Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

PERBANDINGAN ALGORITMA *K-NEAREST NEIGHBOR* DAN *NAÏVE BAYES* DALAM MEMPREDIKSI WAKTU KELULUSAN MAHASISWA

Fridy Mandita[1], Muhammad Fajar Andriansyah[2]
[1,2]Jurusan Teknik Informatika, Universitas 17 Agustus 1945 Surabaya
e-mail: [1]fridymandita@untag-sby.ac.id , [2]fjrandrians@gmail.com

Abstrak

Mahasiswa merupakan aspek penting bagi lembaga pendidikan tinggi terutama tentang waktu kelulusan mahasiswa. Oleh sebab itu, penting mengetahui prediksi lama waktu penyelesaian studi mahasiswa. Penelitian ini mengusulkan pembuatan sistem prediksi tingkat kelulusan mahasiswa dengan menerapkan metode machine learning. Prediksi tersebut dapat menjadi langkah preventif agar mahasiswa dapat memperbaiki proses belajar dan menjadi pendorong bagi mahasiswa untuk dapat lulus tepat waktu. Percobaan model klasifikasi KNN dan Naïve Bayes serta uji waktu kelulusan mahasiswa pada beberapa kelas data yaitu 8 kelas klasifikasi (lulus semester 7 sampai 14) dan 4 kelas klasifikasi (Tidak Lulus, Terlambat, Tepat Waktu, dan Cepat). Penelitian ini juga melakukan handling imbalance class yang bertujuan meningkatkan kinerja model klasifikasi menggunkan metode SMOTE. Berdasarkan keseluruhan uji coba, metode KNN lebih unggul dibandingkan metode Naïve Bayes. Penerapan teknik oversampling SMOTE menghasilkan peningkatan yang signifikan dengan selisih nilai evaluasi (presisi, recall, F1 Score, dan akurasi) antara 12% sampai 41% pada metode Naïve Bayes maupun KNN. Hasil prediksi 4 kelas menggunakan metode KNN dengan SMOTE mendapatkan nilai presisi sebesar 79%, nilai recall sebesar 78%, nilai F1 Score sebesar 78%, dan akurasi sebesar 78%. Sedangkan hasil prediksi 8 kelas menggunakan metode KNN dengan SMOTE mendapatkan nilai presisi, recall, F1 Score, dan akurasi sama-sama sebesar 93%.

Kata kunci: Prediksi Waktu Lulus, Klasifikasi, Naïve Bayes, KNN, SMOTE

Abstract

Students are an important aspect for higher education institutions, especially regarding the time of student graduation. Therefore, it is critical to know the prediction of the time length for completing studies. This study proposes creating a prediction system for student graduation rates; hence it could be a preventive measure for students to improve their learning process. This research used machine learning techniques to compare the K-Nearest Neighbor (KNN) and Naïve Bayes algorithms. The experiment aimed to determine the best model, such as the amount of data collection, the number of classification classes, and the handling of imbalanced classes. Based on all experiments, the KNN method achieved higher results than the Naïve Bayes method. Applying the SMOTE oversampling technique significantly increased the difference in evaluation scores (precision, recall, F1 score, and accuracy) between 12% and 41% in the Naïve Bayes and KNN methods. The results of the 4-class prediction model using the KNN method with SMOTE get a

precision value of 79%, a recall value of 78%, an F1 score of 78%, and an accuracy of 78%. In comparison, the prediction results for eight classes using the KNN method with SMOTE get precision, recall, F1 Score, and accuracy values of 93%.

Keywords: Graduation Time Prediction, Classification, Naïve Bayes, KNN, SMOTE

PENDAHULUAN

Pendidikan merupakan salah satu faktor penting di suatu negara dalam meningkatkan Sumber Daya Manusia (SDM). Kualitas negara yang baik dilihat dari tingkat pendidikan yang bagus dan merata di seluruh wilayah. Semakin maju tingkat pendidikan suatu negara, maka semakin maju pula negara tersebut. Menurut Undang-Undang Nomor 20 Tahun 2003, bahwa pendidikan nasional berfungsi mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat dalam rangka mencerdaskan kehidupan bangsa, bertujuan untuk berkembangnya potensi peserta didik agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, dan menjadi warga negara yang demokratis serta bertanggung jawab [1]. Oleh sebab itu banyak kebijakan-kebijakan yang diambil oleh pemerintah dibidang pendidikan guna untuk memberikan pelayanan yang terbaik dari tingkat dasar sampai universitas. Menurut data Badan Pusat Statistik (BPS) tahun 2021, jumlah perguruan tinggi keseluruhan Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS) di Indonesia adalah 3.115 serta Jawa Timur merupakan penyumbang terbesar kedua setelah Jawa Barat dengan jumlah PTN dan PTS sebanyak 338 [2].

Berdasarkan penelitian terdahulu mengenai prediksi kelulusan mahasiswa, keseluruhan masih melakukan klasifikasi hanya dalam 2 kelas yaitu Lulus Tepat Waktu dan Tidak. Sedangkan dalam prediksi waktu kelulusan seharusnya menghasilkan nilai konkrit berupa lama studi yang dibutuhkan yaitu mulai dari semester 7 sampai semester 14. Oleh sebab itu, penelitian ini menggunakan 8 kelas prediksi waktu kelulusan mahasiswa yaitu semester 7, 8, 9, 10, 11, 12, 13, dan 14, dimana kedelapan kelas tersebut masuk pada kemungkinan waktu kelulusan mahasiswa pada salah satu universitas di Surabaya. Namun terdapat beberapa kendala pada klasifikasi 8 kelas tersebut, diantaranya adalah permasalahan imbalance class. Terdapat beberapa kelas dengan sampel data yang sedikit dibandingkan kelas lain, misalnya jumlah mahasiswa yang lulus pada semester 7 lebih sedikit dibandingkan mahasiswa yang lulus di semester 8. Oleh sebab itu, pada penelitian ini juga mengusulkan penanganan data *imbalance* menggunakan teknik *oversampling*. Data dianggap sebagai *imbalance* jika persentase salah satu kelas yang memiliki jumlah data paling sedikit atau *minority class* kurang dari 35% [3]. Oleh sebab itu, pada penelitian ini juga mengusulkan penanganan data *imbalance* menggunakan teknik *oversampling*.

Penggunaan teknik *oversampling* dalam menangani *imbalance class* dapat meningkatkan kinerja sistem klasifikasi [4]. *Synthetic Minority Over-sampling*

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

Technique (SMOTE) merupakan salah satu teknik oversampling yang banyak digunakan dalam menangani imbalance class. Beberapa penelitian telah membuktikan penggunaan teknik SMOTE mampu menangani *imbalance class* dengan membuat *sample data* sintetis baru dan dapat meningkatkan kinerja klasifikasi.

