

Implementasi *K-Nearest Neighbor* Dalam Memprediksi Keterlambatan Pembayaran Biaya Kuliah Di Perguruan Tinggi

Muhammad Rhosyid Akhmad^{1*}, Taghfirul Azhima Yoga Siswa²

Program Studi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur
 Jl. Ir. H. Juanda No.15 Sidodadi, Samarinda, Indonesia

*Email Corresponding Author: 1811102441067@umkt.ac.id

Abstract

The delay in payment of tuition fees at the Muhammadiyah University of East Kalimantan for the 2020/2021 academic year reached 3,018 out of a total of 5,533 students. The number of latecomers is quite large because it exceeds half of the total students. It is deemed necessary to conduct an analysis related to the late payment, so that prevention and treatment can be carried out as early as possible. The purpose of this study is to determine the parameters of delay in paying tuition fees, implementing the K-Nearest Neighbor algorithm, and evaluating the performance of the algorithm using a confusion matrix. The amount of data used for the algorithm performance testing process is 12,408 records with a ratio of 80% training data and 20% testing data. The results of the evaluation test showed that the accuracy of k=3 was 52.82%, k=5 was 52.49%, k=7 was 52.37%, k=9 was 52.33%, and k=11 was 52.53%. The best data test results were obtained at k = 3, namely 52.82%.

Keywords: Accuracy; K-Nearest Neighbor; Tuition Fee Payment, Confusion matrix

Abstrak

Keterlambatan pembayaran biaya kuliah di Universitas Muhammadiyah Kalimantan Timur tahun ajaran 2020/2021 mencapai 3.018 dari total keseluruhan 5.533 mahasiswa. Jumlah yang terlambat tersebut tergolong cukup banyak karena melebihi separuh dari keseluruhan mahasiswa. Dipandang perlu untuk melakukan analisis berkaitan dengan keterlambatan pembayaran tersebut, agar dapat dilakukan pencegahan dan penanganan sedini mungkin. Tujuan penelitian ini adalah menentukan parameter keterlambatan dari pembayaran biaya kuliah, mengimplementasi algoritma *K-Nearest Neighbor*, dan mengevaluasi kinerja algoritma menggunakan *Confusion Matrix*. Jumlah data yang digunakan untuk proses pengujian kinerja algoritma adalah 12.408 *record* dengan rasio 80% data *training* dan 20% data *testing*. Hasil pengujian evaluasi didapatkan akurasi k=3 sebesar 52,82%, k=5 sebesar 52,49%, k=7 sebesar 52,37%, k=9 sebesar 52,33%, dan k=11 sebesar 52,53%. Hasil pengujian data terbaik didapatkan pada nilai k=3 yaitu 52,82%.

Kata kunci: Akurasi; K-Nearest Neighbor; Pembayaran biaya kuliah, Confusion matrix

1. Pendahuluan

Sumber pendanaan memegang peranan penting dalam keberlangsungan kegiatan operasional perguruan tinggi. Salah satu sumber pendanaan bagi perguruan tinggi swasta adalah dari biaya perkuliahan. Universitas Muhammadiyah Kalimantan Timur (UMKT) merupakan perguruan tinggi swasta milik Muhammadiyah sebagai salah satu amal usaha yang bergerak di bidang pendidikan. Dalam meningkatkan kualitas pembelajaran UMKT terus melakukan upaya perkembangan seperti pembangunan infrastruktur, perbaikan sarana prasarana, dan penambahan jumlah dosen. Dana terbesar yang digunakan adalah keuangan mahasiswa yang bersumber dari pembayaran Sumbangan Pembangunan Pendidikan (SPP). Keterlambatan pembayaran SPP dapat mengakibatkan terhambatnya kegiatan operasional dan mempengaruhi progress akademik karena tidak dapat mengurus Kartu Rencana Studi (KRS) pada awal semester. Berdasarkan data dari bagian keuangan mahasiswa UMKT yang terlambat melakukan pembayaran SPP tahun ajaran 2020/2021 adalah 3.018 dari total keseluruhan 5.533 mahasiswa. Jumlah mahasiswa yang terlambat tersebut cukup banyak dikarenakan melebihi setengah dari total keseluruhan jumlah mahasiswa. Untuk dapat menganalisis permasalahan tersebut, maka perlu dilakukan prediksi agar keterlambatan pembayaran SPP dapat dilakukan pencegahan dan penanganan sedini mungkin.

