminer dan model Support Vector Machine (SVM) memiliki hasil akurasi sebesar 85,06%.

Penelitian selanjutnya diberi judul "Prediksi Kelulusan Siswa Menggunakan Algoritma K-means Data Mining. Penerapan Algoritma Support Vector Machine untuk Model Prediksi Kelulusan Siswa Tepat Waktu" oleh Emy Haryatmi & Sheila Pramita Hervianti [8]. Variabel yang digunakan dalam penelitian ini adalah nama, jurusan, fakultas, tahun masuk perguruan tinggi dan tahun kelulusan mahasiswa. Hasil pengujian kelompok pertama dengan 90% data latih dan 10% data uji menunjukkan bahwa algoritma SVM memberikan nilai akurasi yang sangat baik yaitu 94,4%.

Berdasarkan penelitian diatas, salah satu fungsi penerapan metode Support Vector Machine (SVM) adalah mengklasifikasikan data prediksi siswa untuk menentukan kelulusan tepat waktu. Maka pada penelitian kali ini metode tersebut akan diimplementasikan dan dicari hasil keakuratannya dengan menggunakan model Support Vector Machine (SVM).

2.2 Landasan Teori

2.2.1 Pendidikan dan Faktor-faktor Yang Mempengaruhi terlamba tidaknya mahasiswa lulus

Pendidikan merupakan suatu hal penting yang dibutuhkan setiap individu manusia, melalui pendidikan seseorang dapat mengembangkan dirinya menjadi pribadi yang baik dalam hidupnya. Pendidikan pada perguruan tinggi yang merupakan jenjang pendidikan formal tertinggi mempunyai batas waktu belajar yang telah ditentukan oleh lembaga yang berwenang. Batasan waktu belajar adalah waktu maksimal seorang mahasiswa untuk menyelesaikan pendidikannya pada suatu program studi. Apabila siswa melaksanakan proses pendidikan melampaui batas waktu belajar, maka siswa tersebut dapat dinyatakan gagal [1].

Dalam proses menentukan mahasiswa mana yang dapat menyelesaikan studinya baik tepat waktu atau tidak tepat waktu, banyak faktor yang menjadi penyebabnya. Sebab kelulusan mahasiswa tepat waktu merupakan salah satu penilaian dalam proses akreditasi perguruan tinggi. Faktor-faktor yang dapat mempengaruhi terlambat atau tidaknya mahasiswa lulus adalah sebagai berikut [9];

1. Faktor Fisiologis

Faktor yang berhubungan dengan kondisi fisik seseorang. Kondisi fisik umum yang sangat mempengaruhi aktivitas belajar seseorang. Kondisi fisik yang sehat akan memberikan pengaruh positif terhadap aktivitas belajar individuFaktor psikologis Keadaan psikologis seseorang yang dapat mempengaruhi proses belajar. Beberapa faktor yang mempengaruhi proses belajar adalah kecerdasan siswa, motivasi, minat, sikap, dan bakat.

2. Lingkungan

Keluarga Belajar akan menerima pengaruh dari keluarga berupa: cara orang tua mendidik, suasana di rumah dan keadaan ekonomi keluarga. Pertama, cara orang tua mendidik anaknya mempunyai pengaruh yang besar terhadap pembelajaran anaknya.

3. Masyarakat

Masyarakat merupakan faktor eksternal yang turut mempengaruhi pembelajaran. Pengaruh ini terjadi karena kehadirannya pada lingkungan tersebut. Pertama, aktivitas di masyarakat dapat bermanfaat bagi pengembangan pribadi.

4. Bergaul dengan teman-teman

Inilah pengaruh-pengaruh yang masuk ke dalam jiwa seseorang dengan sangat cepat. Teman yang baik akan memberikan pengaruh yang baik bagi seseorang, begitu pula sebaliknya, teman yang buruk pasti akan memberikan pengaruh yang buruk juga. Misalnya saja orang yang suka begadang, pecandu narkoba, dan sebagainya.

