ANALISIS PERBANDINGAN ALGORITMA C4.5 DAN NAIVE BAYES DALAM PENENTUAN PENJURUSAN SISWA SMA NEGERI 11 PEKANBARU

Yogi Hadinakris[1], Loneli Costaner[2], Ahmad Zamsuri[3]
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer
Universitas Lancang Kuning
Jl. Yos Sudarso KM. 8 Rumbai, Pekanbaru, Riau, telp. 0811 753 2015
e-mail: [1] yogihadinakriss@gmail.com, [2] lonelicostaner@unilak.ac.id,
[3] ahmadzamsuri@unilak.ac.id

Abstrak

Penentuan penjurusan akan berpengaruh pada jenjang akademik berikutnya dan akan membawa pengaruh dalam bidang ilmu atau studi bagi siswasiswi yang akan meneruskan ke jenjang perguruan tinggi berikutnya, sehingga penjurusan yang tidak tepat akan merugikan siswa pada masa depannya. Terdapat kelemahan dalam penentuan penjurusan selama ini, diantaranya berpatokan pada keinginan siswa tanpa meninjau bagaimana latar belakang nilai akademisnya. Akibatnya jurusan yang dipilih terkadang menjadi masalah bagi siswa di kemudian hari, contohnya saja nilai akademik yang kurang maksimal, pada akhirnya pemilihan program studi saat meneruskan ke jenjang perguruan tinggi terkendala akibat jurusan SMA yang tidak tepat. Pada penelitian ini dilakukan klasifikasi penentuan penjurusan siswa dengan algoritma C4.5 dan Naive Bayes, penelitian ini menggunakan data nilai siswa beserta jurusannya, pengujian metode pada penelitian ini dengan menggunakan Pemrograman Python. berdasarkan hasil dari tingkat akurasi algoritma C4.5 dan Naive Bayes tingkat akurasi C4.5 lebih unggul dari Naive Bayes dengan tingkat akurasi sebesar 86,7% sedangkan Naive Bayes memiliki tingkat akurasi sebesar 76%

***Kata Kunci**: : Perbandingan Algoritma, penjurusan, DataMining, C4.5, Naïve Bayes*

Abstract

The determination of majors will affect the next academic level and will bring influence in the field of science or study for students who will continue to the next level of college, so that improper majors will harm students in their future. There are weaknesses in the determination of majors so far, including based on the wishes of students without reviewing the background of their academic grades. As a result, the chosen major sometimes becomes a problem for students in the future, for example, academic grades that are not optimal, in the end the selection of study programs when continuing to the college level is constrained due to improper high school majors. In this study, the classification of determining student majors with the C4.5 and Naive Bayes algorithms was carried out, this study used data on student grades and majors, testing methods in this study using Python programming. based on the results of the accuracy level of the C4.5 and Naive Bayes algorithms, the accuracy level of C4.5 is superior to Naive Bayes with an accuracy level of 86.7% while Naive Bayes has an accuracy level of 76%.