Hasil prediksi lama studi adalah semester akhir masa perkuliahan yang menunjukkan kelulusan mahasiswa. Beberapa penelitian terdahulu telah banyak melakukan prediksi kelulusan mahasiswa, seperti pada penelitian oleh [5] melakukan klasifikasi kelulusan mahasiswa di Fakultas Ekonomi, Universitas Garut menggunakan metode *Decision Tree*. Data berjumlah 999 data dengan 2 kriteria klasifikasi yaitu Lulus Tepat Waktu dan Tidak Tepat Waktu. Klasifikasi menghasilkan nilai akurasi sebesar 79,53% dan *recall* sebesar 19,23%. Selanjutnya penelitian oleh [6] melakukan prediksi kelulusan menggunakan komparasi metode *Decision Tree* dan *Artificial Neural Network*. Klasifikasi kelulusan mahasiswa dibagi menjadi 2 label yaitu mahasiswa dengan lulus tepat waktu dan tidak. Penelitian ini menghasilkan nilai akurasi 74,51% untuk metode *Decision Tree* dan 79,74% untuk *Artificial Neural Network*. Hasil tersebut menunjukkan *Artificial Neural Network* lebih baik dibandingkan metode *Decision Tree*. Selanjutnya penelitian oleh [7] dalam melakukan prediksi waktu kelulusan mahasiswa dengan mengkomparasi 4 metode yaitu algoritma C4.5, Naïve Bayes, KNN, dan SVM. Tingkat keberhasilan klasifikasi diukur dari 2 label kelas yaitu Lulus Tepat Waktu dan Tidak. Hasil terbaik adalah algoritma *Naïve Bayes* dengan nilai akurasi 76,79%. Beberapa penelitian terdahulu rata-rata masih mendapatkan hasil yang kurang maksimal dalam prediksi kelulusan mahasiswa. Oleh sebab itu pemilihan metode yang baik sangat berpengaruh terhadap hasil klasifikasi.

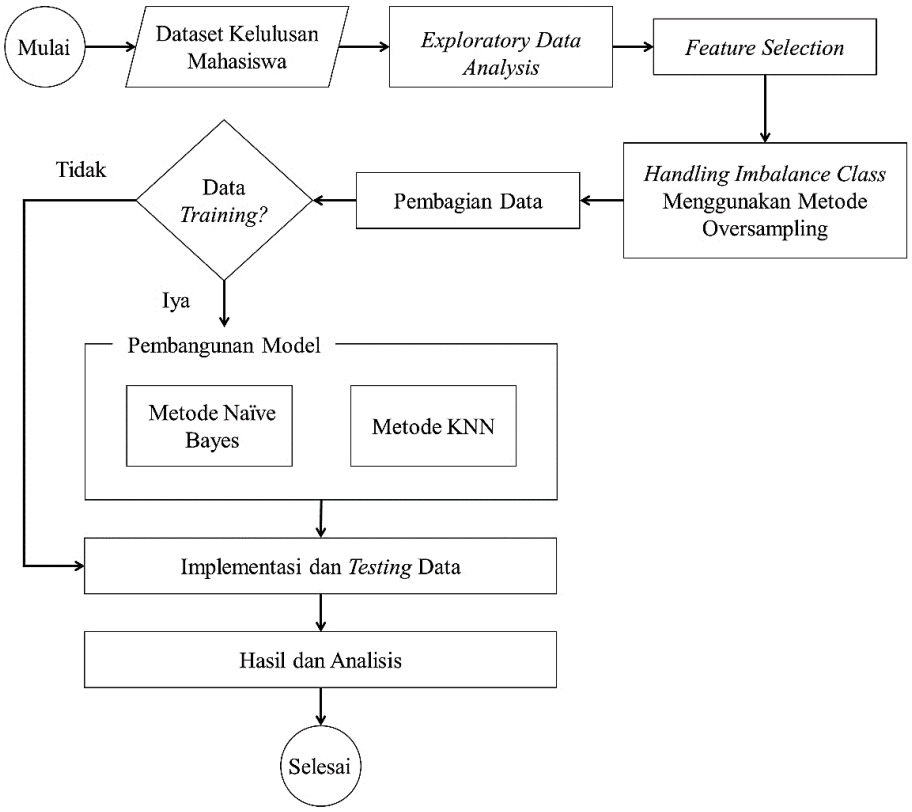
Pada penelitian oleh [8] dalam penelitian tersebut melakukan Klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN) yang digunakan untuk memprediksikan kelulusan tepat waktu mahasiswa. Hasil yang diperoleh menunjukkan pola prediksi tertentu, mahasiswa yang memiliki nilai bagus pada keterampilan di lab, kursus numerik, kursus terkait ilmu komputer. IPK secara konsisten dari tahun pertama hingga tahun keempat memiliki peluang terbaik untuk lulus tepat waktu. Hasil akurasi seluruh data testing adalah 99,83%. Selanjutnya oleh [9] dalam penelitian tersebut melakukan prediksi dua kelas yaitu mahasiswa lulus tepat waktu dan tidak. Algoritma *Naive Bayes* dengan enam tahapan *Cross Industry Standard Process for Data Mining* digunakan dalam penelitian eksperimen ini. Hasilnya model prediksi kelulusan mahasiswa menghasilkan tingkat akurasi sebesar 93,75%.

Selanjutnya penanganan *imbalance class*, penelitian oleh [10] melakukan deteksi penipuan *credit card* menggunakan metode SMOTE. Hasil menunjukkan bahwa penggunaan SMOTE dapat meningkatkan hasil akurasi hingga 7% dibandingkan tanpa SMOTE diberbagai uji coba koefisien k. Selanjutnya penelitian oleh [11] melakukan penanganan *imbalance class* menggunakan metode SMOTE pada klasifikasi *Car Evolution* dengan metode KNN. Terbukti bahwa SMOTE dapat meningkatkan akurasi 9,97% dibandingkan tanpa SMOTE.

Berdasarkan penelitian terdahulu menunjukkan bahwa metode SMOTE memiliki peforma yang baik dalam menangani imbalance class pada sistem klasifikasi.

Berdasarkan hasil studi literatur, penelitian ini melakukan prediksi waktu kelulusan mahasiswa dengan *multiclass* dari label semester 7 sampai semester 14. Penerapan teknik *oversampling* data SMOTE bertujuan untuk menangani masalah *imbalance class* dan meningkatkan kinerja sistem klasifikasi. Metode yang digunakan adalah komparasi dari metode KNN dan *Naïve Bayes* untuk menghasilkan sistem klasifikasi yang optimal dalam memprediksi waktu kelulusan mahasiswa.

METODE PENELITIAN



Gambar 1. *Flowchart* Tahapan Penelitian

Penelitian ini melakukan tahapan dalam memprediksi waktu kelulusan mahasiswa salah satu universitas di Surabaya. Langkah awal adalah pemrosesan dan analisis data, penanganan imbalance dengan algoritma SMOTE, komparasi model klasifikasi dengan metode KNN dan Naïve Bayes, serta langkah akhir adalah evaluasi dan menetukan model paling optimal untuk prediksi waktu kelulusan. Berdasarkan sumber informasi dari universitas, mahasiswa dinyatakan lulus apabila

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

telah menyelesaikan mata kuliah yang berhubungan dengan Tugas Akhir (TA). Serta keputusan universitas tentang batas minimal kuliah adalah 14 semester bagi mahasiswa salah satu universitas di Surabaya. Oleh sebab itu pada penelitian ini melakukan prediksi waktu kelulusan dengan membuat model klasifikasi multiclass, dimana terdapat 8 kelas yaitu kemungkinan waktu kelulusan dari semester 7 sampai semester 14. Penelitian ini juga melakukan uji coba model klasifikasi 4 kelas yaitu tidak lulus, terlambat, tepat waktu, dan cepat. Label tidak lulus tersebut jika mahasiswa tidak menyelesaikan masa perkuliahan. Label terlambat, jika mahasiswa lulus diatas 8 semester yaitu antara 9 sampai 14 semester. Label tepat waktu, jika mahasiswa lulus tepat pada waktu normal perkuliahan yaitu 8 semester. Sedangkan label cepat, jika mahasiswa lulus pada 7 semester.

