

KLASIFIKASI DATA POKOK PENDIDIKAN SEKOLAH DASAR DAN SEKOLAH MENENGAH PERTAMA MENGGUNAKAN ALGORITMA *NAÏVE BAYES* SERTA *C.45*

1st Rafi Pasya Sidik
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
if18.rafidisidik@mhs.ubpkarawang.ac.id &
081285967520

2nd Tatang Rohana
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
tatang.rohana@ubpkarawang.ac.id

3rd Rahmat
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
rahmat@ubpkarawang.ac.id

Abstract— Dapodik digunakan untuk mengumpulkan semua data yang terkait dengan data kurikulum sekolah dan institusi, data siswa, data guru dan staf serta data infrastruktur dari setiap studi sekolah di seluruh Indonesia.. Sarana adalah alat langsung untuk mencapai tujuan pendidikan, lalu prasarana adalah alat tidak langsung untuk mencapai tujuan. Penelitian ini menggunakan Data Dapodik SD dan SMP. Karna pada bidang program dan pelaporan mengalami permasalahan dalam pemberian bantuan sarana dan prasarana pada jenjang SD, SMP sehingga hasil dari klasifikasi dapat ditemukan sekolah yang mengalami sarana dan prasarana yang kurang merata. Penelitian ini dilakukan pengklasifikasian sarana dan prasarana berdasarkan data Dapodik SD, SMP Karawang dengan menggunakan dua algoritma pengklasifikasian. Teknik *Data Mining* klasifikasi yang digunakan yaitu *Naïve Bayes* dan *C.45*. klasifikasi yang akan dilakukan menggunakan 3 kelompok yaitu banyak, sedang, sedikit pada Data Dapodik SD, SMP tahun 2020/2021. Data yang digunakan 100 dibagi menjadi data training 70% dan data testing 30% lanjut perhitungan algoritma *Naïve Bayes* dengan nilai akurasi 63% sedang kan algoritma *C.45* dengan nilai akurasi 73%.

Kata kunci — *C.45*, Dapodik, *Data Mining*, Klasifikasi, Sarana dan Prasarana, *Naive Bayes*.

I. PENDAHULUAN

Dinas Pendidikan Pemuda dan Olahraga memiliki peran penting dalam kelangsungan di dunia pendidikan. Dinas Pendidikan Pemuda dan Olahraga memiliki tugas melayani Pendidikan dan mengawasi dapodik di tiap daerah Kota, Provinsi[1]. Dapodik ialah aplikasi komputer yang dibuat oleh Kementerian Pendidikan dan Kebudayaan Republik Indonesia agar sekolah dapat melaporkan Dapodiknya secara langsung ke jurusan *online* melalui internet tanpa terikat oleh masalah ruang, cara maupun waktu. Sarana adalah alat langsung untuk mencapai tujuan pendidikan. Misalnya buku, perpustakaan, laboratorium, dan lain sebagainya[2]. Infrastruktur adalah alat tidak langsung untuk tujuan seperti lokasi atau tempat, sekolah, lapangan olahraga, uang, dll. [2]

Penelitian ini menggunakan Data Dapodik SD dan SMP. Alasan penelitian ini hanya menggunakan jenjang SD dan SMP saja, karna pada bidang program dan pelaporan mengalami permasalahan dalam pemberian bantuan sarana dan prasarana pada jenjang SD, SMP yang belum tepat karna masih ada beberapa sekolah yang harus bergantian ruang kelas, meja, kursi. Ini menjadi permasalahan tidak meratanya dalam pemberian Sarana dan Prasarana. Solusi dari permasalahan tidak meratanya Sarana dan Prasarana SD, SMP dengan menggunakan klasifikasi Sarana dan Prasarana berdasarkan data Dapodik SD, SMP sehingga hasil dari klasifikasi dapat ditemukan sekolah mana aja yang mengalami Sarana dan Prasarana yang kurang merata. Hasil klasifikasi tersebut dapat diserahkan kepada Pemerintah[1].

