BAB II
LANDASAN TEORI

2.1. Landasan Teori

2.1.1 Sumbangan Pembinaan Pendidikan (SPP)

SPP (Sumbangan Pembinaan Pendidikan) adalah biaya rutin yang dibayarkan oleh siswa atau santri sebagai kontribusi dalam mendukung kegiatan operasional lembaga pendidikan. Di lingkungan pondok pesantren, SPP digunakan untuk mendanai berbagai kebutuhan pembelajaran santri, perawatan fasilitas, dan yang lainnya [5].

SPP memiliki peran penting dalam menjaga keberlangsungan kegiatan pendidikan di pondok pesantren. Keteraturan pembayaran SPP oleh santri memungkinkan pihak pesantren untuk mengelola keuangan secara stabil, merencanakan program-program pembinaan, serta memastikan tersedianya sarana dan prasarana yang memadai [6]. Oleh karena itu, keterlambatan dalam pembayaran SPP dapat berdampak pada terganggunya proses operasional dan pelayanan pendidikan, sehingga diperlukan upaya prediktif untuk mengidentifikasi potensi keterlambatan sebagai langkah antisipasi yang efektif.

2.1.2 Keterlambatan Pembayaran SPP

Keterlambatan pembayaran SPP adalah kondisi ketika santri tidak membayar SPP tepat waktu sesuai jadwal yang telah ditentukan [7]. Masalah keterlambatan ini dapat berdampak pada kelancaran kegiatan operasional pesantren. Oleh karena itu, penting untuk melakukan analisis dan klasifikasi keterlambatan guna membantu pengelolaan keuangan dan perencanaan kebijakan pendidikan.

Untuk mengatasi permasalahan tersebut, dibutuhkan pendekatan berbasis data yang mampu mengidentifikasi pola keterlambatan pembayaran dari data historis yang tersedia. Salah satu metode yang dapat digunakan adalah klasifikasi dalam data mining, yang memungkinkan prediksi terhadap status pembayaran santri di masa mendatang [8]. Dengan demikian, pihak pesantren dapat mengambil

langkah antisipatif lebih awal, seperti memberikan peringatan atau kebijakan khusus kepada wali santri yang berpotensi mengalami keterlambatan, sehingga stabilitas keuangan lembaga tetap terjaga.

2.1.3 Klasifikasi

Klasifikasi adalah salah satu metode dalam data mining yang bertujuan untuk memetakan data ke dalam kelompok atau kategori tertentu. Dalam konteks penelitian ini, klasifikasi digunakan untuk menentukan status pembayaran SPP santri apakah termasuk tepat waktu atau terlambat berdasarkan data historis yang tersedia [9].

Dengan menerapkan algoritma klasifikasi seperti K-Nearest Neighbor (K-NN), sistem dapat mengidentifikasi pola keterlambatan berdasarkan atribut tertentu seperti penghasilan orang tua, pekerjaan, dan data pribadi santri lainnya [10]. Hasil klasifikasi ini diharapkan dapat membantu pihak pesantren dalam melakukan prediksi dini terhadap potensi keterlambatan dan mengambil langkah pencegahan secara lebih efektif dan berbasis data.

2.1.4 Data Mining

Sistem Data mining adalah proses untuk menemukan pola-pola penting dari sejumlah besar data. Data mining membantu dalam proses pengambilan keputusan dengan cara menggali informasi tersembunyi dari data yang ada. Salah satu teknik dalam data mining adalah klasifikasi, yang digunakan dalam penelitian ini untuk menganalisis perilaku pembayaran [11]. Dalam pengembangannya, sistem data mining tidak terlepas dari beberapa landasan teori sebagai berikut:

a. Teori Statistik

   Statistik berperan penting dalam pengolahan data, seperti dalam proses estimasi, distribusi probabilitas, dan analisis variabel. Teknik statistik digunakan untuk mengenali pola dan tren dalam data yang besar [12].

b. Teori Pembelajaran Mesin *(Machine Learning)*

   Machine learning merupakan bagian dari kecerdasan buatan yang memungkinkan sistem belajar dari data historis untuk membuat prediksi atau

klasifikasi. Algoritma *K-Nearest Neighbor (K-NN)* yang digunakan dalam penelitian ini merupakan salah satu bentuk supervised learning [13].