Mula-mula dilakukan pembacaan dan Exploratory Data Analysis (EDA) data menggunakan program Python. Kemudian apabila terdapat data-data yang tidak dibutuhkan maka akan dilakukan data cleaning terlebih dahulu. Setelah dilakukan data cleaning, menentukan variabel atau fitur yang akan paling tepat untuk digunakan. Pemilihan tersebut memanfaatkan uji korelasi terhadap label klasifikasi. Selanjutnya adalah proses pembuatan data synthetic menggunakan SMOTE. Hal ini dilakukan karena terdapat imbalance class pada data waktu lulus mahasiswa. Selanjutnya data akan dibagi menjadi 2 bagian yaitu data training dan data testing. Pembagian data training dan testing menggunakan percobaan K-Fold dengan k=5, 7, dan 10. Data training digunakan sebagai input pada metode Naive Bayes maupun KNN untuk mendapatkan model klasifikasi yang terbaik.

Dataset

Dataset berasal dari rekap data mahasiswa salah satu universitas di Surabaya tahun angkatan 2011 hingga 2015. Tujuan penelitian ini adalah untuk memprediksi waktu kelulusan mahasiswa, sehingga memperlukan data mahasiswa yang sudah dinyatakan lulus dan pada semester berapa mahasiswa tersebut lulus. Data terdapat 2 *file* yaitu laporan *list* IPS dan data nilai mata kuliah mahasiswa. File laporan *list* IPS data mahasiswa berisi tentang rekap nilai tiap semester yang berupa jumlah pengambilan, nilai IPS dan IPK mahasiswa. Sedangkan *file* nilai mata kuliah berisi rekap nilai tiap mata kuliah yang sudah diambil. Tahap selanjutnya adalah pengolahan data. Sesuai dengan topik permasalahan, maka membutuhkan data terkait informasi mahasiswa yang sangat berhubungan dengan waktu kelulusan, seperti nilai IPS dari semester 1 sampai semester 6, IPK sampai semester 6, dan total SKS yang diambil sampai semester 6. Penelitian ini melakukan prediksi waktu kelulusan mahasiswa secara dini yaitu ditujukan pada mahasiswa yang sudah semester 6.

Metode K-Nearest Neighbor (KNN)

Algoritma KNN menerapkan pembelajaran objek klasifikasi berdasarkan perhitungan jarak paling dekat dengan objek tersebut [12]. Algoritma KNN menerapkan perhitungan probabilitas data uji pada kelas data pelatihan *k*. Hasil prediksi ditentukan dari kelas yang memiliki probabilitas tertinggi.

Pseudocode 1. Metode Algoritma KNN

Input: Data *training* X, dimana $X = (x_1, x_2, .., x_n)$, k, label kelas dataset C, data *testing* y.
Output: hasil label kelas data *testing* c_y

1. *begin*
2. *for* setiap x pada dataset X
3. Hitung jarak $D(x, y)$ antara x dan y.

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y)^2}$$

4. distance += (row1[x] - row2[$x + 1$])**2
5. return sqrt(*distance*)
6. *end for*
7. Memilih subset N dari kumpulan data X, dimana N berisi sampel data tetangga terdekat k
8. Menentukan hasil klasifikasi data *testing* y dengan cara:

$$c_y = argmax \sum_{x \in N} I(c = kelas(x))$$

9. output_values = [row[-1] for row in neighbors]
10. prediction = max(set(output_values), key=output_values.count)
11. *end*

Metode Naïve Bayes

Inti dari *classifier* didasarkan pada Teorema Bayes yang dikemukakan oleh ilmuwan Inggris, Thomas Bayes [13]. Algoritma *Naïve Bayes* dapat dikatakan sebagai algoritma yang dibentuk berdasarkan Teorema Bayes dengan asumsi Naïve atau independensi antar fitur.

Pseudocode 2. Metode Algoritmza Naïve Bayes

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

Input: Data fitur X, dimana $X = (x_1, x_2, .., x_n)$
Output: hasil label kelas data *testing* y

1. *begin*
2. Hitung nilai peluang $P(X)$ pada fitur-fitur $x_1, x_2, .., x_n$
3. Hitung nilai peluang $P(y)$ pada seluruh label kelas data.
4. *for* setiap n pada dataset X
   Hitung nilai peluang terjadinya $P(x_n|y)$
5. *end for*
6. Hitung *likelihood* $P(X|y)$ dengan cara:
   $$P(X|y) = P(x_1|y) * P(x_2|y) * ... * P(x_n|y)$$
7. total_rows = sum([summaries[label][0][2] for label in summaries])
8. probabilities = dict()
9. for class_value, class_summaries in summaries.items():
10. probabilities = summaries[class_value][0][2]/float(total_rows)
11. for i in range(len(class_summaries)):
12. mean, stdev, _ = class_summaries[i]
13. probabilities[class_value] *= calculate_probability(row[i], mean, stdev)
14. Menentukan hasil klasifikasi data *testing* y dengan cara:
    $$y = argmax_y\ p(y) \prod_{i=1}^{n} P(x_i|y)$$
15. best_label, best_prob = None, -1
16. for class_value, probability in probabilities.items():
17. if best_label is None or probability > best_prob:
18. best_prob = probability
19. best_label = class_value
20. return best_label
21. *end*

Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode usulan dari Chawla, dkk Tahun 2002 yang beguna untuk menangani permasalahan *imbalance class* dengan teknik *oversampling* [14]. Ide dasar dari SMOTE adalah menambahkan jumlah data pada kelas minor dengan membuat data baru atau *synthetic* berdasarkan data tetangga minor yang terdekat sehingga jumlah sampel data minor sebanding dengan jumlah sampel pada data mayor. Pembuatan data *synthetic* ini dianggap lebih efektif dibandingkan cara tradisional lain seperti mengurangi kelas yang mayor atau menduplikat data pada kelas minor [15].

Pseudocode 3. Metode Algoritma SMOTE

Input: Minoritas data X, dimana $X = (x_1, x_2, .., x_n)$ dan n adalah jumlah data minoritas, persentase SMOTE N, dan nilai parameter KNN k
Output: hasil data sintetis dari kelas minoritas x_{syn}

1. *begin*
2. *for* setiap $i = 1, 2, \dots, n$
3. Hitung jarak antar tetangga pada minoritas data d dengan cara:
   $$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_i - y_i)^2}$$
4. $\widehat{N} = N/100$
5. *while* $\widehat{N} \neq 0$
6. Menentukan data terdekat y_{knn} dengan cara:
   $$y_{knn} = y_{\min\,(d(x,y_1),d(x,y_2),d(x,y_3),\dots,d(x,y_n))}$$

7. Random parameter γ dengan acak antara 0 sampai 1 ($\gamma \in [0,1]$)
8. Hitung data sintetis x_{syn} dengan cara:

$$x_{syn} = x_i + (y_{knn} - x_i)\gamma$$

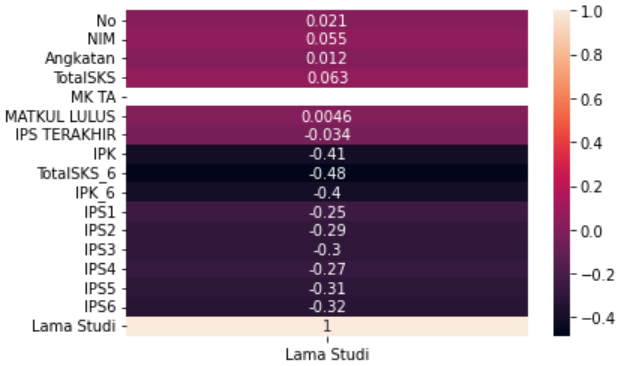
9. $\widehat{N} = \widehat{N} - 1$
10. *end while*
11. *end for*
12. *end*

HASIL DAN PEMBAHASAN

Penelitian ini melakukan prediksi lama waktu lulus mahasiswa dengan 2 jenis model yaitu 8 kelas klasifikasi berdasarkan banyak semester dan 4 kelas berdasarkan kelompok kelulusan. Prediksi tesebut dapat menjadi langkah preventif agar mahasiswa dapat memperbaiki proses belajar dan menjadi pendorong untuk lulus tepat waktu.