K-Nearest Neighbor pada umumnya sudah banyak digunakan peneliti karena mempunyai kelebihan, yaitu nilai akurasi yang tinggi dan tidak ada asumsi pada data [1]. Penelitian sebelumnya yang menggunakan *K-Nearest Neighbor* yaitu prediksi hasil produksi mesin pada PT. Showa Katou Indonesia dengan hasil akurasi 100% [2], prediksi kelulusan mahasiswa magister Teknik Informatika Universitas Amikom Yogyakarta dengan akurasi 98,46% [3], memprediksi prestasi mahasiswa berdasarkan latar belakang pendidikan dan ekonomi dengan hasil akurasi 95.85% [4], prediksi kelulusan siswa pada SMK Anak Bangsa dengan hasil akurasi 93,55%, [5], Prediksi Tingkat Identifikasi Tindakan Berbasis Guru Tentang Komitmen Organisasi Dan Kepuasan Kerja dengan nilai akurasi 93,6% [6], prediksi penjualan produk *Unilever* dengan hasil akurasi 86,66% [7], aplikasi prediksi kelulusan dengan pengujian yang dilakukan menggunakan *K-Fold Cross Validation* didapatkan akurasi tertinggi pada model yang ketiga sebesar 80% [8], Prediksi Potensi Siswa Putus Sekolah Akibat Pandemi Covid-19 dengan nilai akurasi 87.4214% [8], prediksi kelompok UKT mahasiswa dengan hasil akurasi 84,21% [9], dan sebagai sistem prediksi predikat prestasi mahasiswa dengan hasil akurasi 82% [10].

Penelitian ini menguji algoritma *K-Nearest Neighbor* dalam melakukan prediksi keterlambatan pembayaran SPP di perguruan tinggi, agar dapat dilakukan pencegahan dan penanganan sedini mungkin atas kasus yang sama di masa mendatang.

2. Tinjauan Pustaka

Prediksi keterlambatan pembayaran SPP dengan metode *K-Nearest Neighbor* dengan data 236 data siswa Angkatan 2017-2018. Digunakan 30 data sebagai testing dan 236 sebagai training. Nilai akurasi yang didapatkan $K=3$ sebesar 86% [11]. Menganalisis kemungkinan keterlambatan pembayaran SPP dengan algoritma C4.5 (Studi Kasus Politeknik Tecd Bandung). Data yang digunakan dalam penelitian ini berjumlah 102 data yang didapatkan dari kuesioner yang disebarakan kepada mahasiswa program studi Teknik Informatika semester 3,5, dan 7 kelas reguler tahun ajaran 2015/2016. Dari total sampel yang digunakan kemudian dibagi menjadi 4 partisi, partisi 1 untuk 90% data training dan 10% data testing, partisi 2 untuk 80% data training dan 20% data testing, dan partisi 3 untuk 70% data training dan 30% data testing, dan partisi 4 untuk 60% data training dan 40% data testing. Hasil uji coba partisi data set ditemukan tingkat akurasi tertinggi pada data partisi 2 sebanyak 80% data training dan menghasilkan akurasi sebesar 75,00% [12].