c. Teori Basis Data dan Sistem Informasi

Teori ini mendasari bagaimana data disimpan, diakses, dan dikelola. Dalam data mining, sistem basis data sangat penting untuk pengambilan data secara efisien dan terstruktur sebelum dilakukan analisis lebih lanjut [14].

d. Teori Kecerdasan Buatan *(Artificial Intelligence)*

AI menjadi dasar dari pengembangan sistem yang mampu meniru cara berpikir manusia dalam menyelesaikan masalah. Data mining termasuk dalam lingkup AI karena melibatkan pemrosesan cerdas terhadap informasi yang kompleks [15].

e. Teori Informasi

Teori informasi berkaitan dengan pengukuran dan representasi data serta bagaimana informasi tersebut digunakan dalam proses pengambilan keputusan. Ini mencakup aspek nilai informasi, *entropi,* dan efisiensi informasi dalam konteks data mining [16].

2.1.5 *K-Nearest Neighbor* (K-NN)

K-Nearest Neighbor (K-NN) adalah salah satu algoritma klasifikasi yang sederhana namun efektif. K-NN bekerja dengan cara mencari sejumlah tetangga terdekat (berdasarkan nilai k) dari data baru yang akan diklasifikasikan, kemudian menetapkan label berdasarkan mayoritas label dari tetangga tersebut. K-NN banyak digunakan karena sifatnya yang non-parametrik dan mudah diimplementasikan [17].

Rumus *Euclidean Distance*:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Dimana:

p dan q adalah dua titik data

n adalah jumlah atribut atau fitur

p_i dan q_i adalah nilai pada atribut ke-i

Kelebihan dari algoritma *K-Nearest Neighbor (K-NN)* yaitu mudah untuk diimplementasikan, tidak memerlukan proses pelatihan, dan cocok untuk data dengan jumlah fitur yang tidak terlalu besar [18].

Sedangkan kekurangan dari *K-Nearest Neighbor (K-NN)* yaitu sensitif terhadap data yang tidak relevan atau memiliki skala berbeda, waktu prediksi lambat jika data terlalu besar, dan memerlukan pemilihan nilai k yang tepat agar tidak terjadi overfitting atau underfitting [19].

a. Teori Pembelajaran Berdasarkan Ilustrasi (*Instance-Based Learning*) / Pembelajaran Malas (*Lazy Learning*)

   K-NN termasuk dalam kelompok *lazy learning*, yaitu metode yang tidak membangun model eksplisit saat pelatihan, melainkan menyimpan seluruh data latih dan melakukan proses klasifikasi langsung saat data uji diberikan. Algoritma ini mengandalkan contoh-contoh sebelumnya (*instances*) sebagai dasar untuk pengambilan keputusan [20].

b. Teori Jarak (*Distance Metric*)

   Penentuan tetangga terdekat didasarkan pada perhitungan jarak antar titik data. Dalam konteks K-NN, metrik jarak seperti *Euclidean Distance*, Manhattan Distance, atau Minkowski Distance digunakan untuk mengukur kemiripan antara data baru dan data pelatihan. Pemilihan jenis jarak sangat mempengaruhi hasil klasifikasi [21].

c. Teori Penguatan Suara Mayoritas / Rata-rata Berbobot

   Keputusan klasifikasi dalam K-NN dilakukan dengan prinsip mayoritas suara (*majority voting*). Dalam beberapa varian, pemberian bobot berdasarkan jarak juga digunakan, yaitu semakin dekat tetangga, semakin besar bobotnya dalam mempengaruhi keputusan klasifikasi (*weighted voting*). Ini meningkatkan akurasi terutama saat distribusi data tidak merata [22].

d. Asumsi dan Pertimbangan dalam K-NN

   Dalam penerapan algoritma *K-Nearest Neighbor (K-NN)*, terdapat sejumlah asumsi dasar dan pertimbangan penting yang perlu diperhatikan agar

algoritma dapat bekerja secara optimal dan menghasilkan klasifikasi yang akurat. Adapun asumsi dan pertimbangan tersebut sebagai berikut:

1. Data yang serupa akan berada dalam jarak yang dekat (*locality assumption*).
2. Semua fitur dianggap memiliki kontribusi yang sama, sehingga penting untuk melakukan normalisasi atau standardisasi data sebelum digunakan.
3. Nilai *k* harus dipilih dengan hati-hati: terlalu kecil menyebabkan overfitting, terlalu besar bisa menyebabkan underfitting.
4. K-NN bekerja lebih baik pada dataset yang bersih dan memiliki atribut yang relevan.