Penelitian ini dilakukan pengklasifikasian Sarana dan Prasarana berdasarkan data Dapodik SD, SMP Karawang dengan menggunakan dua algoritma pengklasifikasian. Teknik klasifikasi yang digunakan yaitu *Naïve Bayes*[3] dan *C.45*. [3] Klasifikasi yang akan dilakukan menggunakan 3 kelompok yaitu banyak, sedang, sedikit pada Data Dapodik SD, SMP tahun 2020/2021.

II. DATA & METODE

A. Sekolah Dasar

Anak sekolah dasar adalah mereka yang berusia antara 6 dan 12 tahun, atau yang biasa disebut dengan masa intelektual. Pengetahuan anak akan meningkat pesat seiring bertambahnya usia, penguasaan keterampilan akan semakin beragam. Minat anak pada tahap ini terutama terfokus pada apa saja yang bergerak secara fleksibel. Implikasinya, anak cenderung melakukan berbagai kegiatan yang akan membantu perkembangannya di kemudian hari.[4].

B. Sekolah Menengah Pertama

Sekolah Menengah Pertama, disingkat SMP, ialah tahapan pembelajaran formal dasar di Indonesia setelah berhasil pembelajaran dasar (atau sederajat). Siklus sekolah menengah berlangsung selama 3 tahun, dari kelas 7 hingga kelas 9. Saat ini, sekolah menengah adalah program pemerataan 9 tahun (SD, SMP) [5].

C. Klasifikasi

Proses menemukan model (atau fungsi) yang menggambarkan dan membedakan kelas atau konsep data yang akan digunakan untuk untuk memprediksikan kelas objek yang label kelasnya tidak ditentukan [6].

D. Algoritma Naïve Bayes

Algoritma Naive Bayes adalah metode klasifikasi menggunakan probabilitas dan metode statistik yang diusulkan oleh ilmuwan Inggris Thomas Bayes. Naive Bayes menghitung satu set probabilitas serta menambahkan frekuensi serta perpaduan skala per kumpulan data. Algoritma mengaplikasikan teorema Bayes dan mengasumsikan bahwa seluruh properti independen maupun independen dari nilai yang diberikan ke variabel kelas. Naive Bayes mampu membangun model prediktif yang cepat dan juga menyediakan metode baru untuk mengeksplorasi dan memahami data.

Algoritma ini mempunyai keuntungan, karena metode ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menetapkan jangka waktu parameter yang diperlukan dalam proses pengklasifikasian[7]. Dengan kata lain, dalam *Naive bayes*, model yang digunakan adalah “model fitur independen”[8].

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

Keterangan :

- P = Data dengan kelas yang tidak diketahui.
 H = Data Hopotesis X yaitu suatu class yang spesifik.
 P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x (probabilitas posterior).
 P(H) = Probabilitas hipotesis H (probabilitas sebelumnya).
 P(X|H) = Probabilitas X berdasarkan kondisi hipotesis H.
 P(X) = Probabilitas X

E. Algoritma C.45

Algoritma C4.5 adalah algoritma pohon keputusan yang amat efisien buat klasifikasi. Algoritma C4.5 menjadi evolusi atas algoritma Iterative Dichotomiser (ID3) yang terdapat oleh J Ross Quinlan. Pohon penghabisan ini dibangun bersama memotong bahan ala rekursif sehingga setiap putaran menyertakan data melaui kelas yang sama. Formulir split yang digunakan guna memisahkan data terkait pada macam atribut yang digunakan dalam pemisahan. Menurut Tabrani (2016), serangkaian perbaikan pada ID3 telah mencapai puncaknya pada penciptaan sistem pohon keputusan yang realistis dan berdampak, yaitu C4.5 [9].

Secara umum algoritma *decision tree* memiliki tahapan sebagai berikut.

