PENERAPAN ALGORITMA *K-MODES* DAN *C4.5* UNTUK PREDIKSI PEMILIHAN JURUSAN DI UNIVERSITAS PERADABAN PADA SISWA SMA (Studi Kasus: SMA Islam Ta'allumul Huda Bumiayu)

Afif Fauzi[1], Nurul Mega Saraswati[2], Rito Cipta Sigitta Hariyono[3]
[1]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Peradaban,
[2]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Peradaban,
[3]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Peradaban,
[1]afiffauzi2408@gmail.com, nurulmega@peradaban.ac.id, ritocipta@peradaban.ac.id
Jl. Raya Pagojengan KM 03 Paguyangan Brebes

Kata Kunci:

Prediksi,
Pemilihan jurusan,
Algoritma K-Modes,
Algoritma C 4.5,
confusion matrix.

Abstrak

Banyak siswa kelas XII saat ini yang memiliki kecenderungan bahwa mereka belum mengetahui secara pasti minat dan bakatnya serta kelak akan memilih jurusan kuliah dibidang apa selepas masa SMA-nya nanti. Beberapa kasus dalam kurang tepatnya memilih jurusan kuliah yang kurang sesuai dengan kemampuan atau keahlian, maupun minat dan bakat mahasiswa dapat mempengaruhi proses pembelajaran nantinya dibangku perkuliahan. Hal tersebut akan mempengaruhi perkembangan mahasiswa tersebut di dalam suatu jurusan kuliah yang dipilihnya. Untuk mengatasi permasalahan tersebut maka diperlukan suatu metode yang dapat memberikan prediksi jurusan pada siswa. Penulis bermaksud untuk mengimplementasikan metode dari Data Mining yaitu Algoritma K-modes dan Algoritma C4.5 sebagai metode untuk menggali potensi siswa dalam memilih jurusan berdasarkan nilai akademik dari siswa tersebut. Berdasarkan pengujian menggunakan confusion matrix terhadap 158 record data nilai siswa-siswi kelas XII SMA Islam Taallumul Huda Bumiayu angkatan 2018/2019 diperoleh akurasi sebesar 87,50% dengan predikat good classification

Abstract:

Prediction,
College,
Majors,
K-Modes algorithm,
C 4.5 algorithm,
confusion matrix.

Abstract

Many students of the current class XII have the tendency that they do not know exactly the interest and talents and will someday choose the school majors in what after his high school period. Some cases in less precisely choose a course that is lacking in accordance with the skills or expertise, as well as the interests and talents of students can influence the learning process to be built in the course. This will affect the student's progress in a course of study. To overcome the problem, a method that can be predicted to give students a major prediction. The author intends to implement the method of Data Mining, which is K-Modes algorithm and C 4.5 algorithm as a method to explore the potential of students in selecting majors based on the academic value of the student. Based on the testing using confusion matrix of 158 records of the value data of SMA Islam Ta'allumul Huda Bumiayu students period 2018/2019 obtained accuracy of 87.50% with good classification predicate.

Pendahuluan

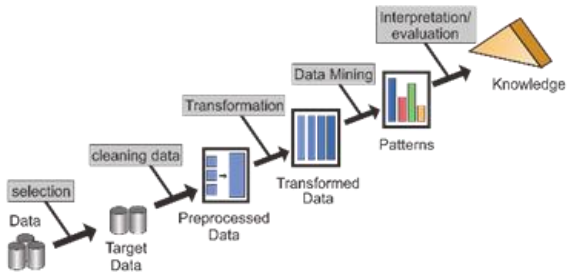
Instansi pendidikan merupakan salah satu contoh dari adanya perkembangan sebuah teknologi. Dalam hal ini, siswa SMA Islam Ta'allumul Huda Bumiayu yang memilih jurusan pada tingkat satuan pendidikan S1 perguruan tinggi yang masih ragu cenderung tinggi. Terutama pada perguruan tinggi swasta yaitu Universitas Peradaban Bumiayu dengan 3 jurusan unggulan yaitu Manajemen, Teknik Informatika dan PGSD. Untuk membantu siswa dalam meminimalisir salah jurusan lebih awal adalah dengan memanfaatkan perkembangan teknologi tersebut seperti penggunaan Data Mining pada analisis