2.1.6 Confusion Matrix

Confusion Matrix merupakan sebuah tabel yang digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi model terhadap nilai aktual dari data. Tabel ini menunjukkan jumlah data yang benar dan salah diklasifikasikan oleh model ke dalam masing-masing kategori uji [23]. Tabel *confusion matrix* ditunjukan pada tabel 2.1 sebagai berikut:

Tabel 2.1. *Confusion Matrix*

	Prediksi Positif	**Prediksi Negatif**
Aktual Positif	True Positive (TP)	False Negative (FN)
Aktual Negatif	False Positive (FP)	True Negative (TN)

1. *True Positive (TP)*

   Jumlah data yang diklasifikasikan benar sebagai positif.

   Contoh: Nasabah layak pinjam dan model memprediksi layak.

2. *False Positive (FP)*

   Jumlah data yang diklasifikasikan salah sebagai positif.

   Contoh: Nasabah tidak layak pinjam tapi model memprediksi layak. *(Juga disebut Type I Error)*

3. *False Negative (FN)*

   Jumlah data yang diklasifikasikan salah sebagai negatif.

   Contoh: Nasabah layak pinjam tapi model memprediksi tidak layak. *(Type II Error)*

4. *True Negative (TN)*

   Jumlah data yang diklasifikasikan benar sebagai negatif.

   Contoh: Nasabah tidak layak pinjam dan model memprediksi tidak layak.

Dari keempat nilai ini, dapat dihitung beberapa metrik evaluasi penting:

1. *Akurasi*

   Mengukur seberapa sering model membuat prediksi yang benar.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2. *Presisi*

   Mengukur seberapa akurat prediksi positif dari model.

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

3. *Recall*

   Mengukur seberapa baik model menemukan semua data positif.

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

4. *FI-Score*

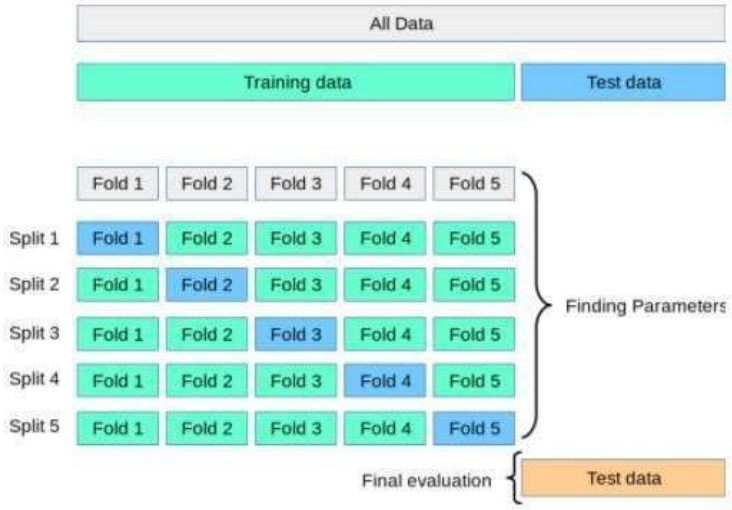
   Merupakan harmonisasi antara presisi dan recall.

$$F1 = 2\,x \frac{Presisi\ X\ Recall}{Presisi + Recall}$$

2.1.7 *K-Fold* Cross-*Validation*

K-Fold Cross-Validation adalah teknik evaluasi model yang digunakan untuk mengukur performa algoritma pembelajaran mesin secara lebih akurat dan stabil. Teknik ini bekerja dengan membagi dataset menjadi *K* bagian (fold) yang kurang lebih sama besar. Proses pelatihan dan pengujian dilakukan sebanyak *K* kali,

di mana pada setiap iterasi satu fold digunakan sebagai data uji, sedangkan *K–1 fold* lainnya digunakan sebagai data latih [24]. Hasil evaluasi dari setiap fold kemudian dirata-rata untuk mendapatkan estimasi performa model yang lebih representatif.