5. Motivasi

Motivasi menurut Slavina (1994) [9] merupakan salah satu faktor yang mempengaruhi efektivitas kegiatan belajar. Motivasi inilah yang mendorong keinginan untuk melakukan kegiatan belajar. Para ahli mendefinisikan motivasi sebagai suatu proses dalam diri individu yang bersifat aktif, mendorong, memberikan arahan, dan memelihara perilaku setiap saat.

2.2.2 *Machine Learning* (ML)

Menurut Gotama (2018) [10], Machine Learning (ML) adalah teknik menyimpulkan data dengan menggunakan pendekatan matematika. Intinya, pembelajaran mesin adalah teknik yang digunakan untuk membuat model (matematis) yang menggambarkan pola dalam data. Inferensi yang dimaksud dalam pembelajaran mesin lebih fokus pada hubungan antar atribut. Selain itu, pembelajaran mesin merupakan salah satu bentuk penggambaran data/sains/pengetahuan dalam model formulasi matematika. Disebut model matematika karena pembelajaran mesin merupakan turunan dari rumusan matematika dan statistik. Pembelajaran mesin ibarat “alat” yang identik dengan rumus matematika.

Secara umum proses machine learning terdiri dari enam proses utama [11], yaitu:

1. Mengumpulkan Data Langkah pengumpulan data merupakan langkah dasar untuk proses pembelajaran mesin. Persiapan Data Setelah data terkumpul dari sumbernya, langkah selanjutnya adalah mempersiapkan data agar dapat digunakan untuk proses pelatihan machine learning.
2. Pemilihan Model Proses selanjutnya adalah pemilihan model yang relevan dengan penelitian. Model umumnya dipilih berdasarkan relevansinya dengan kasus penelitian.
3. Pelatihan Salah satu proses utama pembelajaran mesin adalah pelatihan. Dalam proses ini, data digunakan dalam pengembangan untuk meningkatkan kemampuan model dalam memprediksi.

4. Parameter Evaluasi dan Tuning, Evaluasi akan menggunakan data uji yang belum pernah digunakan untuk memungkinkan melihat bagaimana performa model pada data yang belum digunakan dan melakukan tuning parameter untuk melihat apakah berdasarkan parameter yang dimodifikasi akan menyebabkan peningkatan atau berkurang. dalam hasil evaluasi.
5. Prediksi Proses akhir pembelajaran mesin adalah menggunakan data untuk memberikan jawaban atas pertanyaan. Prediksi merupakan suatu proses yang bertujuan untuk menjawab beberapa pertanyaan

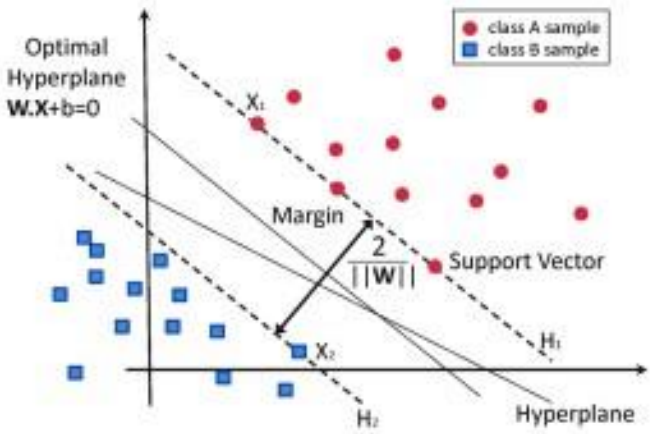
Ikon Terverifikasi Komunitas Dalam machine learning terdapat 2 (dua) metode pembelajaran yang paling banyak digunakan yaitu [10]

1. Unsupervised Learning diterapkan pada data yang tidak memiliki label historis. Tujuan penggunaan metode unsupervised learning adalah untuk memperdalam data dan mengidentifikasi struktur dalam data. Misalnya, Pembelajaran Tanpa Pengawasan biasanya digunakan untuk pengelompokan atau klasifikasi kelas.
2. Pembelajaran yang diawasi (*Supervised Learning)* menggunakan pola yang berguna untuk memprediksi tingkat kelas pada data baru yang tidak berlabel. Data baru harus mempunyai jumlah atribut yang sama dengan data sumber atau data awal. Supervised learning umumnya diterapkan pada aplikasi yang memiliki data historis yang digunakan untuk memprediksi kemungkinan-kemungkinan yang akan terjadi di masa depan.