- Memilih Atribut selaku Root Memilih atribut demi root dapat dihitung menggunakan melihat nilai yang diperoleh melalui setiap atribut. Nilai tertinggi yang dicapai untuk sebagai akar pertama. tampilan padan buat menemukan nilai gain dapat dilihat di perbandingan 2.

$$\text{Gain}(S, A) = \text{Entropy}(s) - \sum_{i=1}^s \frac{|S_i|}{|S|} * \text{Entropy}(s) \quad (2)$$

Keterangan :

- S : Himpunan kasus
 A : Atribut
 n : Jumlah partisi atribut A
 |Si| : Jumlah kasus pada partisi ke – i
 |S| : Jumlah kasus pada S

Sedangkan nilai *entropy* sendiri bisa diperoleh dari persamaan 3 berikut.

$$\text{Entropy}(S) = \sum_{i=0}^n - p_i * \log_2 p_i \quad (3)$$

Keterangan :

- S : Himpunan kasus
 n : jumlah partisi S
 Pi : Proporsi Si terhadap S

- Ulangi metode di atas sepadan seluruh data terbagi.
- metode peniruan dari teknik *decision tree* ini buat berhenti jika :

- Semua data telah terbagi rata.
- Tidak ada lagi atribut yang bisa di bagi lagi.
- Tidak ada data *record* dalam cabang yang kosong

Pada algoritma C4.5 terdapat pembeda dari ID3 yakni adanya rasio gain yang berfungsi sebagai pemecah atribut. Persamaan 4 dibawah memperlihatkan bagaimana RasioGain diperoleh.

$$\text{RasioGain}(s, j) = \frac{\text{Gain}(s, j)}{\text{splitinfo}(s, j)} \quad (4)$$

Keterangan :

- S : Himpunan kasus
 j : fitur ke- j

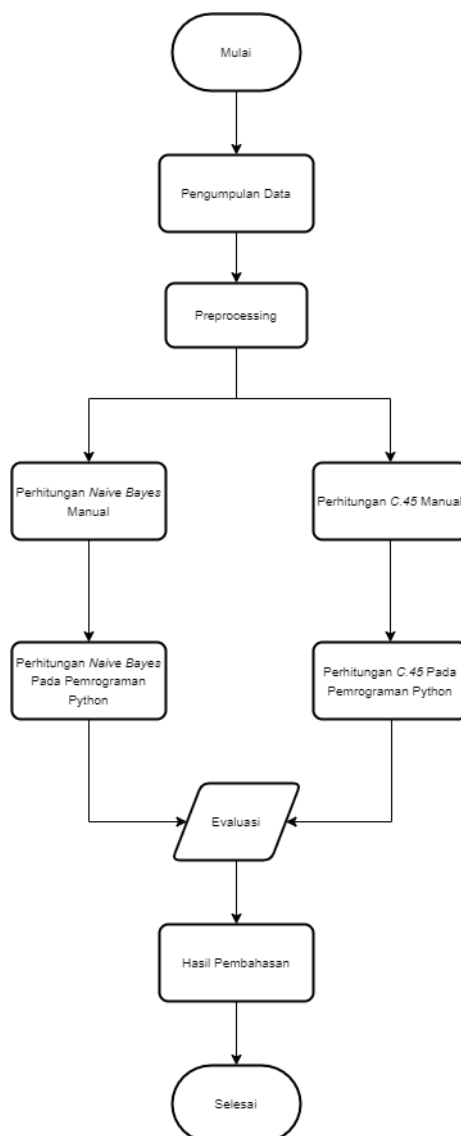
Sedang buat *split* info terima diperoleh pakai rumus dari Persamaan 5.

$$Splitinfo(s, j) = \sum_{i=1}^k p(V_i | s) \log_2 p(V_i | s) \quad (5)$$

Keterangan :

k : jumlah pemecahan

F. Prosedur Penelitian



Gambar 1 Prosedur Penelitian

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang didapatkan dari Dinas Pendidikan Pemuda dan Olahraga data Dapodik Sekolah SD SMP Kabupaten Karawang tahun 2020 - 2021. Yang dapat dilihat pada Tabel 4.1. berikut.