***Keywords**: Algorithm Comparison, majors, DataMining, C4.5, Naïve Bayes*

1. PENDAHULUAN

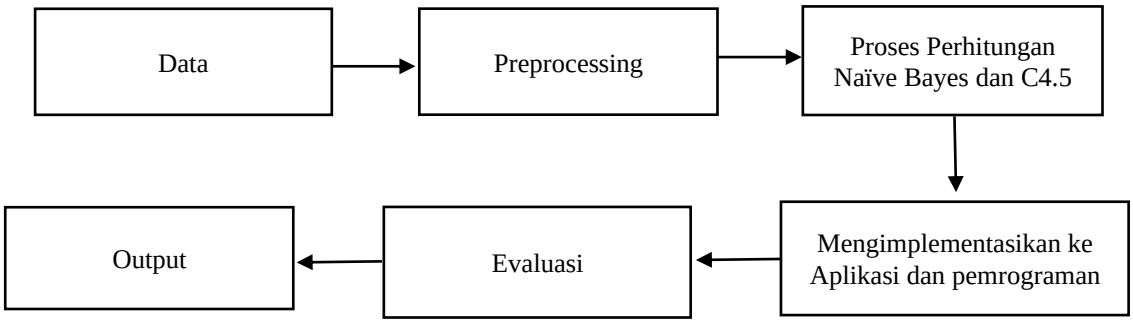
Kemajuan teknologi informasi saat ini, mendapatkan informasi telah menjadi lebih cepat dan mudah, memengaruhi berbagai aspek kehidupan, termasuk pendidikan. SMA Negeri 11 Pekanbaru, yang didirikan pada tahun 1995, memiliki dua jurusan, MIPA dan IPS, yang bertujuan untuk mengarahkan peserta didik dalam meningkatkan kemampuan dan minat mereka. Namun, penentuan jurusan yang tepat sangat penting karena dapat memengaruhi jenjang akademik dan pemilihan program studi di perguruan tinggi. Sayangnya, penentuan jurusan selama ini seringkali hanya berdasarkan keinginan siswa tanpa mempertimbangkan latar belakang nilai akademis, yang bisa berdampak negatif di masa depan.

Pendidikan merupakan upaya sadar dan terencana untuk mencapai suasana belajar dan proses pembelajaran supaya peserta didik dapat secara aktif meningkatkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri, kepribadian, kecerdasan, akhlak mulia, dan keterampilan yang diperlukan dirinya, masyarakat, bangsa dan negara.

Melalui metode ilmiah, prediksi digunakan untuk mengantisipasi peristiwa di masa depan dengan memanfaatkan informasi historis. Beberapa metode prediksi yang umum digunakan termasuk C4.5, naive bayes, K-Nearest Neighbor (K-NN), dan lainnya. Oleh karena itu, penulis melakukan penelitian dengan judul "Analisis Perbandingan Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penentuan Penjurusan Siswa di SMA Negeri 11 Pekanbaru.

2. METODE PENELITIAN

Penelitian ini dilakukan dengan beberapa tahap, berikut tahapan penelitian yang dilakukan seperti pada Gambar 1



Peneliti menggunakan data SMA yang meliputi data nilai Raport siswa kelas X, kemudian data tersebut diolah kembali atau bisa disebut dengan data cleaning untuk menghindari adanya duplikasi dan kesalahan pada data, setelah melakukan proses data Cleaning kemudian melakukan perhitungan manual menggunakan algoritma C4.5 dan Naive Bayes. setelah melakukan perhitungan secara manual, data tersebut diolah menggunakan Python. Kemudian akan memproses data tersebut lalu melakukan evaluasi yang akan menghasilkan Output.

2.1. Data Mining

Data Mining adalah suatu proses pencarian pola dari data-data dengan jumlah yang sangat banyak dan berada dalam suatu tempat penyimpanan yang banyak menggunakan teknologi pengenal pola, teknik statistik, dan matematika. Ada beberapa karakteristik pada Data Mining yaitu sebagai berikut (Sibarani, 2020). :

1. Data mining berkaitan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang belum diketahui sebelumnya.
2. Data mining biasa memerlukan data yang sangat besar.
3. Umumnya data yang besar diperlukan untuk menciptakan hasil lebih dipercaya.
4. Data mining berfungsi untuk membuat keputusan yang kritis, terutama dalam strategi.

Umumnya Data Mining ialah salah satu bagian dari tahap Knowledge Discovery in Database (KDD) yang bertujuan untuk mendapatkan pola atau model dari data dengan menggunakan sebuah algoritma yang spesifik. Berikut proses KDD (Fatmawati, 2019) :

1) Data Selection : Proses memilih data dari sekumpulan data operasional perlu diselesaikan sebelum tahap penggalian informasi KDD dimulai.
2) Preprocessing : Proses cleaning perlu dilakukan sebelum proses Data Mining akan dilaksanakan, tujuannya untuk menghapus duplikasi data, meninjau data yang tidak konsisten, dan memulihkan kesalahan data, seperti kesalahan cetak (tipografi). Juga melangsungkan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.
3) Transformation : Proses Coding pada data yang telah ditentukan, agar data yang ada sesuai dengan proses Data Mining. Proses Coding dalam KDD ialah proses kreatif dan sangat bergantung pada jenis atau pola informasi yang akan ditemukan dalam Database.
4) Data Mining : Proses menemukan pola atau informasi menarik dalam data yang telah ditentukan dengan menerapkan teknik atau metode tertentu.
5) Interpretation / Evaluation : Pola informasi yang didapatkan dari proses Data Mining perlu digambarkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan Interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada pada sebelumnya atau tidak.