Gambar 2.1. Simulasi *K-Fold Cross Validation*

Dalam penerapannya, *K-Fold Cross-Validation* didukung oleh beberapa teori penting sebagai berikut:

a. Teori Generalisasi dan Estimasi Kinerja Model

Cross-validation merupakan salah satu metode utama dalam mengukur kemampuan generalisasi dari suatu model, yaitu sejauh mana model yang dibangun dari data pelatihan mampu memberikan prediksi yang akurat terhadap data yang belum pernah dilihat. Teknik ini membantu memperkirakan kinerja model secara realistis, bukan hanya pada data yang sudah dikenali, tetapi juga pada data yang benar-benar baru [25].

b. Teori Pembagian Data

Dalam pembelajaran mesin, data perlu dibagi secara strategis agar tidak terjadi overlap antara data pelatihan dan pengujian, yang dapat menyebabkan evaluasi kinerja yang bias. *K-Fold Cross-Validation* mengikuti prinsip pembagian data yang adil dan merata, sehingga setiap data memiliki kesempatan yang sama untuk digunakan sebagai data uji, dan membantu mengevaluasi performa model di seluruh distribusi data [26].

c. Teori *Bias-Variance Trade-Off*

Evaluasi model juga erat kaitannya dengan bias dan varians. Model dengan bias tinggi cenderung terlalu sederhana dan tidak mampu menangkap pola dalam data, sementara model dengan varians tinggi terlalu kompleks dan overfit terhadap data latih. K-Fold Cross-Validation digunakan untuk menyeimbangkan kedua hal tersebut dengan memberikan evaluasi dari berbagai subset data, sehingga bisa meminimalisir bias evaluasi tunggal dan mengurangi fluktuasi varians [27].

d. Varian Pembagian Klaster (*Stratified K-Fold*)

Dalam kasus klasifikasi, khususnya ketika terdapat ketidakseimbangan kelas *(imbalanced class),* digunakan varian *Stratified K-Fold*, yaitu teknik pembagian fold yang mempertahankan proporsi label atau kelas pada setiap fold agar konsisten dengan keseluruhan dataset. Hal ini penting untuk memastikan bahwa setiap fold mewakili distribusi data secara adil, dan hasil evaluasi tidak bias terhadap kelas dominan [28].

e. Signifikansi dan Implikasi

Penerapan *K-Fold Cross-Validation* memberikan implikasi signifikan dalam validitas hasil penelitian, karena mengurangi risiko kesalahan pengambilan keputusan akibat evaluasi yang keliru. Teknik ini juga memperkuat keandalan hasil evaluasi, memungkinkan perbandingan antar model atau parameter dilakukan secara objektif. Dalam konteks penelitian pendidikan berbasis data, validasi silang memastikan bahwa sistem klasifikasi yang dibangun memiliki daya prediktif yang stabil dan terpercaya [29].

2.1.8 *Google Colaboratory* (Google Colab)

Google Colaboratory (Google Colab) adalah platform berbasis cloud yang menyediakan lingkungan pemrograman *Phyton* secara online. Colab memungkinkan penulisan dan eksekusi kode Phyton langsung dari browser tanpa perlu instalasi lokal [30]. Keunggulan Google Colab adalah sebagai berikut:

1. Gratis dan mudah digunakan.
2. Mendukung GPU/TPU untuk komputasi cepat.
3. Dapat digunakan untuk kolaborasi dalam penelitian.