Tabel 1 Data Mentah

No.	NPSN	Nama Sekolah	Ruang Kelas	Meja Siswa	Kursi Siswa	Peserta Didik	Terpenuhi
1.	20236045	SDN CICINDE UTARA I	18	251	501	501	Ya
2.	20236617	SDN PAMEKARAN II	6	75	150	150	Ya

No.	NPSN	Nama Sekolah	Ruang Kelas	Meja Siswa	Kursi Siswa	Peserta Didik	Terpenuhi
3.	20236204	SDN GEMPOL KOLOT II	6	70	140	140	Ya
4.	20236195	SDN GEMBONGAN II	7	59	65	144	Tidak
5.	20236537	SDN MEKARASIH I	6	78	99	163	Tidak
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
96.	20246721	SMPN 2 KOTABARU	33	639	1278	935	Ya
97.	20217925	SMPN 3 KOTABARU	12	185	369	246	Ya
98.	20217866	SMPN 4 KOTABARU	26	591	1038	682	Ya
99.	20217907	SMPN 1 KUTAWALUY A	45	660	1320	1380	Tidak
100.	20217929	SMPN 2 KUTAWALUY A	20	692	798	542	Ya

B. Preprocessing Data

Pada tahap ini sebelum data digunakan data harus diubah menjadi kategori dulu. Tiap variabel memiliki range tertentu yang berbeda yaitu.

Tabel 2 Kategori Data

Kategori Data		
Ruang kelas	50 - 31	Banyak
	30 - 21	Sedang
	20 - 1	Sedikit
Meja Siswa	1300 - 700	Banyak
	699 - 300	Sedang
	299 - 1	Sedikit
Kursi Siswa	2500 - 1500	Banyak
	1499 - 500	Sedang
	499 - 1	Sedikit
Pesertas Didik	1380 - 800	Banyak
	799 - 500	Sedang
	499 - 1	Sedikit

Ketika data sudah dikategorikan maka akan melalui tahap Preprocessing yang akan hanya memakai beberapa variabel diantaranya Ruang Kelas, Meja Siswa Kursi Siswa, Peserta Didik, dan Terpenuhi sebagai class/label nya. Saat digunakan pada penghitungan manual dan penerepan algoritma *Naïve Bayes* dengan *C.45* pada aplikasi Python.

C. Perhitungan Manual *Naïve Bayes*

Perhitungan Manual *Naïve Bayes* Menggunakan Microsoft Excel 2021 menghasilkan class prediksi dengan menggunakan Langkah – Langkah perhitungan yang disebutkan sebelumnya sehingga hasil akhir perhitungan manual *Naïve Bayes* dapat dilihat pada tabel 3 dimana class prediksi ya maka kategori ya lebih besar di banding tidak.

Tabel 3 Hasil Perhitungan Manual *Naïve Bayes*

Class Prediction	Ya	Tidak
ya	0,01008	0,00000
ya	0,05701	0,02341
tidak	0,05701	0,18729
ya	0,00806	0,00000
ya	0,00230	0,00005
...	...	...
...	...	...
...	...	...
tidak	0,05701	0,18729
ya	0,01008	0,00000
tidak	0,05701	0,18729
tidak	0,05701	0,18729
tidak	0,05701	0,18729

D. Perhitungan Program Python Algoritma Naïve Bayes

Hasil prediksi dari data testing sebanyak 30 data dengan keterangan 0 untuk “Tidak” dan 1 untuk “Ya”. Dengan akurasi data testing yang terdiri dari 8 *True Positive* dan 11 *False Negative* dengan akurasi data testing 63,33%.

