Klasifikasi Algoritma Naive Bayes Terhadap Data *Imbalance*

Mochammad Ariel Fadhilah F[1]*

[1]Informatika, Institut Teknologi Nasional
Email : ariel.deni47@mhs.itenas.ac.id
Received 24 01 2024 **|** *Revised 31* 01 2024 **|** *Accepted 31* 01 2024

ABSTRAK

Data Imbalance adalah kondisi jumlah data suatu kelas jauh lebih banyak dibandingkan dengan kelas lainnya, data imbalanced dapat menyebabkan prediksi pada kelas minoritas menjadi buruk dan menurunkan kinerja terhadap model yang digunakan. Klasifikasi dilakukan menggunakan algoritma Naive Bayes untuk mendapatkan prediksi burnout pada mahasiswa, data yang digunakan berjumlah 104 data yang memiliki tiga variabel independen untuk melakukan prediksi. Data imbalance akan melalui proses balancing data menggunakan Teknik Synthetic Minority Oversampling Technique (SMOTE), hasil klasifikasi Naive Bayes mengalami penurunan kinerja model accuracy sebesar 9,52%, precission 7,8%, recall 9,52% dan f1-score 8,55%. Hasil evaluasi kinerja model tersebut dipengaruhi oleh jumlah dataset yang kecil.

***Kata kunci**: Klasifikasi, Naive Bayes, Imbalance, SMOTE*

ABSTRACT

Imbalanced data is a condition where the amount of data in a class is much more than other classes, imbalanced data can cause predictions in minority classes to be poor and reduce the performance of the model used. Classification is done using the Naive Bayes algorithm to get a prediction of burnout in college students, the data used amounted to 104 data which has three independent variables to make predictions. Imbalance data will go through a data balancing process using the Synthetic Minority Oversampling Technique (SMOTE), the results of Naive Bayes classification have decreased the performance of the accuracy model by 9.52%, precission 7.8%, recall 9.52% and f1-score 8.55%. The model performance evaluation results are influenced by the small number of datasets.

***Keywords**: Classification, Naive Bayes, Imbalance, SMOTE,*

1. PENDAHULUAN

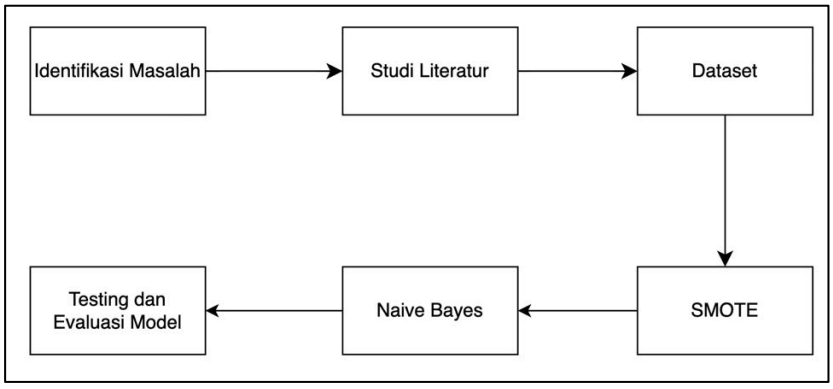
Klasifikasi atau pengelompokan dalam pembelajaran mesin (*machine learning*) merupakan suatu pengelompokan data yang memiliki kelas label atau target. Sehingga algoritma-algoritma dalam menyelesaikan masalah klasifikasi dikategorikan ke dalam *supervised learning* (Ning et al., 2019). Klasifikasi dengan menggunakan data yang tidak seimbang dapat menyebabkan kinerja model menurun dan prediksi yang dilakukan oleh model akan cenderung pada kelas mayoritas (Pujianto et al., 2020). Algoritma Naive Bayes digunakan

untuk memprediksi tingkat *burnout* mahasiswa dengan menggunakan tiga variabel yaitu *Exhaustion*, *Cynicism,* dan *Professional Efficacy* (Schaufeli et al., 2002), terdapat tiga kelas pada target dengan nilai yang tidak seimbang dengan jumlah kelas 'Rendah' = 12, 'Sedang' = 75, dan 'Tinggi' =17 data. Nilai pada kelas tersebut menunjukan bahwa data tersebut memiliki nilai *imbalance*. Untuk menyeimbangkan data tersebut, SMOTE digunakan untuk menambahkan data minoritas agar sama dengan kelas mayoritas (Deng et al., 2021). Penelitian ini bertujuan untuk membandingkan evaluasi kinerja yang diberikan oleh model Naive Bayes dengan menggunakan data *imbalance* dan data setelah proses *balancing* menggunakan teknik SMOTE.