Exploratory Data Analysis (EDA)

Tahap awal adalah pembacaan dan pengolahan data. Perngolahan tersebut berupa pengambilan beberapa variabel atau fitur dari keseluruhan data yang menghasilkan 18 variabel seperti NIM, Angkatan, Nama, Total SKS yang telah ditempuh, informasi pengambilan Mata Kuliah Tugas Akhir (MK TA), jumlah mata kuliah lulus, nilai IPS pada semester paling akhir, nilai IPK, total SKS yang diambil sampai semester 6, nilai IPK sampai semester 6, nilai IPS semester 1 sampai IPS semester 6. Total keseluruhan data sebanyak 2.116 mahasiswa dari angkatan 2011 sampai 2015.

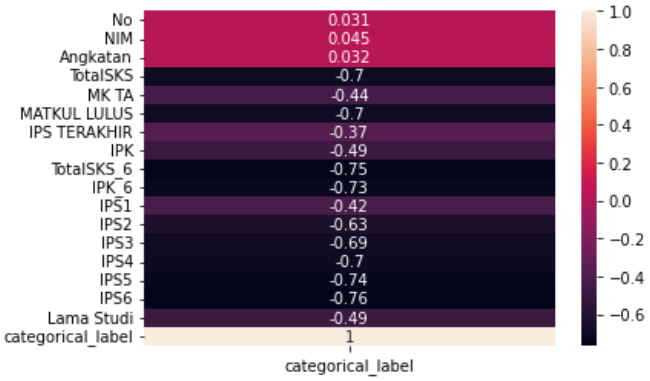


Gambar 2. Uji Korelasi Seluruh Variabel Terhadap Lama Studi Mahasiswa 8 Kelas

Hasil uji korelasi Gambar 2, menunjukkan bahwa terdapat 3 variabel yang memiliki hubungan yang kuat dengan lama studi mahasiswa yaitu nilai IPK, Total SKS sampai semester 6, dan nilai IPK sampai semester 6 dengan masing-masing nilai korelasi adalah -0,41, -0,48, dan -0,4. Sedangkan nilai IPS semester 1 sampai semester 6 memiliki rentang nilai korelasi antara -0,25 sampai -0,32. Berdasarkan

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

uji korelasi ini variabel yang dipilih sebanyak 8 yaitu Total SKS sampai semester 6, IPK sampai semester 6, dan nilai IPS dari semester 1 sampai semester 6. Sedangkan nilai IPK tidak digunakan karena IPK dapat diperoleh sampai semester akhir mahasiswa sehingga tidak relevan dengan penelitian ini yang digunakan sebagai prediksi dini lama studi mahasiswa.

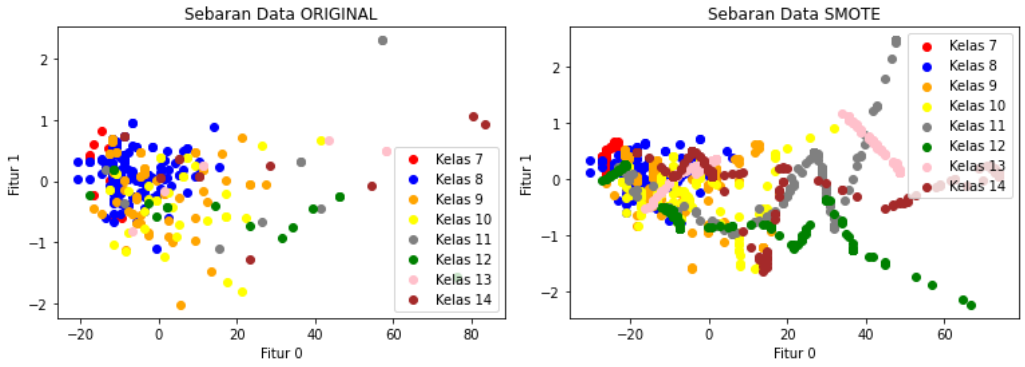


Gambar 3. Uji Korelasi Seluruh Variabel Terhadap Lama Studi Mahasiswa 4 Kelas

Hasil uji korelasi data dengan jumlah label 4 kelas dapat dilihat pada Gambar 3. Sebelum dilakukan uji korelasi, label data diubah dalam bentuk numerik agar dapat dihitung nilai korelasi antar variabel. Hasil uji korelasi, grafik tersebut menunjukkan bahwa 7 variabel atau parameter yang memiliki nilai korelasi yang terbaik yaitu Total SKS, Jumlah Mata Kuliah yang Lulus, Total SKS sampai semester 6, nilai IPK sampai semester 6, IPS semester 4, IPS semester 5, dan IPS semester 6. Hasil uji korelasi 4 kelas, hampir seluruh variabel mengalami peningkatan nilai korelasi dibandingkan pada lama studi 8 kelas. Agar dapat dibandingkan dengan data sebelumnya yaitu 8 kelas maka pengambilan variabel sama dengan data sebelumnya. Data yang digunakan sebagai *input* model adalah Total SKS sampai semester 6, IPK sampai semester 6, dan nilai IPS dari semester 1 sampai semester 6.

Hasil Penerapan SMOTE

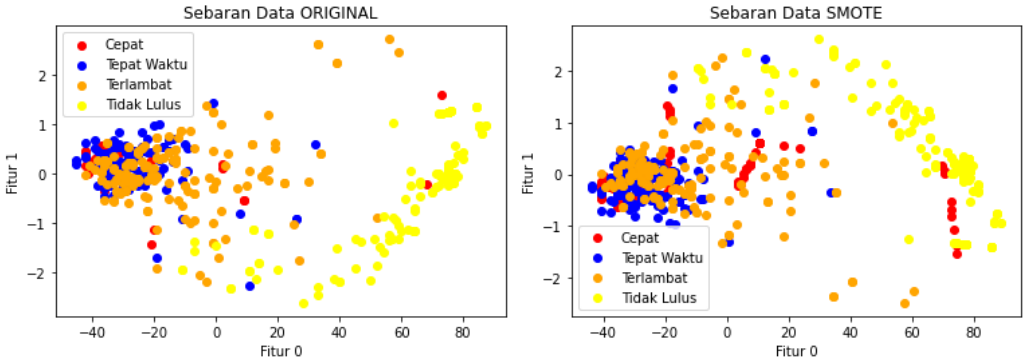
Berdasarkan penerapan teknik *oversampling* SMOTE menghasilkan beberapa data *synthetic* pada setiap uji coba data. Uji coba seperti pada data 8 kelas klasifikasi dengan jumlah 100, 257, dan semua data, serta data 4 kelas klasifikasi dengan jumlah data 100, 257, dan semua data. Agar dapat melihat dengan jelas sebaran data baru yang dibuat, maka dilakukan *plotting data* yang menerapkan metode *Principal Component Analysis* (PCA). Metode tersebut berguna untuk mereduksi 8 fitur data menjadi 2 fitur sehingga dapat direpresentasikan dalam bentuk plot 2 dimensi *(x,y)*.