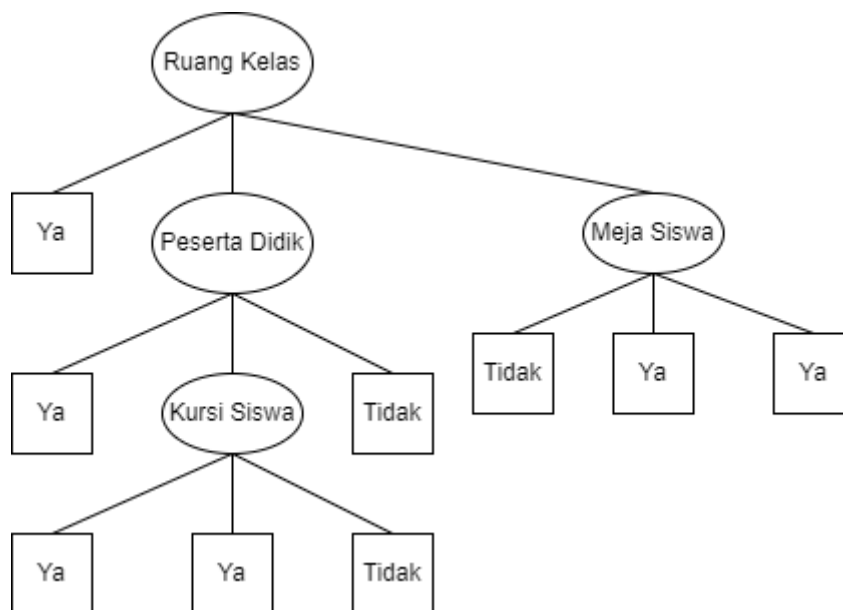
```
In [10]: #Prediksi terhadap model training yang telah dibuat
prediction = NaiveBayes.predict(X_test)
accuracies = accuracy_score(y_test, prediction)

print(prediction)
print(accuracies)

[1 1 0 1 1 0 0 1 1 0 1 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0 0]
0.6333333333333333
```

Gambar 1 Hasil Perhitungan Program Python Algoritma Naïve Bayes

E. Perhitungan Manual C.45



Gambar 2 Pohon Keputusan Perhitungan Manual C.45

Pada Gambar 4.3 diatas ruang kelas menjadi menjadi *node* 1. Kemudian, dilanjutkan pada meja siswa menjadi *node* 1.2, lalu pada peserta didik, kursi siswa menjadi *node* 1.3 dan *node* 1.4.

Berdasarkan Pohon Keputusan diatas menghasilkan *rule* sebagai berikut.

1. IF Ruang Kelas, Sedikit = Meja Siswa, Sedikit = Ya
2. IF Ruang Kelas, Sedikit = Meja Siswa, Sedang = Ya
3. IF Ruang Kelas, Sedikit = Meja Siswa, Banyak = Tidak
4. IF Ruang Kelas, Banyak = Ya

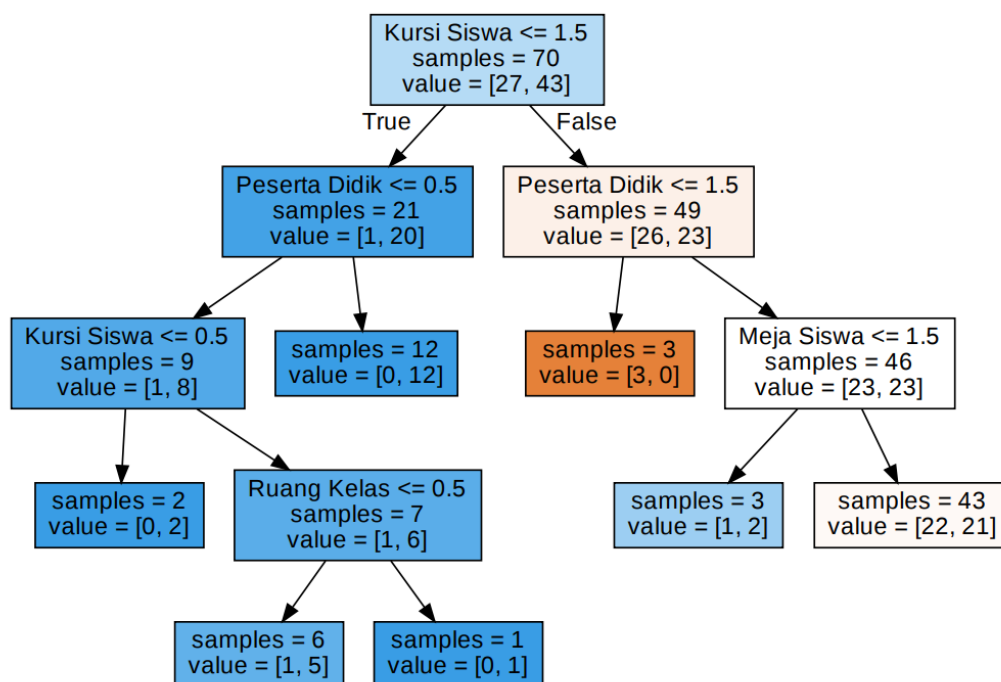
5. IF Ruang Kelas, Sedang = Peserta Didik, Sedikit = Tidak
6. IF Ruang Kelas, Sedang = Peserta Didik, Banyak = Ya
7. IF Ruang Kelas, Sedang = Peserta Didik, Sedang = Kursi Siswa, Sedikit = Tidak
8. IF Ruang Kelas, Sedang = Peserta Didik, Sedang = Kursi Siswa, Sedang = Ya
9. IF Ruang Kelas, Sedang = Peserta Didik, Sedang = Kursi Siswa, Banyak = Ya