2.2. Naive bayes

Naive Bayes ialah suatu metode yang digunakan untuk menjalankan klasifikasi probabilistik sederhana yaitu dengan cara memperkirakan probabilitas sesuai dengan total frekuensi dan kombinasi yang ada pada data. Algoritma ini berpegang pada prinsip dari teori Bayes yaitu dengan anggapan bahwa semua atribut itu tidak berkaitan satu dengan yang lain. Berikut merupakan persamaan yang dipakai dalam algoritma Naïve Bayes (Kadafi, 2018)

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

Keterangan :
X = Data dengan kelas yang belum diketahui
H = Hipotesa data kelas
P(H|X) = Probabilitas hipotesa H berdasar kondisi X
P(H) = Probabilitas hipotesa H
P(X|H) = Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) = Probabilitas X

2.2. Algoritma C4.5

Algoritma C4.5 adalah salah satu teknik klasifikasi/Prediksi pada Machine Learning yang diterapkan pada proses Data Mining dengan merangkai sebuah pohon keputusan (Decision Tree) yang digambarkan dalam bentuk aturan. Algoritma C4.5 diciptakan oleh Ross Quinlan dan diterapkan untuk membuat pohon keputusan dan sering dikarakteristikkan sebagai pengklasifikasi statistik. C4.5 merupakan pengembangan dari algoritma ID3 yang memerlukan entropi informasi, atribut kontinu dan diskret, atribut kategorial dan numerik, dan missing values. (Lukhayu Pritalia, 2018).

Berikut beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 yaitu (Monalisa & Hadi, 2020) :

1. Mempersiapkan data training. Data training umumnya dari data histori yang sudah terjadi sebelumnya dan telah dikelompokkan ke dalam kelas-kelas tertentu
2. Memilih akar dari pohon . Akar akan didapatkan dari atibut yang terpilih dengan cara menghitung nilai Gain dari masing-masing atribut, nilai Gain tertinggi akan menjadi akar pertama. Sebelum menghitung nilai Gain dari atribut, terlebih dahulu menghitung nilai entropy yakni:

$$\boldsymbol{entropy}\ (\boldsymbol{A})\ =\ -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{|\boldsymbol{Si}|}{|\boldsymbol{S}|} \boldsymbol{log_2} \frac{|\boldsymbol{Si}|}{|\boldsymbol{S}|}$$

   Keterangan :
   - S : Himpunan kasus
   - A : Atribut
   - N : Jumlah partisi atribut A
   - |Si|: Jumlah kasus pada partisi ke i
   - |S|: Jumlah kasus dalam S

3. Kemudian hitung Gain :

$$\boldsymbol{Gain(S,A) = entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)}$$

   Keterangan :
   S : himpunan kasus
   A : atribut
   n : jumlah atribut A
   |si| : jumlah kasus pada partisi ke-i
   |s| : jumlah kasus dalam S
4. Ulangi langkah ke-2 hingga semua tupel terpartisi
5. Proses partisi pohon keputusan akan berhenti saat :
   a. Semua tupel dalma node N mendapat kelas yang sama
   b. Tidak ada atribut di dalam tupel yang dipartisi lagi.
   c. Tidak ada tupel di dalam cabang yang kosong

2.3. Confusion Matrix

Confusion Matrix terdiri dari informasi tentang kelas sebenarnya dan kelas prediksi dari suatu proses klasifikasi. Umumnya confusion matrix membedakan hasil klasifikasi yang dijalankan oleh suatu sistem dengan hasil klasifikasi yang sebenarnya.