Dalam mendukung kelancaran proses validasi dan pengembangan model klasifikasi keterlambatan pembayaran SPP berbasis algoritma *K-Nearest Neighbor* (K-NN), penelitian ini tidak hanya memanfaatkan konsep dasar klasifikasi dan evaluasi model, tetapi juga didukung oleh berbagai teori dan pendekatan teknologi modern. Beberapa teori tersebut adalah sebagai berikut:

a. Teori Konfigurasi Awan (*Cloud Computing*)

Cloud computing merupakan model komputasi modern yang memungkinkan pengguna untuk mengakses sumber daya komputasi secara fleksibel melalui internet, tanpa perlu memiliki infrastruktur fisik secara langsung. Dalam konteks penelitian ini, cloud computing digunakan untuk menjalankan pemodelan algoritma *K-Nearest Neighbor (K-NN)* melalui platform *seperti Google Colaboratory.* Platform tersebut menyediakan lingkungan pemrograman berbasis cloud yang mendukung *Python,* memungkinkan pelaksanaan proses komputasi dan analisis data yang cukup besar tanpa membutuhkan perangkat keras lokal dengan spesifikasi tinggi. Keunggulan ini menjadikan *cloud computing* sangat efisien dalam menunjang eksperimen dan validasi model berbasis data mining di bidang pendidikan [31].

b. Teori Sistem Terdistribusi dan Visualisasi

Sistem terdistribusi adalah sekumpulan komputer yang terhubung dalam suatu jaringan dan bekerja bersama untuk menyelesaikan suatu tugas secara paralel. Dalam implementasi teknik validasi seperti *K-Fold Cross Validation*, prinsip ini mendukung proses pembagian dan pengolahan data secara efisien untuk mempercepat pelatihan dan evaluasi model. Di sisi lain, visualisasi data memegang peran penting dalam menjelaskan hasil analisis secara intuitif. Penggunaan grafik akurasi, *confusion matrix*, dan visual interaktif lainnya mempermudah peneliti dalam memahami performa model dan pola data, sehingga mempermudah pengambilan keputusan berbasis analitik [32].

c. Teori Notebook Komputasi Interaktif

Notebook komputasi interaktif seperti *Jupyter Notebook* dan Google Colab memberikan lingkungan terpadu untuk menulis kode, menjalankan analisis data, menampilkan grafik, serta menyisipkan penjelasan dalam satu dokumen. Dalam

penelitian ini, penggunaan notebook tersebut sangat mendukung proses eksplorasi dan dokumentasi model K-NN secara real time. Setiap tahapan, mulai dari preprocessing, pelatihan, validasi, hingga visualisasi, dapat dijalankan dan disesuaikan secara langsung. Hal ini tidak hanya meningkatkan produktivitas peneliti, tetapi juga membuat proses dokumentasi menjadi lebih transparan, terstruktur, dan mudah dipahami oleh pihak lain yang ingin mereplikasi penelitian ini [33].

d. Teori Pembelajaran dan *Deep Learning*

Pembelajaran mesin *(machine learning)* merupakan bagian dari kecerdasan buatan yang memungkinkan komputer untuk belajar dari data dan membuat prediksi tanpa pemrograman eksplisit. Meskipun algoritma K-NN termasuk dalam kategori *machine learning* tradisional dan bukan *deep learning*, prinsip pembelajaran tetap berlaku, yakni dengan memanfaatkan data historis untuk membangun pola klasifikasi. *Deep learning* sendiri merupakan bagian lanjutan dari machine learning yang menekankan pada penggunaan jaringan saraf berlapis untuk analisis data kompleks. Dalam konteks penelitian ini, pemahaman terhadap konsep pembelajaran berulang dan evaluasi performa secara iteratif menjadi dasar penting dalam membangun dan menyempurnakan model klasifikasi [34].

e. Teori Kolaborasi dan Versi Kontrol

Pengelolaan proyek berbasis data secara kolaboratif membutuhkan sistem yang mampu melacak perubahan dan mendokumentasikan perkembangan kerja secara sistematis. Sistem kontrol versi seperti *Git* dan *platform* kolaborasi seperti *GitHub* menjadi solusi yang sangat efektif dalam mendukung kerja tim dalam penelitian ini. Dengan adanya fitur *commit, branch*, dan *merge*, setiap anggota tim dapat bekerja secara paralel, melakukan eksperimen mandiri, dan kembali menyatukan hasil kerja tanpa risiko kehilangan data atau konflik. Teori ini memperkuat aspek *reproducibility* dan transparansi dalam penelitian ilmiah, serta menjadi fondasi penting dalam pengembangan sistem klasifikasi berbasis data yang kompleks dan dinamis [35].