tersebut. penelitian yang dilakukan Kusrini, Data Mining merupakan suatu proses menemukan hubungan yang berarti, pola dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan, dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika[1]. Algoritma K-modes termasuk dalam salah satu algoritma clustering, Dalam pengelompokan datanya sangat cocok untuk memproses data yang tidak memiliki label data sekalipun, K-modes merupakan pengembangan dari algoritma K-means dengan kelebihan tidak hanya bekerja baik pada tipe data numerik namun juga dapat untuk fitur kategorikal[2]. pada penelitian yang berjudul "Perbandingan kinerja Algoritma C4.5 dan Naïve Bayes untuk klasifikasi penerima beasiswa", penelitian tersebut menghasilkan tingkat akurasi dari model algoritma C4.5 yakni sebesar 96,40% sedangkan model algoritma Naïve Bayes menghasilkan tingkat akurasi sebesar 95,11% [3].
Sesuai dengan penelitian terdahulu yang telah dijelaskan di atas maka dapat disimpulkan bahwa Algoritma C4.5 memiliki performa yang lebih baik dari algoritma klasifikasi lainnya dengan hasil tingkat akurasi yang lebih besar sehingga penulis bermaksud membuat sistem prediksi untuk calon mahasiswa yang akan melanjutkan ke Perguruan Tinggi terutama Perguruan Tinggi Swasta Universitas Peradaban Bumiayu dengan pendekatan Data Mining menggunakan Algoritma K-Modes dan C4.5.
Learning data set yang di gunakan adalah data ujian nasional dan ujian akhir sekolah serta nilai raport siswa SMA Islam Ta'allumul Huda Bumiayu. Sistem yang akan dibangun hanya menginformasikan prediksi 3 jurusan unggulan pada Universitas Peradaban yaitu Manajemen Teknik Informatika dan PGSD dengan menggunakan model webview yang dapat berjalan pada semua platform android.
Tujuan dari Penelitian yang ingin dicapai dalam penelitian ini adalah mengetahui seberapa besar tingkat akurasi dari algoritma K-Modes dan algoritma C4.5 dalam memprediksi jurusan kuliah berdasarkan nilai akademik siswa SMA Islam Ta'allumul Huda Bumiayu.
potensi yang dapat di peroleh pada penelitian ini Seperti memberikan refrensi seputar prediksi potensi siswa terhadap jurusan yang ada di Universitas Peradaban yang dapat di akses secara real time kapan pun dan dimanapun, meminimalisir adanya stigma salah jurusan terhadap mahasiswa di Universitas Peradaban, dan lain sebagainya.

Landasan Teori

1. Data Mining

Data Mining terdiri dari algoritma inti yang memungkinkan seseorang untuk mendapatkan wawasan dan pengetahuan mendasar dari data yang sangat besar. Data Mining adalah bagian dari proses penemuan pengetahuan yang lebih besar, yang mencakup tugas pra-pemrosesan seperti ekstraksi, pembersihan, penggabungan, pengurangan data dan konstruksi fitur, serta langkah-langkah pasca-pemrosesan seperti interpretasi pola dan model [4]. Di dalam Data Mining terdapat beberapa pengelompokan antara lain deskripsi, estimasi, prediksi, klasifikasi, pengklasteran dan asosiasi. Tahapan-tahapan pada Data Mining dalam Proses penggalian informasi pada basis data yang besar yaitu dengan menerapkan KDD (knowledge data discovery).



Gambar 1. Tahapan KDD dalam Data Mining

Adapun proses pada knowledge data discovery (KDD) secara garis besar dijelaskan sebagai berikut ;

a) Data selection

Seleksi data dari sekumpulan data operasional sebelum tahap penggalian informasi dalam KDD dimulai.

b) Pre-processing/cleaning

Sebelum proses Data Mining dilaksanakan,Proses pembersihan pada data yang menjadi fokus KDD dengan tujuan untuk membersihkan data dari duplikasi, inkonsisten data, dan tipografi (kesalahan cetak).

c) Transformation

Coding adalah transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses Data Mining.

d) Data Mining

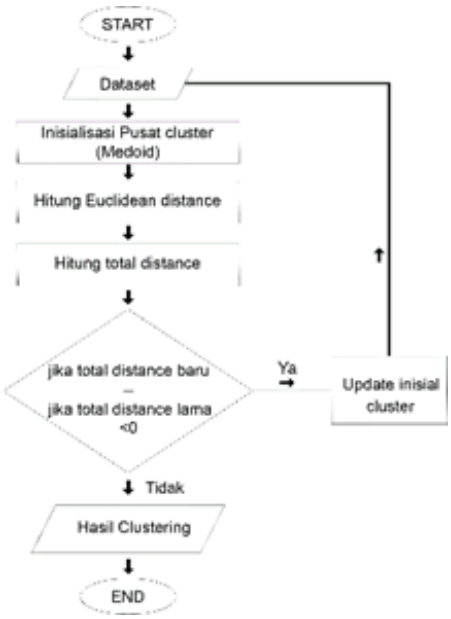
Proses pencarian pola atau informasi pada suatu data menggunakan teknik atau metode tertentu, pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

e) Interpretation / Evaluation

Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelum nya [5].