2.2.3 *Support Vector Machine* (SVM)

Support Vector Machine merupakan sistem pembelajaran yang menggunakan hipotesis berupa fungsi linier pada fitur berdimensi tinggi dan dilatih

menggunakan algoritma pembelajaran berdasarkan teori optimasi [7]. SVM *(Support Vector Machine)* merupakan salah satu metode dalam *Supervised Learning* yang digunakan untuk mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. Performa SVM sangat bergantung pada pilihan nilai parameter yang memadai, termasuk misalnya parameter kernel dan regularisasi. Pemilihan parameter SVM secara umum merupakan masalah optimasi dimana teknik pencarian digunakan untuk menemukan konfigurasi parameter yang memaksimalkan kinerja SVM [3].dimana:



Gambar 2. 1 *Support Vector Machine* (SVM)

max : jarak maximum antar dua garis
w,b

w, x , b dan y: komponen dari persamaan garis lurus

Hyperplane yang merupakan pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur *hyperplane* tersebut. dan carilah titik maksimumnya. Margin adalah jarak antara *hyperplane* dengan pola terdekat dari masing-masing kelas. Pola terdekat ini disebut vektor pendukung. Garis padat pada gambar menunjukkan *hyperplane* terbaik yang terletak tepat di tengah-tengah kedua kelas, sedangkan titik merah dan kuning pada lingkaran hitam merupakan vektor pendukung. Upaya

mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM [12]. SVM memiliki beberapa parameter yang umum digunakan yaitu [13];

1. Kernel

Penggunaan kernel bertujuan untuk mengubah data menjadi ruang berdimensi tinggi, dengan membuat data non-linier dipisahkan secara linier. Ada beberapa fungsi kernel yang digunakan oleh aplikasi di SVM yaitu;

a. Kernel linier adalah fungsi kernel yang paling sederhana
b. Kernel polinomial adalah fungsi kernel yang digunakan saat datatidak dipisahkan secara linear.
c. Kernel fungsi basis radial (RBF) adalah fungsi kernel yang digunakan dalam analisis ketika data tidak dipisahkan secara linier. RBF memiliki 2 parameter yaitu gamma dan cost.
d. Gamma Menentukan seberapa besar pengaruh satu sampel terhadap kumpulan data pelatihan, dengan nilai rendah berarti “jauh” dan nilai tinggi berarti “dekat”. Menurut Patel (2017) [13]. jika gammanya tinggi berarti titik-titik yang berada di sekitar garis pemisah normal akan diperhitungkan dalam perhitungan dan gamma yang rendah berarti titik-titik yang jauh dari garis pemisah normal akan diperhitungkan dalam perhitungan garis pemisah.
e. Parameter biaya (C)

Ini merupakan parameter yang berfungsi sebagai pengoptimal SVM untuk menghindari kesalahan klasifikasi setiap sampel dalam dataset pelatihan. Nilai C yang tinggi cenderung menghasilkan batasan keputusan yang lebih ketat.

f. *Degree*

Parameter *degree* terdapat pada kernel *polynomial* dimana semakin tinggi nilai derajatnya semakin kompleks modelnya. Derajat digunakan untuk mentransformasi input data ke dalam ruang fitur yang lebih tinggi.

2.2.4 Klasifikasi

Klasifikasi adalah teknik yang melihat perilaku dan atribut kelompok yang telah ditentukan. Teknik ini dapat memberikan klasifikasi data baru dengan memanipulasi data yang sudah ada yang telah diklasifikasikan dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan [[7].