Untuk motivasi dalam penelitian ini, terdapat penelitian dalam literatur dan analisis terkait pembelajaran Naive Bayes dengan menggunakan teknik SMOTE. Pada penelitian (Putri et al., 2021) yang menggunakan algoritma Naive Bayes dan Random Forest untuk klasifikasi data *imbalance* pada status vaksinasi HB agar algoritma tersebut tidak bias terhadap kelas mayoritas dengan menggunakan metode *balancing data Synthetic Minority Oversampling Technique* (SMOTE), penelitian menggunakan data Survei Sosial Ekonomi Nasional pada tahun 2017 dengan kasus berjumlah 2264 dan 14 variabel mendapatkan hasil penggunaan SMOTE pada algoritma Naive Bayes dan Random Forest meningkatkan akurasi identifikasi status non-vaksinasi Hepatitis-B sebesar 30,08% dan 26,09%.

2. METODOLOGI

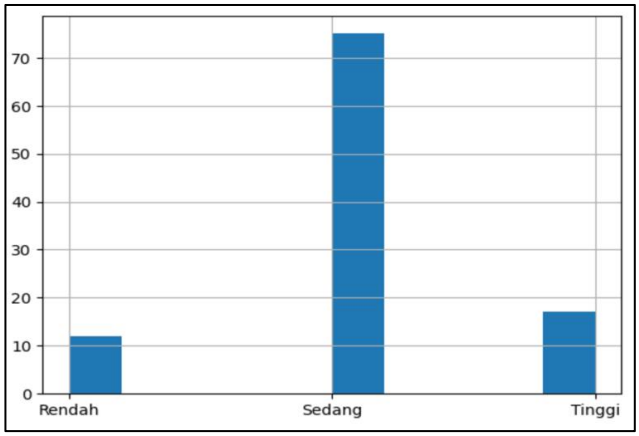
Tahapan yang dilakukan pada penelitian ini terdiri dari beberapa proses analisis kebutuhan, desain sistem, implementasi kode program dan pengujian sistem. Penelitian ini menggunakan algoritma Naive Bayes dengan memanfaatkan library *python3* sebagai dasar dari algoritmanya untuk mengembangkan dan melatih model. Alur penelitan digambarkan pada Gambar 1.



Gambar 1. Alur penelitian

2.1. Dataset

Dataset yang digunakan merupakan dataset *Academic Burnout* yang berasal dari kuesioner *Maslach Burnout Inventory – Student Survey* (MBI-SS) berjumlah 104 data yang diperoleh dari penelitian (Amarsa et al., 2023) dengan jumlah kelas 'Rendah' = 12, 'Sedang' = 75, dan 'Tinggi' =17 data. Grafik jumlah data pada target dapat dilihat pada Gambar 2.



Gambar 2. Jumlah data pada kelas target

2.2 SMOTE

Synthetic Minority Oversampling Technique (SMOTE) adalah salah satu metode *oversampling* untuk menambahkan data sintesis secara acak pada kelas minoritas untuk menghindari data *imbalance* (WIJAYANTI et al., 2021). Teknik SMOTE akan mencari nilai *k* tetangga terdekat pada setiap data di kelas minoritas seperti yang ditunjukan pada Persamaan (1).

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \quad (1)$$

Keterangan:

x_{syn} : data sintesis hasil replikasi

x_i : data yang akan direplikasi

x_{knn} : data yang memiliki jarak terdekat dari data yang direplikasi

δ : bilangan *random* antara 0 dan 1

2.3 Naive Bayes

Algoritma Naive Bayes adalah salah satu algoritma yang memprediksi masa depan peluang berdasarkan pengalaman sebelumnya yang dikenal sebagai teorema *bayes,* teorema *bayes* digunakan untuk pendekatan statistik dalam pengenalan pola dengan asumsi independensi yang kuat (Sun & Ma, 2020). Persamaan (2) menunjukan cara menggunakan algoritma Naive Bayes.

$$P(A \mid B) = \frac{P(B|A).P(A)}{P(B)} \quad (2)$$

Keterangan:

A = hipotesis data B merupakan suatu *class* spesifik

B = data dengan class belum diketahui

P(A|B) = Probabilitas hipotesis A berdasarkan kondisi B (*posterior probability*)

P(A) = Probabilitas hipotesis A (*prior probability*)

P(B|A) = Probabilitas B berdasar kondisi hipotesis A

P(B) = Probabilitas dari B

2.3 Evaluasi Model

Setelah model Naive Bayes telah melakukan tahap pelatihan, model dievaluasi menggunakan *confussion matrix* seperti yang ditunjukan pada "Tabel 1" untuk menghasilkan persamaan (3), (4), (5) dan (6) *accuracy, precision, recall* dan *f1-score*.