2.4. *Python*

Python merupakan salah satu bahasa pemrograman yang dinamis dan mempunyai sistem manajemen memori yang otomatis seperti bahasa pemrograman lainnya. Python biasanya digunakan melalui script atau kode-kode, meskipun bahasa pemrograman ini lebih banyak dimanfaatkan untuk yang umumnya tidak banyak menggunakan script. Ada banyak alasan mengapa Python sukses dan berkembang. Python memiliki sintaksis yang jauh lebih ringkas dari bahasa-bahasa pemrograman populer yang ada saat ini seperti Java, C, dan C++. Oleh karena itu, python jauh lebih mudah untuk dipelajari.

2.5. Pohon Keputusan

Pohon keputusan yaitu suatu kerangka yang berfungsi untuk memecahkan berbagai kumpulan data yang besar untuk diubah menjadi berbagai himpunan record yang lebih kecil yakni dilakukan dengan mengimplementasikan sederetan aturan keputusan. Data dalam Pohon keputusan biasanya dalam bentuk tabel dengan atribut dan record. Atribut yaitu memiliki fungsi sebagai kriteria dalam pembentukan Pohon. Proses pada pohon keputusan yaitu mengubah bentuk data (tabel) menjadi model pohon, kemudian mengubah model pohon menjadi rule, dan menyederhanakan rule.

2.6. Data yang di perlukan

Data yang digunakan pada penelitian ini adalah data siswa SMA Negeri 11 Pekanbaru Kelas X

NO	Bahasa Indonesia	Matematika	IPA	IPS	Bahasa Inggris	Jurusan
1	B	B	A	B	B	MIPA
2	C	C	B	C	C	IPS
3	C	C	C	C	C	MIPA
4	D	C	C	B	D	MIPA
5	B	A	B	B	B	MIPA
6	B	C	C	B	C	MIPA
7	B	B	B	B	B	IPS
8	B	B	B	B	B	MIPA
9	B	B	B	B	B	IPS

3. HASIL DAN PEMBAHASAN

Hal pertama yang dilakukan untuk melakukan penerapan Algoritma C4.5 dan Naïve Bayes adalah dengan mengimport Library yang digunakan yaitu *Pandas*, *Numpy*, *Matplotib*, *Seaborn*, *Sklearn*, *Standartcaler*, *Naïve Bayes,* dan *C4.*5 Kemudian memasukkan data yang digunakan

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB

from sklearn import tree
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
import sklearn.metrics as metrics
from sklearn.tree import export_text
```

Setelah mengimport *Library* yang digunakan kemudian melakukan pengujian metode *naïve bayes* dan *C4.5* dengan *Split* validation 80% dan menampilkan hasil dari pengujian metode kedua metode tersebut.

```
[50] print('Akurasi Algoritma C4.5 Untuk Data Testing Adalah :','{:.3f}'.format(metrics.accuracy_score(y_test,ypred
     print(classification_report(y_test,ypred_test_c45,))

Akurasi Algoritma C4.5 Untuk Data Testing Adalah : 0.867
              precision    recall  f1-score   support

         IPS       1.00      0.83      0.90        23
        MIPA       0.64      1.00      0.78         7

    accuracy                           0.87        30
   macro avg       0.82      0.91      0.84        30
weighted avg       0.92      0.87      0.88        30
```

```
[68] print('Akurasi Algoritma Naive Bayes Untuk Data Testing Adalah :','{:.3f}'.format(metrics.accuracy_score(y_tes
     print(classification_report(y_test,y_pred_testNB,))

Akurasi Algoritma Naive Bayes Untuk Data Testing Adalah : 0.767
              precision    recall  f1-score   support