Maka hasil *Confusion Matrix* Perhitungan Manual Algoritma C.45 pada Excel dengan data testing 30% sama dengan Algoritma Naïve Bayes Sebagai berikut.

Tabel 4 Hasil Confusion Matrix Perhitungan Manual C.45

Prediksi	Aktual	
	Ya	Tidak
Ya	11	2
Tidak	6	11

F. Perhitungan Program Python Algoritma C.45



Gambar 3 Pohon Keputusan Program Python Algoritma C.45

Gambar 3 Pohon Keputusan di atas merupakan hasil perhitungan program Python dengan Algoritma C.45 dengan data testing 30%.

```

confusion matrix
[[11  2]
 [ 6 11]]
Tingkat Akurasi Algoritma c.45
Akurasi :          precision    recall  f1-score   support

      0       0.65      0.85      0.73        13
      1       0.85      0.65      0.73        17

 accuracy          0.73        30
 macro avg       0.75      0.75      0.73        30
 weighted avg    0.76      0.73      0.73        30

```

Tingkat Akurasi: 73 persen

Gambar 4 Hasil Perhitungan Program Python Algoritma C.45

Hasil Perhitungan Program Python Algoritma C.45 dengan data testing yang berbeda mendapat akurasi sebesar 73% dari 11 TP, 11 TN, 2 FP, 6 FN.

G. Evaluasi

Setelah melakukan perhitungan manual dan perhitungan program python selanjutnya adalah melakukan evaluasi dengan menggunakan metode *Confusion Matrix*, dalam metode ini akan mencari nilai performa akurasi, presisi, *recall*.

1. Algoritma Naïve Bayes
 - a. Perhitungan Manual

Tabel 5 *Confusion Matrix* Perhitungan Manual Naïve Bayes

Aktual		
Prediksi	Ya	Tidak
Ya	11	10
Tidak	1	8

- b. Perhitungan Program Python

Tabel 6 *Confusion Matrix* Perhitungan Program Python Algoritma Naïve Bayes

Aktual		
Prediksi	Ya	Tidak
Ya	8	1
Tidak	10	11

2. Algoritma C.45
 - a. Perhitungan Manual

Tabel 7 *Confusion Matrix* Perhitungan Manual C.45

Aktual		
Prediksi	Ya	Tidak
Ya	11	2
Tidak	6	11

- b. Perhitungan Program Python

Tabel 8 *Confusion Matrix* Perhitungan Program Python Algoritma C.45

Aktual		
Prediksi	Ya	Tidak
Ya	11	2
Tidak	6	11

H. Hasil Pembahasan

Tabel 9 Hasil Evaluasi Algoritma Naïve Bayes dan C.45

Jenis Perhitungan Algoritma Naïve Bayes		
<i>Confusion Matrix</i>	Manual	Python

<i>True Positive</i>	11	8
<i>True Negative</i>	1	11
<i>False Positive</i>	10	1
<i>False Negative</i>	8	10

Jenis Perhitungan Algoritma C.45		
<i>Confusion Matrix</i>	Manual	Python
<i>True Positive</i>	11	11
<i>True Negative</i>	11	11
<i>False Positive</i>	2	2
<i>False Negative</i>	6	6

Pada Tabel 9 Hasil hasil penelitian di atas adalah hasil perhitungan manual excel dan program python dengan data testing 30% hasil akurasi 63.33% untuk Algoritma *Naïve Bayes* dan hasil alurasi 73.33% untuk Algoritma C.45.