Tabel 1. Confussion Matrix

Confussion Matrix		**Prediction**	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} x100\% \quad (3)$$

$$Precision = \frac{TP}{TP + FP} x100\% \quad (4)$$

$$Recall = \frac{TP}{TP + FN} x100\% \quad (5)$$

$$F1\ Score = \frac{2\ x\ Precision\ x\ Recall}{(Precision + Recall)} \quad (6)$$

3. HASIL DAN PEMBAHASAN

Dalam pengujian untuk mengevaluasi kinerja dan mengetahui akurasi klasifikasi Naive Bayes dengan menggunakan data *imbalance* dan data setelah proses *balancing* menggunakan Teknik SMOTE.

3.1 Hasil dan Pembahasan pada Model Naive Bayes

Tahap ini akan memberikan hasil kinerja model Naive Bayes dengan menggunakan data *imbalance* dan data setelah dilakukan proses *balancing* seperti pada Gambar (3). Hasil kinerja model dapat dilihat pada "Tabel 2".

Tabel 2. Jumlah data latih

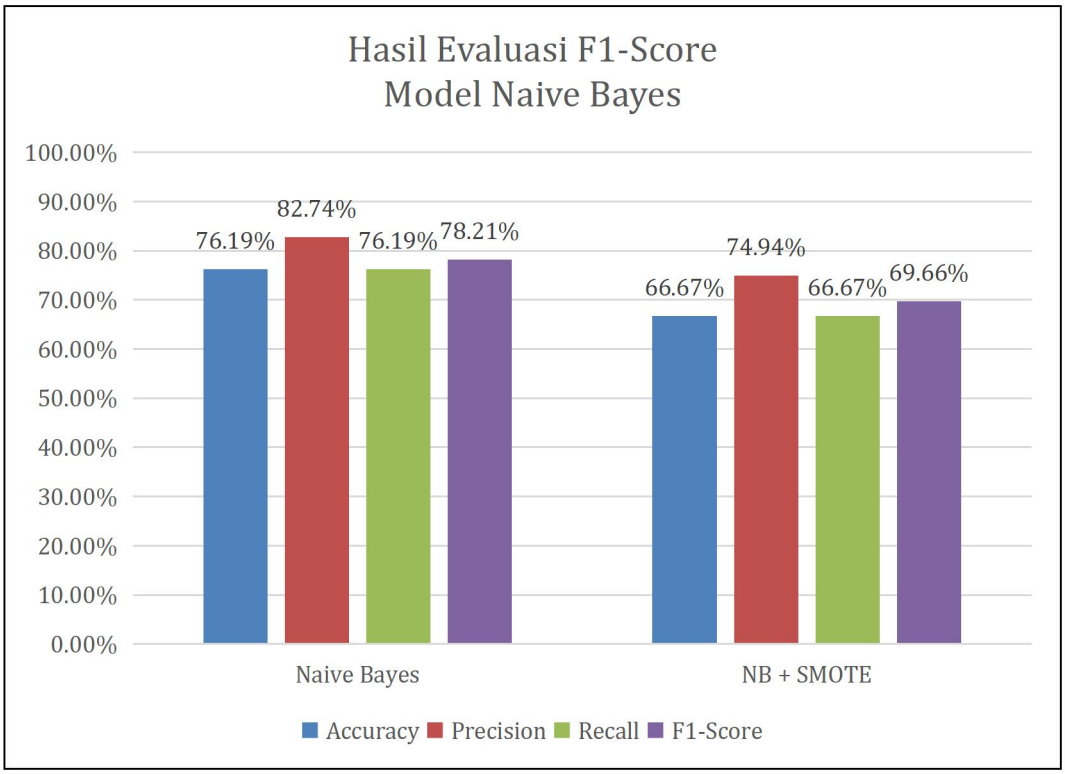
Data	**Kategori**		
	Rendah	**Sedang**	**Tinggi**
Imbalance	10	58	15
Setelah proses *balancing*	58	58	58

"Tabel 2" menunjukan data yang digunakan saat menggunakan data *imbalance* serta data setelah dilakukan proses *balancing* menggunakan SMOTE, nilai pada kelas minoritas yaitu kelas *Rendah* dan kelas *Tinggi* ditambahkan dengan data sintesis. Sehingga data minoritas memiliki nilai yang sama dengan kelas mayoritas yaitu kelas *Sedang* dan "Tabel 3" memperlihatkan hasil evaluasi kinerja model Naive Bayes.