2.1.9 Phyton

Phyton merupakan bahasa pemrograman tingkat tinggi yang banyak digunakan dalam analisis data, machine learning, dan data mining [36]. Phyton memiliki banyak pustaka yang mendukung proses klasifikasi sebagai berikut:

1. *Pandas* untuk manipulasi data.
2. *Scikit-learn* untuk implementasi algoritma klasifikasi seperti *K-Nearest Neighbor (K-NN).*
3. *NumPy* untuk komputasi numerik.
4. *Matplotlib* dan *Seaborn* untuk visualisasi data.

Selain itu, penggunaan Python dalam penelitian ini juga didukung oleh berbagai teori sebagai landasan konseptual, sebagai berikut:

a. Teori Bahasa Pemrograman dan Desain Kompiler/Interpreter

   Python dikembangkan dengan prinsip interpretasi, di mana kode dieksekusi baris demi baris. Hal ini memberikan fleksibilitas dalam eksperimen data secara langsung, serta memudahkan proses debugging dan pengembangan model klasifikasi [37].

b. Paradigma Pemrograman

   Python mendukung berbagai paradigma pemrograman seperti imperatif, objektif, dan fungsional. Fleksibilitas ini memudahkan peneliti untuk menyesuaikan gaya pemrograman dengan kebutuhan proyek, baik dalam eksplorasi data, pembuatan fungsi, maupun pengembangan modular [38].

c. Teori Struktur Data dan Algoritma

   Python menyediakan struktur data dasar seperti list, tuple, set, dan dictionary, serta mendukung implementasi berbagai algoritma penting dalam klasifikasi dan optimasi. Pustaka seperti NumPy dan Scikit-learn memperluas kapabilitas ini untuk keperluan komputasi ilmiah [39].

d. Teori Komunitas dan Ekosistem Open Source

   Python memiliki komunitas pengguna yang besar dan aktif di seluruh dunia. Ribuan pustaka open source tersedia secara gratis dan terus diperbarui,

menjadikan Python ekosistem yang sangat mendukung penelitian ilmiah dan pengembangan teknologi [39].

e. Penerapan Ilmiah dan Data Science

Python saat ini menjadi bahasa utama dalam dunia data science dan AI karena kemudahan penggunaan, dukungan pustaka yang lengkap, serta dokumentasi yang luas. Bahasa ini memungkinkan integrasi langsung antara eksplorasi data, visualisasi, pemodelan machine learning, dan pelaporan hasil dalam satu lingkungan kerja .

2.2. Penelitian Terdahulu

Dalam penelitian ini, peneliti menggali informasi dari penelitian-penelitian sebelumnya guna mendapatkan teori yang berkaitan dengan judul yang digunakan diantaranya sebagai berikut.

1. Penelitian yang dilakukan oleh T. A. Y. Siswa dan R. P. Wibowo membahas tantangan dalam studi *data mining* terkait keterlambatan pembayaran SPP, khususnya pada *dataset* berdimensi tinggi yang sering menghasilkan akurasi di bawah 60%. Selain itu, penelitian tentang hubungan antar atribut dalam pemodelan klasifikasi masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk menganalisis peningkatan akurasi berbagai algoritma klasifikasi, yaitu *K-Nearest Neighbor*, *Naive Bayes, C4.5*, *Random Forest*, dan *Logistic Regression*, dengan mengoptimalkannya menggunakan perbandingan algoritma seleksi fitur seperti *Mutual Information, Forward Selection, Backward Elimination*, dan *Recursive Elimination*. Data yang digunakan adalah data pembayaran SPP mahasiswa dari tahun 2019 hingga 2021, dengan pembagian data menggunakan metode *5-fold cross-validation*. Hasilnya menunjukkan bahwa algoritma *Backward Elimination* memberikan peningkatan akurasi tertinggi sebesar 0,52%. Sementara itu, algoritma *Random Forest* dan *C4.5* menunjukkan akurasi tertinggi secara keseluruhan, yaitu 62,6%, dengan *presisi* 65%, *recall* 63%, dan *F1-score* 61% [40].