IV. KESIMPULAN

Hasil Klasifikasi Data Dapodik pada jenjang SD, dan SMP dengan menggunakan data testing hasil akurasi menggunakan manual dan python dengan Algoritma *Naïve Bayes* dengan Akurasi 63,33% dan hasil akurasi Menggunakan manual dan python dengan Algoritma C.45 memiliki Akurasi 73,33%.

V. PENGAKUAN

Naskah ilmiah ini adalah sebagian dari penelitian Tugas Akhir milik Rafi Pasya Sidik, dengan judul Klasifikasi Data Pokok Pendidikan Sekolah dasar dan Sekolah Menengah Pertama Menggunakan Algoritma *Naïve Bayes* Serta C.45 yang dibimbing oleh Tatang Rohana dan Rahmat.

DAFTAR PUSTAKA

- [1] F. L. Y. Adha, M. L. Hamzah, I. Maita, Megawati, dan A. Marsal, "Analisis Penerimaan Pengguna Dapodik Sekolah Dasar Kecamatan Tampan Menggunakan Model TAM dan EUCS," *J. Sains, Teknol. dan Ind.*, vol. 18, no. 2, hal. 196–205, 2021.
- [2] I. M. Sinta, "Manajemen Sarana Dan Prasarana," *J. Isema Islam. Educ. Manag.*, vol. 4, no. 1, hal. 77–92, 2019, doi: 10.15575/isema.v4i1.5645.
- [3] C. Fadlan, S. Ningsih, dan A. P. Windarto, "Penerapan Metode *Naïve Bayes* Dalam Klasifikasi Kelayakan Keluarga Penerima Beras Rastra," *J. Tek. Inform. Musirawas*, vol. 3, no. 1, hal. 1, 2018, doi: 10.32767/jutim.v3i1.286.
- [4] A. FAHLEVI, "Analisis Sistem Informasi Aplikasi Dapodik Sekolah dasar," *Simki.Unpkediri.Ac.Id*, hal. 2–12, 2019, [Daring]. Tersedia pada: http://simki.unpkediri.ac.id/mahasiswa/file_artikel/2019/13.1.03.03.0014.pdf
- [5] M. Irvan, Y. Purnama, dan R. Vhalery, "Model Prediktif Untuk Akreditasi Sekolah Tingkat Sekolah Menengah Pertama (Smp)," *Res. Dev. J. Educ.*, vol. 5, no. 2, hal. 03, 2019, doi: 10.30998/rdje.v5i2.3747.
- [6] Triase dan Samsudin, "Implementasi Data Mining Dalam Mengklasifikasikan Ukt (Uang Kuliah Tunggal) Pada Uin Sumatera Utara Medan," *J. Teknol. Inf.*, vol. 4, no. 2, hal. 370–376, 2020, [Daring]. Tersedia pada: <http://www.jurnal.una.ac.id/index.php/jurti/article/view/1711>
- [7] Anas, F. Tempola, dan A. Khairan, "Hybrid fuzzy dan *Naive Bayes* Dalam Penentuan Status UKT," *PROtek*, vol. 06, no. 1, hal. 2–7, 2019.
- [8] S. Suprianto, "Implementasi Algoritma *Naive Bayes* Untuk Menentukan Lokasi Strategis Dalam Membuka Usaha Menengah Ke Bawah di Kota Medan (Studi Kasus: Disperindag Kota Medan)," *J. Sist. Komput. dan Inform.*, vol. 1, no. 2, hal. 125, 2020, doi: 10.30865/json.v1i2.1939.
- [9] T. Akhir dan A. Nirwana, "PADA BANK MENGGUNAKAN ALGORITMA DECISION TREE C4 . 5," 2021.