(a) (b)

Gambar 4. Plot Perbandingan Sebaran (a) Data Original dan (b) Data *Synthetic* Pada Uji Coba 8 Kelas Klasifikasi dengan Pengambilan 100 Data

Penerapan teknik *oversampling* SMOTE pada pengambilan 100 data menghasilkan data sebanyak 1.120. Terdapat 831 data *synthetic* yang dihasilkan menggunakan teknik *oversampling* SMOTE, sedangkan data original atau asli hanya 289 data. Penerapan teknik *oversampling* SMOTE menghasilkan 74,19% data baru yang digunakan sebagai pelatihan model. Hal ini tidak bagus dalam pembuatan data, karena data pelatihan lebih banyak berasal dari data *synthetic* daripada data asli. Data *original* terlihat sebaran paling dominan adalah warna biru yaitu kelas 8 semester, sedangkan data kelas 10 semester, warna kuning, terlihat lebih menyebar dibandingkan data kelas yang lain. Hasil sebaran data *synthetic* SMOTE menunjukkan bahwa data baru dibuat diantara beberapa data asli yang sudah ada dengan jarak terdekat antar data. Seperti contoh pada kelas 11 semester, pada Gambar 4(a) terdapat data dengan label kelas 11 semester berada diatas, oleh sebab itu arah sebaran data synthetic yang dibuat sejalan dengan data-data yang sudah ada sehingga hasil pada Gambar 4(b) terlihat seperti garis lurus yang naik ketas. Hal ini menunjukkan bahwa konsep SMOTE dalam membuat data baru berhasil yaitu data baru dibuat berdasarkan data terdekat kelas minoritas.



(a) (b)

Gambar 5. Plot Perbandingan Sebaran (a) Data Original dan (b) Data *Synthetic* Pada Uji Coba 4 Kelas Klasifikasi dengan Pengambilan 100 Data

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

Penerapan teknik *oversampling* SMOTE pada pengambilan 100 data menghasilkan data sebanyak 680. Terdapat 180 data *synthetic* yang dihasilkan menggunakan teknik *oversampling* SMOTE, sedangkan data *original* atau asli sebanyak 500 data. Penerapan teknik *oversampling* SMOTE menghasilkan 26,47% data baru yang digunakan sebagai pelatihan model. Jumlah mahasiswa yang paling banyak terlihat adalah mahasiswa dengan tingkat kelulusan kelas Terlambat dan kelas Tidak Lulus. Sedangkan tingkat kelulusan kelas Cepat hampir tidak terlihat serta penyebarannya yang luas. Hasil sebaran data *synthetic* SMOTE Gambar 5(b) menunjukkan bahwa data baru dibuat diantara beberapa data asli yang sudah ada dengan jarak terdekat antar data. Seperti contoh pada kelas Cepat, pada Gambar 5(a) atas sebelah kanan, terdapat ruang kosong antar data asli sehingga arah sebaran data *synthetic* yang dibuat sejalan dengan data-data yang sudah ada sehingga hasil pada Gambar 5(b) terlihat seperti garis lurus yang naik ketas kemudian balik lagi kebawah. Hal ini menunjukkan bahwa konsep SMOTE dalam membuat data baru berhasil yaitu data baru dibuat berdasarkan data terdekat kelas minoritas. Berdasarkan grafik tersebut juga telah membuktikan bahwa sebaran data mahasiswa dengan label kelas Tepat Waktu dan Terlambat memiliki titik sebaran yang hampir sama atau tumpang tindih. Oleh sebab itu, hasil data *synthetic* antara data dengan label kelas Terlambat hampir sama dengan label Tepat Waktu.

Prediksi Waktu Kelulusan Mahasiswa

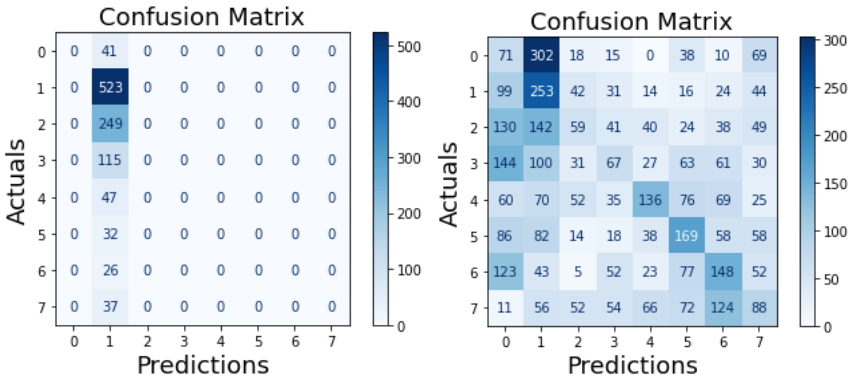
Pembuatan model prediksi waktu kelulusan mahasiswa menggunakan metode KNN dan *Naïve Bayes* dengan beberapa uji coba. Pada penelitian ini, melakukan uji coba jumlah pengambilan data untuk mendapatkan model yang terbaik. Oleh sebab itu, penelitian ini melakukan 4 jenis tahap uji coba yaitu uji coba jumlah kelas, uji coba jumlah data yang diambil, uji coba metode klasifikasi, dan uji coba *handling imbalance data*. Hasil keseluruhan uji coba dapat dilihat pada Table 1. Hasil keseluruhan uji coba, dapat diketahui bahwa model prediksi waktu kelulusan terbaik adalah menggunakan metode KNN + teknik *oversampling* SMOTE dengan pengambilan 257 data. Susunan model tersebut terbukti paling baik pada kedua uji coba jumlah kelas (8 kelas maupun 4 kelas). Pengambilan model terbaik dipilih berdasarkan hasil beberapa nilai evaluasi seperti, nilai akurasi, presisi, *recall*, dan *F1 Score*. Nilai akurasi berguna untuk mengukur model yang terprediksi benar (positif maupun negatif) pada keseluruhan data. Nilai presisi berguna untuk mengukur model yang terprediksi benar positif dibandingkan keseluruhan data prediksi positif. Nilai *recall* berguna untuk mengukur model yang terprediksi benar positif dibandingkan keseluruhan data aktual positif. Sedangkan nilai *F1 Score* berguna untuk mengukur model yang terprediksi salah sehingga nilai FP dan FN diperhitungkan dalam nilai ini. Oleh sebab itu nilai *F1 Score* juga baik dalam mengevaluasi model dengan kasus *imbalance data*.