Tabel 3. Hasil evaluasi kinerja model Naive Bayes

Data	**Evaluasi Kinerja Model**			
	Accuracy	**Precission**	**Recall**	**F1-Score**
Imbalance	76,19%	82,74%	76,19%	78,21%
Sesudah proses *balancing*	66,67 %	74,94 %	66,67 %	69,66 %

Berdasarkan pengamatan pada "Tabel 3", model Naive Bayes dengan penggunaan data yang telah dilakukan proses *balancing* menggunakan Teknik SMOTE mengalami penurunan *accuracy* sebesar 9,52%, *precission* 7,8%, *recall* 9,52% dan *f1-score* 8,55%. Gambar (3) menunjukan grafik antara model Naive Bayes dan Naive Bayes dengan Teknik SMOTE.



Gambar 3. Hasil Evaluasi Model Naive Bayes

4. KESIMPULAN

Pada penelitian ini, dimana proses pengujian menggunakan 104 data *academic burnout* dengan tiga kelas pada target kelas 'Rendah' = 12, 'Sedang' = 75, dan 'Tinggi' =17 data. Data tersebut memiliki nilai yang tidak seimbang atau *imbalance* karena jarak antara kelas yang memiliki nilai yang jauh. Penggunaan model Naive Bayes menghasilkan evaluasi kinerja *accuracy* sebesar 76,19%, *precision* 82,74%, *recall* 76,19% dan *f1-score* 78,21%. Sementara, hasil penggunaan model Naive Bayes dengan teknik SMOTE menghasilkan evaluasi kinerja *accuracy* sebesar 66,67%, *precision* 74,94%, *recall* 66,67% dan *f1-score* 69,66%. Penggunaan teknik SMOTE pada model Naive Bayes menghasilkan penurunan *accuracy* sebesar 9,52%, *precission* 7,8%, *recall* 9,52% dan *f1-score* 8,55%. Nilai tersebut dihasilkan dari dataset yang digunakan tidak cukup banyak untuk melatih model dengan baik.

References

Amarsa, R. R., Ramdhani, R. N., Taufiq, A., & Saripah, I. (2023). Kecenderungan Academic Burnout Pada Mahasiswa Bimbingan dan Konseling Universitas Pendidikan Indonesia. *Jurnal Bimbingan Dan Konseling*, *7*(2), 395–405. https://doi.org/https://doi.org/10.31316/gcouns.v7i03.4477

Deng, M., Guo, Y., Wang, C., & Wu, F. (2021). An oversampling method for multi-class imbalanced data based on composite weights. *PLoS ONE*, *16*(11 November). https://doi.org/10.1371/journal.pone.0259227

Ning, B., Junwei, W., & Feng, H. (2019). Spam message classification based on the naïve Bayes classification algorithm. *IAENG International Journal of Computer Science*, *46*(1).
Pujianto, U., Agung Prasetyo, W., & Rakhmat Taufani, A. (2020). Students Academic Performance Prediction with k-Nearest Neighbor and C4.5 on SMOTE-balanced data. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*. https://doi.org/10.1109/ISRITI51436.2020.9315439
Putri, V. M., Masjkur, M., & Suhaeni, C. (2021). Performance of SMOTE in a random forest and naive Bayes classifier for imbalanced Hepatitis-B vaccination status. *Journal of Physics: Conference Series*, *1863*(1). https://doi.org/10.1088/1742-6596/1863/1/012073
Schaufeli, W. B., Martinez, I. M., Pinto, A. M., Salanova, M., & Bakker, A. B. (2002). Burnout and engagement in university students a cross-national study. Journal of cross-cultural psychology. In *Journal of cross-cultural psychology* (Vol. 33, Issue 5).
Sun, Y., & Ma, Y. (2020). Application of Classification Algorithm Based on Naive Bayes in Data Analysis of Fitness Test. *Journal of Physics: Conference Series*, *1648*(4). https://doi.org/10.1088/1742-6596/1648/4/042078
WIJAYANTI, N. P. Y. T., N. KENCANA, E., & SUMARJAYA, I. W. (2021). SMOTE: POTENSI DAN KEKURANGANNYA PADA SURVEI. *E-Jurnal Matematika*, *10*(4). https://doi.org/10.24843/mtk.2021.v10.i04.p348