Klasifikasi merupakan suatu cara pengelompokan objek berdasarkan ciri-ciri yang dimiliki oleh objek klasifikasi tersebut. Dalam prosesnya, klasifikasi dapat dilakukan dengan banyak cara, baik secara manual maupun dengan bantuan teknologi. Klasifikasi manual adalah klasifikasi yang dilakukan oleh manusia tanpa bantuan algoritma komputer yang cerdas. Sedangkan klasifikasi dilakukan dengan bantuan teknologi, memiliki beberapa algoritma antara lain *Naïve Bayes, Support Vector Machine, Decission Tree, Fuzzy* dan Jaringan Syaraf Tiruan [12]

2.2.5 *Confusion Matrix*

Confusion Matrix adalah tabel dengan 4 campuran nilai prediksi dan nilai aktual yang berbeda. Berdasarkan Tabel 1 dapat dijelaskan bahwa terdapat 4 istilah yang mewakili hasil proses klasifikasi pada matriks konfusi yaitu True Positif (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), *matriks Confusion* sering juga disebut dengan matriks kesalahan. berikut adalah matriks kebingungan untuk menjelaskan ukuran kinerja klasifikasi[14]

Tabel 2.1 *Confusion Matrix*

Aktuali	**Predikki**	
	Positive	**Negative**
Positive	**True Positive**	**False Negative**
Negative	**False Positive**	**True Negative**

Accuracy = Didefinisikan selaku jenjang korelasi antara nilai prediksi dengan nilai aktual

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision = Merupakan tingkatan ketepatan antara data yang diharapkan oleh pengguna dengan jawaban yang diberikan oleh system

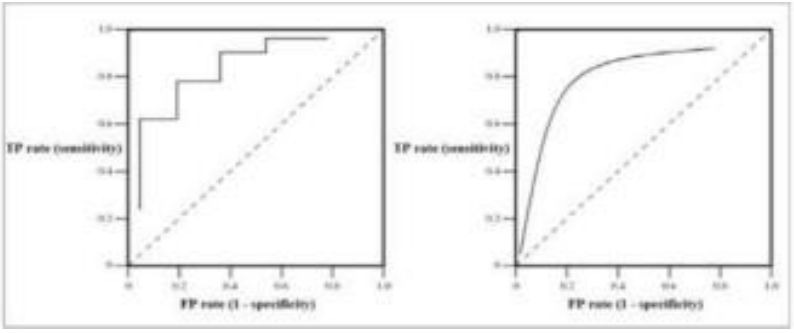
$$\frac{TP}{TP+FP} \quad (2)$$

Recall = Merupakan tingkatan keberhasilan 22embal dalam menciptakan kembali suatu data

$$\frac{TP}{TP+FN} \quad (3)$$

ROC = adalah ukuran numerik untuk membedakan kinerja model dan menunjukkan seberapa sukses dan benar model memberi peringkat pada pengamatan positif dan negatif

Kurva ROC menunjukkan keakuratan dan membandingkan klasifikasi secara visual dengan nilai positif palsu (spesifisitas) sebagai garis horizontal dan nilai positif sebenarnya (sensitivitas) sebagai garis vertikal [15]



Gambar 2. 2 Kurva ROCKurva ROC

Garis diagonal yang membagi ruang ROC menggambarkan bahwa spasi di atas garis diagonal menunjukkan klasifikasi baik dan spasi di bawah garis diagonal menunjukkan klasifikasi buruk, sedangkan tebakan yang benar-benar acak terdapat di sepanjang garis diagonal dari kiri bawah ke kanan atas [15].

Menurut Ir. Adi Sucipto, n.d. (2019)[16] Cara umum menghitung luas di bawah kurva ROC adalah Area Under Curve (AUC) dimana luas di bawah kurva mempunyai nilai yang selalu berada pada angka 0,0 dan 1,0. Namun yang menarik untuk dihitung adalah bagi yang luasnya diatas 0,5 maka semakin tinggi luasnya maka semakin baik, seperti pada petunjuk di bawah ini.