Selanjutnya menganalisis kemungkinan keterlambatan pembayaran SPP menggunakan algoritma *Support Vector Machine*. Data yang digunakan merupakan siswa SMP Perintis 2 Bandar dengan akurasi terbaik sebesar 97.8469%, *error rate* 0.021531, *false positive rate* 0.005208, *recall* 0.964602, *specificity* 0.994792, dan *precision* 0.99543 [13]. Optimasi Metode *Naïve Bayes* Dengan *Feature Selection Information Gain* Untuk Prediksi Keterlambatan Pembayaran Sumbangan Pembinaan Pendidikan Sekolah. Data yang digunakan merupakan data dapodik sekolah tahun 2017/2018 dengan dataset uji sebanyak 30 *record*. Hasil penelitian menunjukkan bahwa akurasi tertinggi diperoleh dengan menggabungkan *Naïve Bayes* Metode dengan *Information Gain Method* diperoleh akurasi 90% [14].

Dari beberapa penelitian terkait sebelumnya dapat disimpulkan bahwa penelitian tentang keterlambatan pembayaran SPP sudah pernah dilakukan, namun metode penyelesaian yang digunakan berbeda-beda untuk setiap studi kasusnya. Pada dasarnya metode *K-Nearest Neighbor* telah digunakan namun hanya menggunakan satu nilai k dan lebih banyak menganalisa data numerik. Penelitian ini merujuk pada penelitian sebelumnya dengan menggunakan metode *K-Nearest Neighbor* namun dengan menggunakan beberapa uji coba untuk menentukan nilai k yang terbaik, kemudian perbedaan selanjutnya pada analisis pada tipe data kategorikal (*ordinal encoding*) untuk dapat diuji coba dalam memperoleh nilai akurasinya.

3. Metodologi

3.1. Pengumpulan Data

Pengumpulan data pada penelitian ini adalah Bagian Administrasi Akademik berupa data keseluruhan mahasiswa dan Bagian Keuangan berupa data pembayaran SPP mahasiswa periode 2019 – 2021. Data mahasiswa yang diperoleh dari Bagian Administrasi Akademik adalah sebanyak 10.959. Kemudian data yang diperoleh dari Bagian Keuangan sebanyak 8.833 mahasiswa terlambat dan 30.811 mahasiswa tepat waktu dengan total 39.644 yang melakukan pembayaran SPP selama periode 2019-2021. Data pembayaran SPP yang diperoleh dari Bagian Keuangan terdapat 3 atribut diantaranya yaitu nim, nama, dan label (terlambat atau tepat) dapat dilihat pada Tabel 1. Sedangkan data mahasiswa yang diperoleh dari Bagian Administrasi Akademik terdapat 5 atribut diantaranya yaitu nim, nama, fakultas, prodi, angkatan, gender, pendapatan ayah, pendapatan ibu, pendidikan ayah, dan pendidikan ibu dapat dilihat pada Tabel 2.

Tabel 1. Ilustrasi Data Pembayaran SPP Tepat Waktu Dan Terlambat

No	NIM	Nama Mahasiswa	Label
1	1811102411001	Abdul Muis	Tepat
2	1811102411002	Adam Muh. Agussalim	Tepat
3	1811102411003	Ahmad Yusuf	Tepat
4	1811102411004	Angga Septian Budi Saputra	Tepat
5	1811102411005	Annisa Yuliah	Tepat
⋮	⋮	⋮	⋮
39644	17111024430021	Robby Hadi Cahyono	Terlambat

Tabel 2. Ilustrasi Data Keseluruhan Mahasiswa

No	NIM	Nama	Fakultas	Prodi	Angkatan	Gender	Penghasilan Ayah	Penghasilan Ibu	Pendidikan Ayah	Pendidikan Ibu
1	17111024110001	Abdul Hair	Ilmu Keperawatan	Keperawatan	2017	L	Rp. 2,000,000 - Rp. 4,999,999	Kurang dari Rp. 500,000	SD	SD
2	17111024110002	Adela Rosnadia	Ilmu Keperawatan	Keperawatan	2017	P	Rp. 2,000,000 - Rp. 4,999,999	Kurang dari Rp. 500,000	S1	SMA
3	17111024110003	Affiah Nida Fauziah	Ilmu Keperawatan	Keperawatan	2017	P	Rp. 2,000,000 - Rp. 4,999,999	Rp. 2,000,000 - Rp. 4,999,999	D3	S1
4	17111024110004	Afina Agma Fazalina	Ilmu Keperawatan	Keperawatan	2017	P	Kurang dari Rp. 500,000	Kurang dari Rp. 500,000	Tidak sekolah	Tidak sekolah
5	17111024110005	Afira Khofifah	Ilmu Keperawatan	Keperawatan	2017	P	Rp. 2,000,000 - Rp. 4,999,999	Rp. 2,000,000 - Rp. 4,999,999	S1	S1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10959	2111102444012	Dea Thania	Sains Dan Teknologi	Teknik Geologi	2021	P	Rp. 5,000,000 - Rp. 20,000,000	Kurang dari Rp. 500,000	SMA	SMA