         IPS       0.71      0.86      0.77        14
        MIPA       0.85      0.69      0.76        16

    accuracy                           0.77        30
   macro avg       0.78      0.77      0.77        30
weighted avg       0.78      0.77      0.77        30
```

Berdasarkan hasil prediksi dari pengujian algoritma C4.5 dan Naive Bayes dengan nilai Data training sebesar 80% dan Data Testing Sebesar 20% dapat diketahui bahwa tingkat akurasi algoritma C4.5 memiliki akurasi sebesar 86,7%, Precision 83%, dan Recall 100% sedangkan Naive Bayes memiliki akurasi sebesar 76,7%, Precision 85%, Recall 71%. berdasarkan perbandingan Algoritma C4.5 dan Naive Bayes Dapat disimpulkan bahwa Algoritma C4.5 unggul dibandingkan Naive Bayes dengan tingkat akurasi mencapai 86,7% Precision 83% dan Recall 100%

4. KESIMPULAN

Berdasarkan hasil perbandingan dari Algoritma C4.5 dan Naive Bayes dalam penentuan penjurusan dapat disimpulkan yakni, Berdasarkan hasil uji kinerja Algoritma C4.5 dan Naive bayes dalam penentuan penjurusan siswa dengan Split Validation 80%, untuk Algoritma C4.5 memiliki nilai akurasi sebesar 86,% dengan nilai Recall sebesar 100% dan Precision sebesar 83%. Dan untuk Algoritma Naive Bayes memiliki tingkat akurasi sebesar 76% dengan nilai Recall sebesar 71% dan nilai Precision sebesar 8%. Berdasarkan hasil dari pengujian kedua Algoritma tersebut, maka dapat disimpulkan bahwa Algoritma C4.5 lebih unggul dibandingkan dengan Algoritma Naive bayes, dengan tingkat akurasi Algoritma C4.5 sebesar 86%.

UCAPAN TERIMAKASIH

Jurnal ini diselesaikan dan dipublikasikan oleh penulis atas bantuan dari berbagai pihak. Oleh karena itu penulis mengucapkan terima kasih kepada Ketua Jurusan Teknik Informatika, Universitas Lancang Kuning dan SMA Negeri 11 Pekanbaru. Terima kasih kepada dosen pembimbing saya yang telah memberikan arahan dalam penyelesaian jurnal ini.

DAFTAR PUSTAKA

[1] Anam, C., & Santoso, H. B. (2018). Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa. Energy - Jurnal Ilmiah Ilmu-Ilmu Teknik, 8(1), 13–19.

[2] Fatmawati, K. R. & A. (2019). Jurnal Ilmiah Komputer dan Informatika (KOMPUTA). 8(1).

[3] Hikmatulloh, H., Wintana, D., & Susilawati, S. (2020). Sistem Pakar Analisa Kerusakan Sepeda Motor Matic Dengan Metode Dempster Shafer Dan Pemrograman Python. Klik - Kumpulan Jurnal Ilmu Komputer, 7(1)

[4] Kadafi, A. R. (2018). Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA. Jurnal ELTIKOM, 2(2), 67–77. https://doi.org/10.31961/eltikom.v2i2.86

[5] Lukhayu Pritalia, G. (2018). Penerapan Algoritma C4.5 untuk Penentuan Ketersediaan Barang E-commerce. Indonesian Journal of Information

[6] Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Edik Informatika, 2(2), 213–219. https://doi.org/10.22202/ei.2016.v2i2.1465

[7] Monalisa, S. M., & Hadi, F. (2020). Algoritma C4.5 dalam Penentuan Jurusan Siswa Baru. Ultimatics : Jurnal Teknik Informatika, 12(2), 108–113. https://doi.org/10.31937/ti.v12i2.1838

[8] Putri, N. E., Nirwana, H., & Syahniar, S. (2019). Hubungan kondisi lingkungan keluarga dengan hasil belajar siswa sekolah menengah atas. JPGI (Jurnal Penelitian Guru Indonesia), 3(2), 98. https://doi.org/10.29210/02268jpgi0005

[9] Sibarani, A. J. P. (2020). Implementasi Data Mining Menggunakan Algoritma Apriori Untuk Meningkatkan Pola Penjualan Obat. JATISI (Jurnal Teknik Informatika Dan Sistem Informasi), 7(2), 262–276.

[10] Wanto, A., & Windarto, A. P. (2017). Analisis Prediksi Indeks Harga Konsumen Berdasarkan Kelompok Kesehatan Dengan Menggunakan Metode Backpropagation. Jurnal & Penelitian Teknik Informatika Sinkron, 2(2), 37–43. https://zenodo.org/record/1009223#.Wd7norlTbhQ