Tabel 1. Hasil Keseluruhan Uji Coba Prediksi Waktu Kelulusan Mahasiswa

Jumlah Kelas	Jumlah Data	Metode Klasifikasi	Handling Imbalance	Presisi	Recall	F1 Score	Akurasi
4 Kelas	100 Data	Naïve Bayes	Original	0,53	0,54	0,52	0,65
			SMOTE	0,47	0,55	0,48	0,55
		KNN	Original	0,60	0,60	0,59	0,64
			SMOTE	0,76	0,76	0,74	0,76
	257 Data	Naïve Bayes	Original	0,55	0,56	0,54	0,70
			SMOTE	0,59	0,56	0,50	0,56
		KNN	Original	0,60	0,59	0,58	0,66
			SMOTE	0,79	0,78	0,78	0,78
	Semua Data	Naïve Bayes	Original	0,35	0,47	0,38	0,61
			SMOTE	0,51	0,51	0,48	0,51
		KNN	Original	0,53	0,51	0,50	0,59
			SMOTE	0,72	0,71	0,69	0,71
8 Kelas	100 Data	Naïve Bayes	Original	0,06	0,13	0,08	0,48
			SMOTE	0,50	0,45	0,46	0,45
		KNN	Original	0,29	0,28	0,28	0,54
			SMOTE	0,87	0,87	0,87	0,87
	257 Data	Naïve Bayes	Original	0,07	0,13	0,09	0,54
			SMOTE	0,31	0,30	0,29	0,30
		KNN	Original	0,30	0,27	0,28	0,56
			SMOTE	0,93	0,93	0,93	0,93
	Semua Data	Naïve Bayes	Original	0,06	0,13	0,08	0,49
			SMOTE	0,24	0,24	0,24	0,24
		KNN	Original	0,23	0,20	0,20	0,45
			SMOTE	0,88	0,88	0,88	0,88

Penting untuk tidak hanya melihat pada suatu nilai evaluasi saja. Misalkan pada percobaan klassifikasi 8 kelas, 257 data, dan metode *Naïve Bayes*, terlihat hasil model tanpa SMOTE memiliki nilai akurasi 54%, dimana yang lebih tinggi dibandingkan model dengan SMOTE yaitu 30%. Namun pada nilai evaluasi yang lain, seperti presisi, *recall*, dan *F1 Score*, memiliki nilai yang buruk dengan masing-masing 7%, 13%, dan 9%. Nilai-nilai tersebut menunjukkan bahwa model sangat baik dalam memprediksi salah satu kelas yang mayor (jumlah paling banyak). Sedangkan pada kelas lain dapat dikatakan gagal dalam memprediksi kelas sehingga nilai presisi, *recall*, dan *F1 Score* sangat buruk. Penjelasan tersebut dijelaskan pada hasil *confusion matrix* Gambar 6.

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517



(a) (b)

Gambar 6. Hasil *Confusion Matrix* dari Model Klasifikasi 8 Kelas dengan Pengambilan 257 Data Menggunakan Metode *Naive Bayes* dan (a) Tanpa SMOTE (b) dengan SMOTE

Berdasarkan Gambar 6(a) diatas terlihat bahwa model dengan tanpa SMOTE hanya dapat memprediksi kelas ke-1 atau dalam hal ini yang dimaksud adalah 8 semester. Kelas mayoritas selalu mendapatkan hasil yang paling baik pada kasus *imbalance class* dan terbukti dengan tingginya nilai akurasi, sedangkan nilai presisi, *recall*, dan *F1 Score* sangat buruk. Berbeda halnya hasil model klasifikasi dengan penerapan SMOTE yaitu pada Gambar 6(b), seluruh kelas mendapatkan data synthetic yang dapat membantu menyeimbangkan jumlah data antar kelas. Terlihat bahwa model tidak hanya dapat memprediksi kelas ke-1 (8 semester), namun juga di kelas ke-4 (11 semester), ke-5 (12 semester), dan ke-6 (13 semester). Sehingga meskipun nilai akurasi lebih rendah 30% namun nilai presisi, *recall*, dan *F1 Score* dapat dikatakan lebih baik dibandingkan model tanpa SMOTE yaitu masing-masing 31%, 30%, dan 29%.

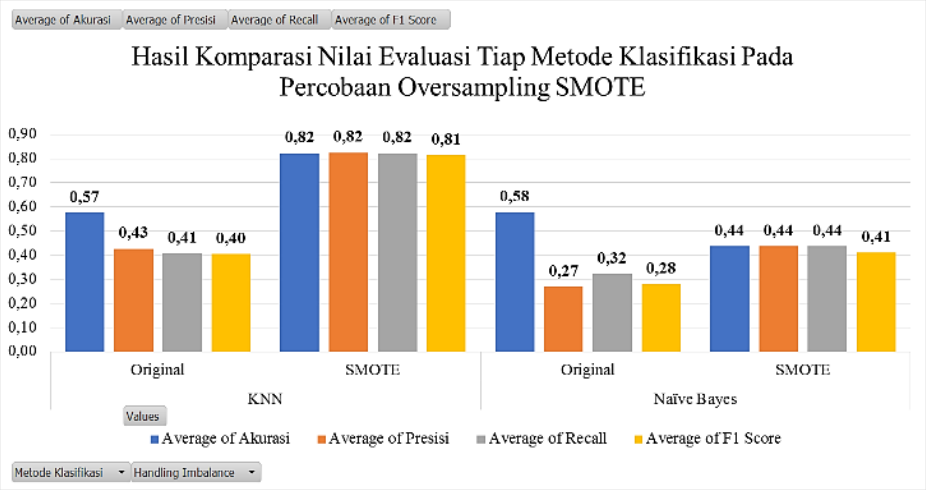
Penelitian ini menghasilkan model yang baik dan dapat dijadikan model prediksi yang tepat dalam permasalahan waktu kelulusan mahasiswa. Pemilihan model klasifikasi paling baik adalah menggunakan metode KNN dengan selisih nilai evaluasi paling tinggi mencapai 6% pada percobaan 8 kelas, 100 data, dan tanpa SMOTE dibandingkan metode *Naïve Bayes*. Penerapan *handling imbalance class* menggunakan SMOTE menjadikan model semakin baik. Hal ini disebabkan karena model semakin banyak mempelajari data dengan berbagai variasi dengan jumlah kelas yang sama, sehingga model akan dapat mengenali dengan baik prediksi disetiap kelas dan tidak hanya cenderung baik pada kelas prediksi tertentu. Penerapan model paling signifikan antara SMOTE dan tanpa SMOTE dengan selisih akurasi mencapai 43% pada 8 kelas, semua data, dan metode KNN. Hasil model prediksi 4 kelas waktu kelulusan terbaik yaitu mendapatkan nilai presisi sebesar 79%, nilai *recall* sebesar 78%, nilai *F1 Score* sebesar 78%, dan akurasi sebesar 78%. Hasil model prediksi 8 kelas waktu kelulusan terbaik yaitu mendapatkan nilai presisi, *recall*, *F1 Score*, dan akurasi sama-sama sebesar 93%.

Selain adanya permasalahan *imbalance class*, perbedaan data antara mahasiswa yang lulus dengan 8 semester ataupun diatas 8 semester (9, 10, 11, dan seterusnya) juga tidak jauh berbeda. Kebanyakan mahasiswa memiliki nilai IPK dan IPS tiap semesternya baik namun tidak sedikit yang memilih lulus pada

semester-semester akhir atau diatas semester normal yaitu 8 semester. Setelah dilakukan analisis lebih lanjut, terdapat beberapa alasan diantara lain seperti kelompok mahasiswa-mahasiswa Pendidikan Jarak Jauh (PJJ), kesulitan menyelesaikan Mata Kuliah Tugas Akhir, mahasiswa dengan status sedang bekerja, dan lain-lain. Dimana parameter-parameter tersebut belum ada pada pengumpulan data penelitian ini sehingga tidak bisa *include* pada pembelajaran model klasifikasi.

Hasil Komparasi Penerapan SMOTE

Selanjutnya melakukan analisis komparasi antara penerapan data *oversampling* SMOTE dan data asli atau *original*. Analisis ini memanfaatkan *pivot table* pada Excel yang menyatukan seluruh percobaan pada Tabel 1 dengan pengambilan beberapa analisis uji coba. Keseluruhan nilai merupakan hasil rata-rata dari beberapa uji coba.