3.2. Persiapan Data

Proses persiapan data dilakukan dengan beberapa tahapan yaitu integrasi data, pembersihan data, seleksi data, reduksi data, dan transformasi data dengan prosesnya sebagai berikut :

1) Integrasi Data

Tahap intergrasi data dilakukan untuk menggabungkan data dari Bagian keuangan dan Bagian Administrasi Akademik. Data yang didapatkan dari bagian keuangan memiliki atribut yaitu nim, nama, label kemudian data yang didapatkan dari Bagian Administrasi Akademik memiliki atribut yaitu nama, nim, fakultas, program studi, angkatan, pendidikan ayah, penghasilan ayah, pendidikan ibu, penghasilan ibu. Data digabungkan dengan cara melihat kecocokan nim sehingga atribut keseluruhan menjadi nama, nim, fakultas, program studi, angkatan, pendidikan ayah, penghasilan ayah, pendidikan ibu, penghasilan ibu dan label.

2) Pembersihan Data

Tahap pembersihan data dilakukan untuk menghapus data yang tidak lengkap atau kosong. Data awal sebelum dibersihkan berjumlah 39.644 baris data dan setelah dilakukan pembersihan menjadi 29.545 baris data dengan jumlah data tepat 23.341 dan terlambat berjumlah 6.204.

3) Seleksi Data dan Reduksi Data

Tahap seleksi dan reduksi data dilakukan agar memaksimalkan pemodelan yang akan dilakukan. Atribut yang digunakan setelah diseleksi yaitu menjadi pendapatan ayah, pendapatan ibu, pendidikan ayah, pendidikan ibu dan label. Kemudian direduksi dengan melakukan penyeimbangan data dikarenakan jumlah data tepat adalah 23.341 dan terlambat 6.204 tidak seimbang. Sehingga data tepat dikurangi sejumlah data terlambat dengan total akhir dataset adalah 12.408. Proses reduksi merujuk pada penelitian [15] dimana dataset yang memiliki kelas target yang paling banyak (mayoritas) akan menyebabkan pemodelan menjadi bias terhadap kelas target minoritas.

4) Transformasi

Tahap transformasi data dilakukan agar algoritma K-Nearest Neighbor dapat melakukan pengolahan data. Transformasi pada penelitian ini menggunakan metode ordinal encoding. Atribut yang diubah menjadi numerik yaitu pendapatan ayah, pendapatan ibu,

pendidikan ayah, dan pendidikan ibu dengan prosesnya dapat dilihat pada Tabel 3 hingga Tabel 6.

Tabel 3. Transformasi Pendapatan Ayah

Pendapatan Ayah	Transformasi
Kurang dari Rp. 500,000	0
Rp. 500,000 - Rp. 999,999	1
Rp. 1,000,000 - Rp. 1,999,999	2
Rp. 2,000,000 - Rp. 4,999,999	3
Rp. 5,000,000 - Rp. 20,000,000	4
Lebih dari Rp. 20,000,000	5

Tabel 4. Transformasi Pendapatan Ibu

Pendapatan Ibu	Transformasi
Kurang dari Rp. 500,000	0
Rp. 500,000 - Rp. 999,999	1
Rp. 1,000,000 - Rp. 1,999,999	2
Rp. 2,000,000 - Rp. 4,999,999	3
Rp. 5,000,000 - Rp. 20,000,000	4
Lebih dari Rp. 20,000,000	5