• 0,9- 1,00 = klasifikasi sangat baik

• 0,8 – 0,9 = klasifikasi baik

• 0,7 – 0,8 = klasifikasi rata-rata

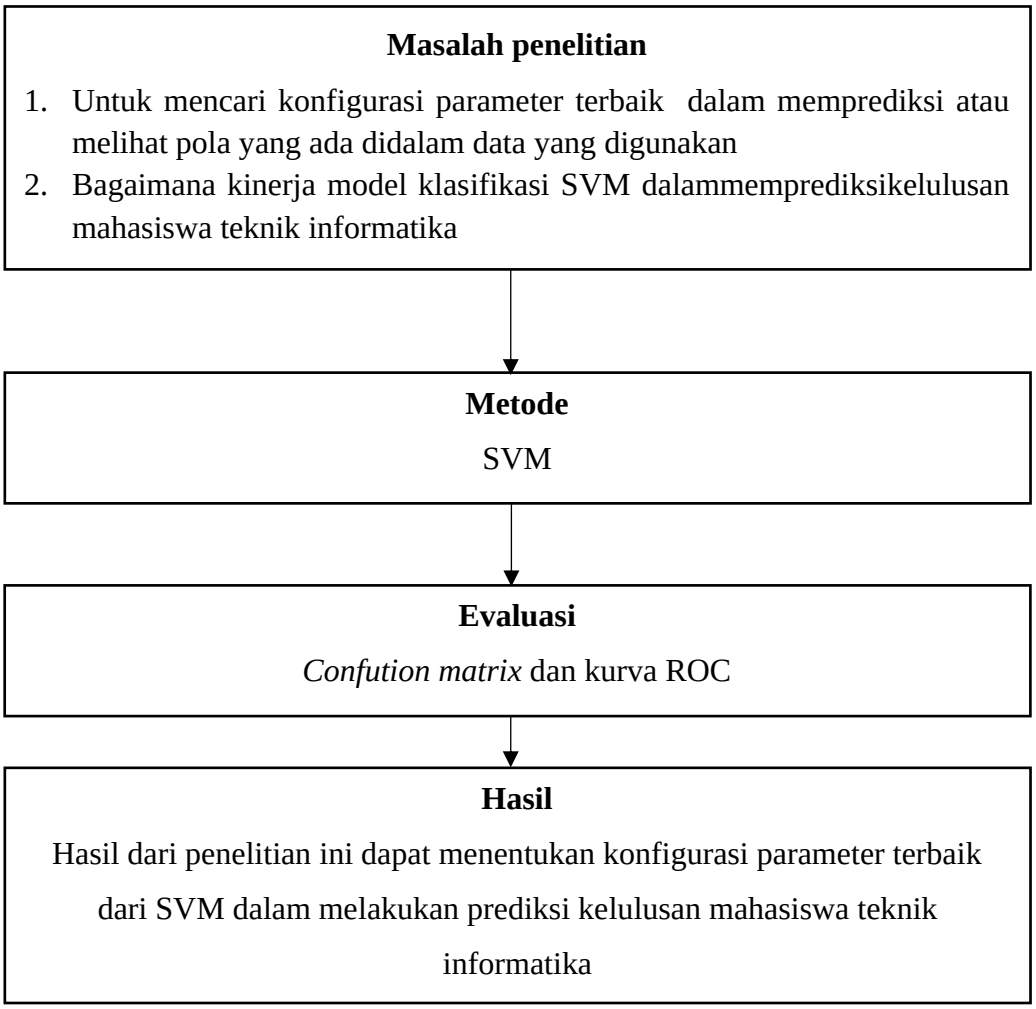
• 0,6 – 0,7 = klasifikasi rendah

• 0,5 – 0,6 = kegagalan

2.2.6 *One Hot Encoding*

One Hot Encoding adalah salah satu teknik untuk menyeragamkan data kategorikal dan numerik. Cara kerja *One Hot Encoding* adalah dengan membuat array 1 dimensi dengan selama ada jenis fitur dan memiliki konten biner antara 0 dan 1. *One Hot Encoding* dapat mewakili data yang diketik kategoris menjadi lebih ekspresif. Pada algoritma pembelajaran SVM tidak bisa bekerja dengan data kategorikal. Oleh karena itu, data kategorikal harus diubah menjadi data numerik yang berharga bilangan bulat 0 dan1. Menu selanjutnya mengubah setiap nilai di dalamnya kolom menjadi kolom baru dan diisi dengan nilai angka 0 dan 1 untuk fitur yang mempunyai nilai kategori tersebut.[17]

2.3 Kerangka Pikir



Gambar 2. 3 kerangka pikir