Gambar 7. Hasil Komparasi Nilai Evaluasi Metode KNN dan *Naïve Bayes* Pada Uji Coba Teknik *Oversampling* SMOTE

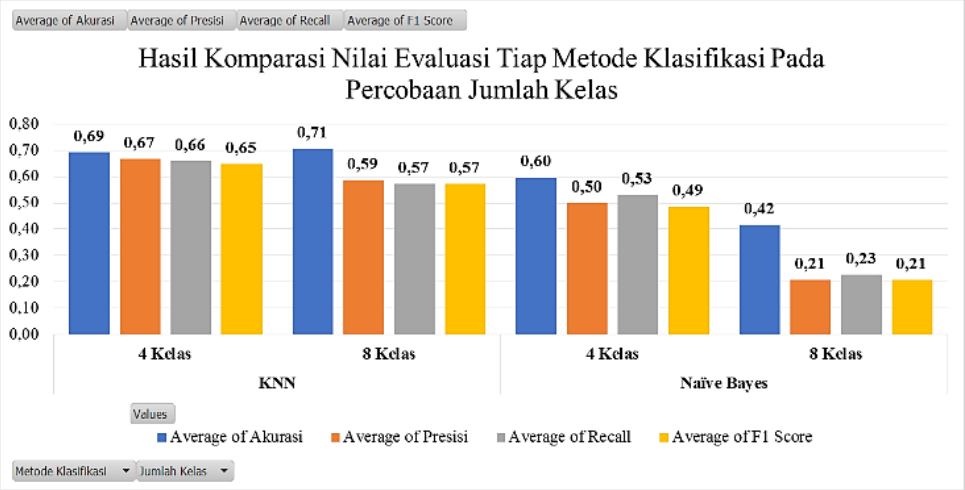
Berdasarkan Gambar 7, menunjukkan bahwa metode KNN dengan penerapan teknik *oversampling* SMOTE adalah model yang terbaik. Rata-rata nilai presisi sebesar 82%, nilai *recall* sebesar 82%, nilai *F1 Score* sebesar 81%, dan akurasi sebesar 82%. Teknik *oversampling* SMOTE dapat meningkatkan kinerja model klasifikasi. Meskipun dapat terlihat bahwa nilai rata-rata akurasi pada model *Naïve Bayes* tanpa penerapan *oversampling* SMOTE memiliki nilai lebih baik dibandingkan model *Naïve Bayes* dengan SMOTE, namun model tersebut tidak dapat dikatakan baik karena rata-rata nilai yang lain masih rendah. Sehingga penerapan *oversampling* SMOTE masih dikatakan lebih baik dibandingkan tanpa penerapan *oversampling* SMOTE. Terbukti pada metode *Naïve Bayes* maupun KNN, penerapan teknik *oversampling* SMOTE dapat meningkatkan nilai evaluasi (presisi, *recall*, *F1 Score*, dan akurasi) antara 12% sampai 41%. Berdasarkan hasil tersebut dapat dikatakan bahwa permasalahan *imabalance class* sangat mempengaruhi kinerja model klasifikasi. Masalah *imbalance class* menyebabkan

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

model yang dihasilkan akan lebih cenderung terhadap mayor *class* sedangkan kelas-kelas yang minoritas juga tidak dapat disepelekan begitu saja. Oleh sebab itu perlu dilakukan penanganan dengan salah satunya adalah teknik *oversampling* SMOTE. Penggunaan teknik *oversampling* menghasilkan data-data baru pada kelas yang minoritas sehingga pembelajaran model dapat seimbang pada setiap kelas.

Hasil Komparasi Metode KNN dan Naïve Bayes

Berdasarkan Gambar 8, menunjukkan bahwa metode KNN lebih unggul dibandingkan metode *Naïve Bayes* dalam permasalahan prediksi waktu lulus mahasiswa. Terbukti hasil antara 4 kelas dan 8 kelas terlihat tidak terlalu signifikan serta model klasifikasi KNN dapat dikatakan lebih stabil dan lebih baik dibandingkan *Naïve Bayes*. Pada metode KNN, terlihat nilai akurasi 8 kelas lebih baik 2% dibandingkan klasifikasi 4 kelas. Namun secara keseluruhan nilai evaluasi (presisi, *recall*, *F1 Score*, dan akurasi), klasifikasi 4 kelas adalah yang baik. Pada metode *Naïve Bayes*, terlihat signifikan perbedaan nilai akurasi, presisi, *recall*, dan *F1 Score* mencapai 30% akurasi, presisi, *recall*, dan *F1 Score* mencapai 30%.



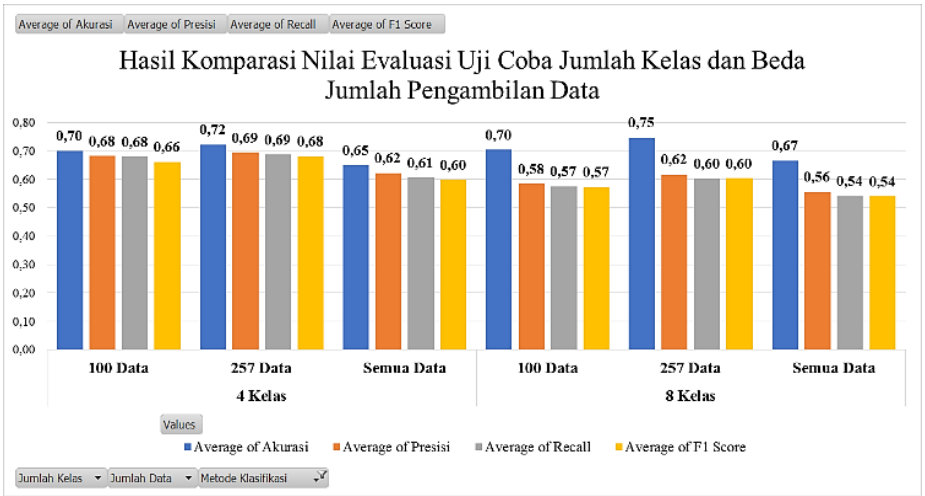
Gambar 8. Hasil Komparasi Nilai Evaluasi Metode KNN dan *Naïve Bayes* Pada Uji Coba Beda Kelas Klasifikasi

Berdasarkan hasil perbandingan antara metode KNN dan *Naïve Bayes*, menunjukkan bahwa data memiliki kedekatan yang baik antar beda kelas sehingga cocok menggunakan metode KNN dibandingkan *Naïve Bayes*. Berbeda dengan metode KNN yang berorientasi pada kedekatan antar data, metode *Naïve Bayes* berorientasi pada probabilitas tiap kejadian sehingga jumlah data tiap kelas sangat berpengaruh. Sehingga pada kasus prediksi lama waktu kelulusan tidak cocok menggunakan metode *Naïve Bayes* karena adanya permasalahan *imbalance class* yang cukup signifikan.

Hasil Komparasi Beda Kelas dan Pengambilan Data

Hasil uji coba pada Gambar 9 menunjukkan bahwa berdasarkan nilai akurasi, percobaan dengan jumlah 8 kelas lebih unggul dibandingkan 4 kelas

klasifikasi. Namun nilai evaluasi yang lain seperti presisi, *recall*, *F1 Score* memiliki rata-rata nilai dibawah 60%. Oleh sebab itu, pengujian dengan 4 kelas klasifikasi dapat dinyatakan lebih baik dibandingkan 8 kelas. Dapat disimpulkan bahwa penentuan jumlah kelas sangat berpengaruh terhadap hasil klasifikasi. Semakin banyak jumlah kelas maka semakin sulit model dalam mengenali data tiap kelas klasifikasi.