Tabel 5. Transformasi Pendidikan Ayah

Pendidikan Ayah	Transformasi
Tidak Sekolah	0
TK	1
SD	2
SMP	3
SMA	4
D1	5
D2	6
D3	7
D4 / S1	8
S2	9
S3	10

Tabel 6. Transformasi Pendidikan Ibu

Pendidikan Ibu	Transformasi
Tidak Sekolah	0
TK	1
SD	2
SMP	3
SMA	4
D1	5
D2	6
D3	7
D4 / S1	8
S2	9
S3	10

3.3. Pembagian Data

Tahap pembagian data yaitu membagi dataset yang telah diolah pada proses persiapan data dengan rasio 80:20. Pembagian tersebut terbagi menjadi 80% *data training* dan 20% *data testing*.

3.4. Pemodelan Algoritma *K-Nearest Neighbor*

Tahap pemodelan merupakan mencari pola atau informasi menarik dalam data yang dipilih berdasarkan proses dari persiapan data. Teknik yang digunakan untuk modeling prediksi adalah menggunakan algoritma *K-Nearest Neighbor* dimana metode ini memiliki atribut yang di inialisasi sebagai nilai *k*, yaitu jumlah tetangga yang dijadikan acuan pada algoritma dengan prosesnya sebagai berikut:

- a. Menentukan nilai *k*
- b. Hitung Jarak (*d*) (*test data* dan *train data*) menggunakan rumus *Eucliden Distance*

$$Eucliden\ Distance = \sqrt{\sum_{i=1}^P (a_k - b_k)^2} \dots\dots\dots (1)$$

- c. Sorting (*d*) (dari yang terkecil sampai terbesar)
- d. Klasifikasi *test data* mayoritas

3.4. Evaluasi *Confusion Matrix*

Tahap evaluasi yaitu untuk melihat tingkat performa dari pola yang di hasilkan algoritma. Parameter yang digunakan untuk evaluasi algoritma adalah *confusion matrix* yang dapat dilihat pada tabel 7 dengan melihat nilai akurasi.

Tabel 7. *Confusion Matrix*

		Predicted	
		Negatif	Positif
Aktual	Negatif	TP	FP
	Positif	FN	TN

Keterangan:

- TP (*True Positive*) merupakan jumlah data point berlabel yes yang nilainya diidentifikasi benar
- TN (*True Negative*) merupakan jumlah data point berlabel no yang nilainya diidentifikasi salah.
- FP (*False Positive*) merupakan jumlah data point berlabel yes yang nilai sebenarnya diidentifikasi salah.
- FN (*False Negative*) merupakan jumlah data point berlabel no yang nilai sebenarnya teridentifikasi benar.

Mencari nilai akurasi dapat menggunakan rumus sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (2)$$

4. Hasil dan Pembahasan

4.1 Implementasi *K-Nearest Neighbor* Menggunakan Python

```
import pandas as pd
df = pd.read_csv('dataset_olah.csv')
df
```

	Penghasilan_Ayah	Penghasilan_Ibu	Pendidikan_Ayah	Pendidikan_Ibu	Label
0	2	2	8	4	Terlambat
1	3	0	4	3	Terlambat
2	2	0	4	2	Terlambat
3	2	1	4	3	Terlambat
4	2	0	4	0	Terlambat
...
12403	3	0	4	5	Tepat
12404	3	3	8	9	Tepat
12405	5	3	4	4	Tepat
12406	2	0	4	3	Tepat
12407	2	2	8	4	Tepat