2. Penelitian terdahulu yang relevan dengan klasifikasi keterlambatan pembayaran SPP sebagian besar menggunakan algoritma *K-Nearest Neighbor (KNN)* sebagai metode utama. Salah satunya adalah penelitian oleh K. Samruddhi, R. Ashok Kumar yang menerapkan KNN murni untuk prediksi harga mobil bekas dengan evaluasi *K-Fold Cross Validation* sebagai metode pengujian performa. Hasil penelitian tersebut menunjukkan akurasi sekitar 85% dengan rata-rata hasil cross-validation di kisaran 80–82%, yang menegaskan bahwa KNN dapat menghasilkan performa cukup baik meskipun belum optimal. Selain itu, penggunaan cross-validation memberikan gambaran yang lebih menyeluruh mengenai kemampuan generalisasi model, karena performa tidak hanya diukur dari satu skenario pembagian data, melainkan dari beberapa kombinasi data latih dan uji. Dengan pendekatan ini, KNN mampu memanfaatkan pola kedekatan antar data secara efektif untuk menyelesaikan kasus klasifikasi dengan karakteristik data numerik dan kategorik yang sederhana [41].
3. Begitu juga penelitian yang dilakukan oleh M. D. Anggraeni, K. Kusrini, dan M. R. Arief mengulas sistem pembiayaan di SMK Ma'arif Salam, sebuah lembaga pendidikan swasta yang sebagian besar operasionalnya dibebankan pada siswa melalui pembayaran SPP, praktik umum di sekolah swasta yang bertanggung jawab atas kebijakan pembiayaan lokal, berbeda dengan sekolah negeri yang didanai pemerintah. Studi ini menggunakan Algoritma K-Nearest Neighbor (KNN) untuk klasifikasi. Dengan menetapkan nilai k=3, penelitian ini menghasilkan *dataset* murni. Proses klasifikasi dilakukan dengan membagi data menjadi 80% data pelatihan (406 *record*) dan 20% data pengujian (102 *record*). Hasil perhitungan menggunakan algoritma KNN menunjukkan tingkat akurasi sebesar 82,35% [42].

4. Penelitian lain membahas tantangan biaya sumbangan pembinaan pendidikan, khususnya terkait pembayaran SPP di sekolah swasta seperti SMK Wirasaba Karawang. Dengan menganalisis 725 data pembayaran SPP siswa selama satu semester di tahun 2023, yang menunjukkan sekitar 22% siswa mengalami keterlambatan, penelitian ini bertujuan untuk meningkatkan efisiensi administrasi melalui pendekatan klasifikasi guna memprediksi keterlambatan pembayaran. Harapannya, pemahaman pola keterlambatan dapat menghasilkan solusi preventif yang efektif, sehingga berkontribusi pada pengembangan sistem administrasi pendidikan yang lebih efisien dan meningkatkan layanan pendidikan secara keseluruhan di Indonesia. Kontribusi utama dari penelitian ini adalah pengembangan sistem administrasi pendidikan yang lebih efisien melalui pendekatan teknologi informasi, didukung oleh analisis pola keterlambatan pembayaran berdasarkan data aktual dari SMK Wirasaba Karawang [43].
5. Kemudian penelitian yang dilakukan oleh R. W. Abdullah, Kusrini, dan E. T. Luthfii membahas masalah keterlambatan pembayaran SPP di SMK Al-Islam Surakarta, yang berdampak pada terganggunya kegiatan operasional sekolah karena sebagian besar dana SPP digunakan untuk pengembangan fasilitas dan infrastruktur. Untuk mengatasi hal ini, diperlukan tindak lanjut terhadap orang tua/wali siswa yang terlambat membayar SPP, yang dapat diprediksi menggunakan metode K-Nearest Neighbor (KNN). Untuk memprediksi keterlambatan pembayaran SPP, parameter yang digunakan adalah Penghasilan, Pendidikan, Tanggungan Keluarga, dan Usia dalam menghitung jarak terdekat antara data pelatihan dan data pengujian. Tujuan dari penelitian ini adalah untuk menentukan akurasi hasil prediksi keterlambatan pembayaran SPP dengan metode KNN. Hasilnya, didapatkan akurasi sebesar 86%. Prediksi ini diharapkan dapat digunakan oleh sekolah untuk memberikan surat pemberitahuan pembayaran SPP kepada calon wali siswa, sehingga

mereka tidak mengalami keterlambatan saat jadwal pembayaran tiba [44].