Gambar 9. Hasil Komparasi Nilai Akurasi Pada Tiap Uji Coba Jumlah Kelas dan Jumlah Pengambilan Data

Pengambilan 100 data atau 257 data merupakan pengambilan data sejumlah tersebut pada setiap data Angkatan Mahasiswa. Sedangkan keseluruhan data merupakan semua data disemua Angkatan dengan jumlah total 2.116 data. Pada 2 percobaan beda kelas (4 kelas dan 8 kelas), pengambilan 100 dan 257 data lebih baik dibandingkan memakai keseluruhan data. Hal ini terjadi karena memungkinkan sekali jika semua data digunakan, maka data-data *noise* juga semakin banyak sehingga tidak menjamin model akan lebih baik.

Hasil Komparasi Beda Nilai K-Fold

Penelitian ini juga melakukan perbandingan nilai *K-Fold* pada pembagian data *training* dan *testing*. Nilai k perpengaruh terhadap berapa banyak jumlah formasi pembagian data, sehingga setiap data memiliki kesempatan untuk menjadi data *training* maupun data *testing*. Penelitian ini menguji nilai k dengan 5, 7, dan 10. Hasil perbandingan dapat ditunjukkan pada Tabel 4.15. Berdasarkan hasil uji coba Tabel 4.15 diatas menunjukkan bahwa k=10 menghasilkan nilai yang paling baik dibandingkan nilai k=7 atau k=5.

Tabel 2. Perbandingan Nilai K-Fold Pada Uji Coba 8 Kelas Klasifikasi

K-Fold	Metode Klasifikasi	*Handling Imbalance*	Presisi	Recall	F1 Score	Akurasi
k=7	Naïve Bayes	Original	0,41	0,13	0,08	0,49
		SMOTE	0,24	0,24	0,22	0,24

Jurnal Dinamika Informatika
Volume 9, No 2, September 2020
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

	KNN	Original	0,24	0,20	0,20	0,45
		SMOTE	0,85	0,86	0,85	0,86
k=5	Naïve Bayes	Original	0,06	0,13	0,08	0,49
		SMOTE	0,24	0,24	0,23	0,23
	KNN	Original	0,21	0,19	0,19	0,45
		SMOTE	0,85	0,86	0,85	0,86
k=10	Naïve Bayes	Original	0,06	0,13	0,08	0,49
		SMOTE	0,24	0,24	0,24	0,24
	KNN	Original	0,23	0,20	0,20	0,45
		SMOTE	0,88	0,88	0,88	0,88

KESIMPULAN

Penelitian ini melakukan prediksi waktu kelulusan mahasiswa. Terdapat dua hasil prediksi waktu kelulusan yaitu berdasarkan tingkat waktu (Cepat, Tepat, Terlambat, dan Tidak Lulus) dan berdasarkan jumlah semester yang ditempuh (semester 7, semester 8, sampai semester 14). Data yang digunakan adalah informasi tentang rekap nilai IPS tiap semester mahasiswa salah satu uniersitas di Surabaya. Uji coba dilakukan untuk mendapatkan model terbaik seperti uji coba metode, penerapan teknik *oversampling*, dan pengambilan jumlah data. Model prediksi waktu kelulusan terbaik adalah menggunakan metode KNN dengan penambahan teknik *oversampling* SMOTE dan pengambilan data sebanyak 257. Hasil prediksi 8 kelas waktu kelulusan terbaik yaitu mendapatkan nilai presisi, *recall*, *F1 Score*, dan akurasi yang sama yaitu sebesar 93%. Metode KNN lebih unggul dibandingkan metode *Naive Bayes* pada kasus prediksi waktu kelulusan. Sehingga data nilai mahasiswa memiliki nilai kedekatan yang lebih baik pada tiap waktu kelulusan dibandingkan dengan nilai probabilitas kejadiannya. Penelitian ini juga membuktikan bahwa pada metode *Naïve Bayes* maupun KNN, penerapan teknik *oversampling* SMOTE dapat meningkatkan nilai evaluasi (presisi, *recall*, *F1 Score*, dan akurasi) antara 12% sampai 41%.

SARAN

Berdasarkan keseluruhan uji coba yang telah dilakukan, penulis menyadari adanya banyak kekurangan. Oleh sebab itu, penulis menyarankan perbaikan hasil prediksi yang lebih baik lagi dalam memprediksi waktu kelulusan mahasiswa. Pengambilan data menjadi poin penting pada penelitian ini, sehingga harus dipastikan bahwa data merupakan hasil terupdate rekap dan dilakukan pengecekan outlier terlebih dahulu. Penambahan parameter-parameter lain seperti usia, gender, status pekerjaan, dan kelompok mahasiswa (Reguler atau PJJ) dapat membantu meningkatkan model prediksi.

DAFTAR PUSTAKA

[1] T. Noor, “rumusan tujuan pendidikan nasional pasal 3 undang-undang sistem pendidikan nasional No 20 Tahun 2003,” *Wahana Karya Ilm. Pendidik.*, vol. 3, no. 01, 2018.

[2] Badan Pusat Statistik, "Jumlah Perguruan Tinggi, Tenaga Pendidik dan Mahasiswa(Negeri dan Swasta) di Bawah Kementerian Riset, Teknologi dan Pendidikan Tinggi/Kementerian Pendidikan dan Kebudayaan Menurut Provinsi, 2021," *BPS*, 2021.

[3] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 321–330, 2014.

[4] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019.

[5] F. F. Roji and D. Ramdani, "The Study of Decision Tree Algorithm For Classification of Student Graduation (Case Study: Faculty of Economics, University of Garut)," *RISTEC Res. Inf. Syst. Technol.*, vol. 2, no. 1, pp. 62–72, 2021.

[6] E. P. Rohmawan, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Desicion Tree Dan Artificial Neural Network," *J. Ilm. Matrik*, vol. 20, no. 1, pp. 21–30, 2018.

[7] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4, 5, Naive Bayes, Knn Dan Svm," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019.

[8] A. P. Salim, K. A. Laksitowening, and I. Asror, "Time Series Prediction on College Graduation Using KNN Algorithm," in *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*, 2020. doi: 10.1109/ICoICT49345.2020.9166238.

[9] C. Mulyadi and L. Sugiarto, "Penggunaan Algoritma Naïve Bayes untuk Prediksi Ketepatan Waktu Lulus Mahasiswa Diploma 3 STMIK Cipta Darma Surakarta," *Teknomatika*, vol. 11, no. 01, pp. 21–30, 2021.

[10] R. Siringoringo, "Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan k-nearest neighbor," *J. Inf. Syst. Dev.*, vol. 3, no. 1, 2018.

[11] F. D. Astuti and F. N. Lenti, "Implementasi SMOTE untuk mengatasi Imbalance Class pada Klasifikasi Car Evolution menggunakan K-NN," *JUPITER (Jurnal Penelit. Ilmu dan Teknol. Komputer)*, vol. 13, no. 1, pp. 89–98, 2021.

[12] F. Ramadani and F. Mandita, "Implementasi Metode K-Nearest Neighbor Pada Sistem Informasi Jasa Layanan Sedot Wc Berbasis Web Untuk Meningkatkan Pemasaran," in *Senakama: Prosiding Seminar Nasional Karya Ilmiah Mahasiswa*, 2022, pp. 795–807.

[13] F.-J. Yang, "An implementation of naive bayes classifier," in *2018 International conference on computational science and computational intelligence (CSCI)*, IEEE, 2018, pp. 301–306.

[14] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[15] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018.