12408 rows × 5 columns

Gambar 1. Mengimport Dataset

Pengujian data pada penelitian ini menggunakan pemrograman *Python* untuk pemodelan algoritma *K-Nearest Neighbor*. Langkah pertama yaitu mengimport dataset yang akan digunakan untuk pengujian data dapat dilihat pada gambar 1. Pada Gambar 1, dataset yang akan digunakan dengan lima atribut yaitu penghasilan ayah, penghasilan ibu, pendidikan ayah, pendidikan ibu, dan label target. Jumlah data yang digunakan adalah 12.408. Selanjutnya proses membagi data menjadi x dan y yang dapat dilihat pada Gambar 2.

```
x = df[['Penghasilan_Ayah', 'Penghasilan_Ibu',
        'Pendidikan_Ayah', 'Pendidikan_Ibu']]
y = df['Label']
```

Gambar 2. Membagi Atribut Data Menjadi X dan Y

Gambar 2 variabel x memiliki atribut penghasilan ayah, penghasilan ibu, pendidikan ayah, pendidikan ibu dan variabel y memiliki atribut label. Berikutnya merupakan proses membagi yang dapat dilihat pada gambar 3.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print ("Jumlah train = ", len(y_train))
print ("Jumlah test = ", len(y_test))
```

```
Jumlah train = 9926
Jumlah test = 2482
```

Gambar 3. Pembagian Data Dalam Komposisi 80:20

Gambar 3 merupakan proses membagi data menjadi training dan testing menggunakan fungsi *train_test_split* dengan rasio 80:20. Jumlah data training adalah 9.926 dan testing 2.482. Setelah itu melakukan pemodelan algoritma *K-Nearest Neighbor* dengan rumus *Euclidian Distance* menggunakan fungsi *K-Neighbors Classifier*. Nilai *k* yang akan digunakan dalam penelitian ini adalah 3, 5, 7, 9 dan 11 dimana prosesnya dapat dilihat pada gambar 4.

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier (n_neighbors=3, metric='euclidean')
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)
```

Gambar 4. Pemodelan *K-Nearest Neighbor*

4.2 Evaluasi

Hasil evaluasi dari pemodelan algoritma *K-Nearest Neighbor* yang dibuat sebelumnya dapat dilihat pada tabel 8.

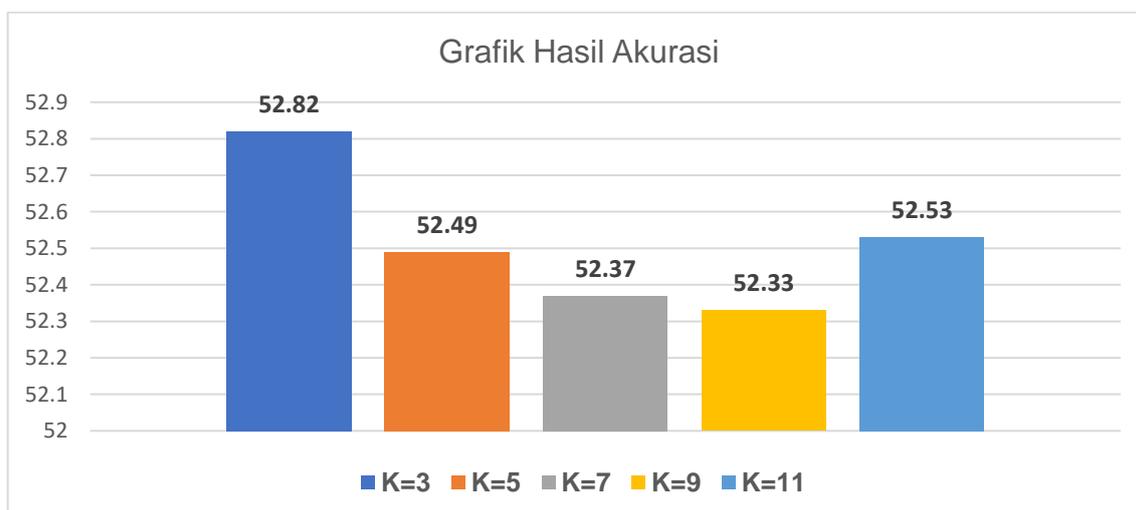
Tabel 8. Hasil Evaluasi <i>Confusion Matrix</i>		
Nilai K	Hasil Evaluasi	Hasil Akurasi
k=3	$\begin{bmatrix} 831 & 413 \\ 758 & 480 \end{bmatrix}$	0.5282030620467365
k=5	$\begin{bmatrix} 558 & 686 \\ 493 & 745 \end{bmatrix}$	0.5249798549556809
k=7	$\begin{bmatrix} 701 & 543 \\ 639 & 599 \end{bmatrix}$	0.523771152296535
k=9	$\begin{bmatrix} 777 & 467 \\ 716 & 522 \end{bmatrix}$	0.5233682514101531
k=11	$\begin{bmatrix} 772 & 472 \\ 706 & 532 \end{bmatrix}$	0.5253827558420628

Perhitungan untuk mendapatkan hasil akurasi berdasarkan hasil evaluasi menggunakan *confusion matrix* adalah sebagai berikut:

$$\begin{aligned} \text{Akurasi } k=3 &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{831+480}{831+480+758+413} = \frac{1311}{2482} = 52,82\% \\ \text{Akurasi } k=5 &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{558+745}{558+745+493+686} = \frac{1303}{2482} = 52,49\% \\ \text{Akurasi } k=7 &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{701+599}{701+599+639+543} = \frac{1300}{2482} = 52,37\% \\ \text{Akurasi } k=9 &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{777+522}{777+522+716+467} = \frac{1299}{2482} = 52,33\% \\ \text{Akurasi } k=11 &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{772+532}{772+532+706+472} = \frac{1304}{2482} = 52,53\% \end{aligned}$$

4.3 Pembahasan

Grafik hasil akurasi pada gambar 5 menunjukkan bahwa percobaan menggunakan beberapa nilai *k* didapatkan hasil akurasi terbaik oleh *k=3* sebesar 52,82%. Berbeda dengan penelitian [2] menggunakan *k=5* dengan hasil akurasi 100% dan penelitian [9] menggunakan *k=13* dengan hasil akurasi 84,21%. Perbedaan penelitian ini dengan penelitian sebelumnya adalah tipe data yang diolah. Data pada penelitian ini betipe kategorikal dan tipe data pada penelitian sebelumnya bertipe numerik sehingga mempengaruhi hasil akurasi yang didapatkan.



Gambar 5. Grafik Hasil Akurasi

Penelitian ini sejalan dengan penelitian [11] yang sama-sama melakukan penelitian tentang prediksi keterlambatan pembayaran memakai atribut NIS, penghasilan orang tua, tanggungan orang tua, pendidikan ayah, pendidikan ibu, umur ibu, dan keterangan menggunakan *K-Nearest Neighbor*. Jumlah data yang diolah penelitian terdahulu tersebut sebanyak 236 record dengan hasil akurasi tertinggi pada *k=3* sebesar 86%. Pada penelitian ini prediksi keterlambatan pembayaran menggunakan atribut penghasilan ayah, penghasilan ibu, pendidikan ayah, pendidikan ibu, dengan jumlah data yang digunakan jauh lebih besar, mencapai 12.408 record. Namun demikian, akurasi maksimum yang diperoleh hanya 52,8%. Temuan ini menjadi menarik untuk dikaji lebih jauh, sebab secara teori semestinya dengan jumlah data set yang besar, kinerja komputasi juga akan lebih baik [16].

5. Simpulan

Dari 5 kali percobaan *K-Nearest Neighbor* yang dilakukan dengan variasi *K=3*, *K=5*, *K=7*, *K=9*, dan *K=11*, akurasi kinerja terbaik terjadi pada nilai *K=3*, yaitu sebesar 52,82%. Dari pemodelan algoritma *K-Nearest Neighbor* yang telah berhasil dibuat dimana *data training* pada penelitian ini dapat dijadikan sebagai data untuk pembuatan sistem prediksi keterlambatan biaya kuliah. Rekomendasi pada penelitian selanjutnya untuk dapat meningkatkan nilai akurasi dengan mencoba menggunakan dataset lain berupa tipe data numerik (non-transformasi).

Daftar Referensi

- [1] J. Suntoro, *Data Mining Algoritma Dan Implementasi dengan Pemograman PHP*. Jakarta: PT Gramedia Jakarta, 2019.
- [2] D. W. Farkhina, D. Utari, A.M Siregar, "Implementasi Algoritme K-Nearest Neighbor (KNN) untuk Prediksi Hasil Produksi," *Sci. Student J. Information, Technol. Sci.*, vol. 1, no. 1, pp. 21–25, 2020.
- [3] E. S. Susanto, K. Kusriani, & H. Al Fatta, "Prediksi Kelulusan Mahasiswa Magister Teknik Informatika Universitas Amikom Yogyakarta Menggunakan Metode K-Nearest Neighbor," *Respati*, vol. 13, no. 2, pp. 67–72, 2018.
- [4] D. Prasetyawan, R. Gatra, "Algoritma K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 1, pp. 56–67, 2022.
- [5] S. R. Rani, S. R. Andani, & D. Suhendro, "Penerapan Algoritma K-Nearest Neighbor untuk Prediksi Kelulusan Siswa pada SMK Anak Bangsa," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, pp. 670-676, 2019.
- [6] D. Şengür, M. Turban, "Prediction Of The Action Identification Levels Of Teachers Based On Organizational Commitment And Job Satisfaction By Using K-Nearest Neighbors Method," *Firat Univ. Turkish J. Sci. Technol.*, vol. 13, no. 2, pp. 61–68, 2018.
- [7] A. Alfani, F. Rozi, & F. Sukmana, "Prediksi Penjualan Produk Unilever Menggunakan Metode K-Nearest Neighbor," *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 6, no. 1, pp. 155–160, 2021.
- [8] J. S. I. Darmayanti, P. Subarkah, & L.R. Anunggilarsa, "Prediksi Potensi Siswa Putus Sekolah Akibat Pandemi Covid-19 Menggunakan Algoritme K-Nearest Neighbor," *J. Sains dan Teknol.*, vol. 10, no. 1, pp. 230–238, 2021.
- [9] S. Sukanto, Y. Adriyani, & R. Aulia, "Prediksi Kelompok UKT Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *JUITA J. Inform.*, vol. 8, no. 1, pp. 121-130, 2020.
- [10] R. K. Dinata, H. Akbar, & N. Hasdyna, "Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 104–111, 2020.
- [11] R. W. Abdullah, Kusriani, & E. T. Luthfi, "Prediksi Keterlambatan Pembayaran Spp Sekolah Dengan Metode K-Nearest Neighbor (Studi Kasus Smk Al-Islam Surakarta)," *J. Inf. Interaktif*, vol. 4, no. 3, pp. 1–18, 2019.
- [12] Y. D. Atma, A. Setyanto, "Perbandingan algoritma c4.5 dan k-nn dalam identifikasi mahasiswa berpotensi drop out," *Metik J. ISSN 2580-1503*, vol. 2, no. 2, pp. 31–37, 2018.
- [13] I Rosiana, "Analisis Kemungkinan Keterlambatan Pembayaran Spp Menggunakan Algoritma Support Vector Machine (Studi Kasus: Smp Perintis ...)," vol. 2, no. 9, pp. 1–11, 2020.
- [14] M. Muqorobin, K. Kusriani, & E. T. Luthfi, "Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah," *J. Ilm. SINUS*, vol. 17, no. 1, pp. 1-14, 2019.
- [15] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, & M. F. Mushtaq, "Imbalance class problems in data mining: A review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1552–1563, 2019.
- [16] W. Widystuti, B. Darmawan. "Pengaruh jumlah data set terhadap akurasi pengenalan dalam deep convolutional network". *Konferensi Nasional Sistem Informasi (KNSI) 2018*, pp. 634-638, 2018.