2. Algoritma K-Modes

Algoritma K-Modes merupakan pengembangan dari Algoritma K-Means dalam hal data yang di analisa dengan fitur berupa kategorikal karena merupakan pengembangan dari Algoritma K-Means[2], di dalam cara kerjanya pun mengadopsi dari Algoritma K-Means dengan melakukan modifikasi sebagai berikut :



Gambar 2. Flowchart Algoritma K-Modes

Berdasarkan gambar 2 diatas maka Berikut merupakan langkah-langkah maupun urutan kerja algoritma K-Modes :

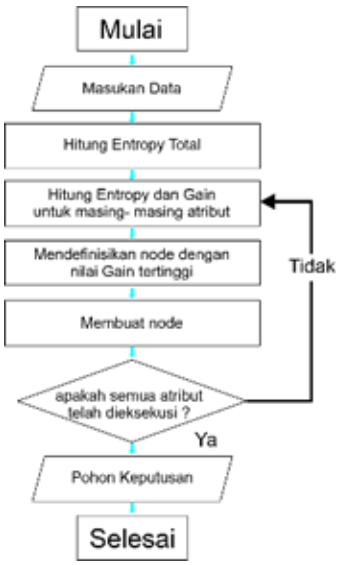
a) Pilih K data sebagai inisialisasi centroid (modus), satu untuk setiap Cluster.
b) Alokasikan data (objek) ke Cluster terdekat dengan titik pusat (modus) menggunakan ukuran jarak Euclidian Distance dengan persamaan 1.

$$d_{ij} = \sqrt{\sum_{a=1}^{p} = (x_{ia} - x_{ja})^2} = \sqrt{(x_i - x_j)'(x_i - x_j)} \dots\dots\dots\dots (1)$$

c) Pilih secara acak objek pada masing-masing Cluster sebagai medoid baru .
d) Hitung jarak setiap objek yang berada pada masing-masing Cluster dengan medoid baru.
e) Hitung total simpangan (S) dengan menghitung nilai total distance baru – total distance lama. Jika S < 0 maka tukar objek dengan data Cluster untuk membentuk medoid baru.
f) Ulangi langkah ke 3 - 5 sampai tidak terjadi perubahan medoid [6]

3. Algoritma C4.5

Algoritma Data Mining C4.5 adalah salah satu algoritma yang digunakan untuk melakukan klasifikasi,segmentasi atau pengelompokan dan bersifat prediktif, dengan algoritma ini mesin (komputer) di berikan sekelompok data utuk di pelaari atau sering di sebut sebagai learning dataset, dasar dari algoritma C4.5 itu sendiri ialah pembentukan pohon keputusan (Decision tree) [7].
Secara umum untuk langkah-langkah dalam membangun sebuah pohon keputusan menggunakan algoritma C4.5 adalah sebagai berikut :



Gambar 3. Flowchart Algoritma C4.5

Berdasarkan gambar 3 maka tahapan-tahapan dari algoritma C4.5 adalah sebagai berikut :

a) Hitung nilai Entropy.
b) Hitung Nilai Gain untuk masing-masing atribut.
c) Pilih akar (root) berdasarkan atribut yang memiliki nilai Gain tertinggi sedangkan atribut yang memiliki nilai Gain lebih rendah dari akar (root) di pilih sebagai cabang.
d) Hitung kembali nilai Gain dari masing-masing atribut dengan tidak mengikutsertakan atribut yang sebelumnya sudah terpilih menjadi akar (root).
e) Atribut dengan gain tertinggi berikutnya dipilih menjadi cabang
f) Ulangi tahapan ke-4 dan ke-5 sampai semua atribut yang tersisa menghasilkan nilai Gain = 0 [8].

untuk menentukan sebuah atribut sebagai root node atau akar adalah berdasarkan dengan nilai Gain tertinggi pada keseluruhan atribut, untuk mendapatkan nilai Gain caranya adalah dengan melakukan perhitungan menggunakan rumus persamaan 2.

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

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

Sebelum mendapatkan nilai Gain tertinggi pada masing-masing atribut terdapat satu langkah perhitungan lagi yaitu mencari nilai Entropy, Berikut ini adalah perhitungan nilai entropy berdasarkan rumus persamaan 3 berikut :

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2 pi \quad \cdots\cdots\cdots\cdots\cdots (3)$$

Dengan :
S : Himpunan kasus
n : Jumlah partisi dalam dalam S
Pi : proporsi dari Si terhadap S

Setelah mengetahui entrophy total maka langkah selanjutnya adalah menghitung entrophy dari masing-masing nilai pada suatu kasus, setelah di ketahui entropy dari masing-masing kasus pada keseluruhan atribut maka selanjutnya lakukan penghitungan gain pada masing-masing atribut[9].

4. Confusion Matrix

Confusion matrix adalah suatu metode yang biasa digunakan untuk menghitung akurasi pada konsep Data Mining, Confusion matrix digambarkan pada sebuah tabel yang terdiri atas banyaknya data uji yang diklasifikasikan benar (positive) dan banyaknya data uji yang salah (negative)[10]. Tabel confusion matrix dapat dilihat pada tabel 2.1

Tabel 2.1 Confusion Matrix

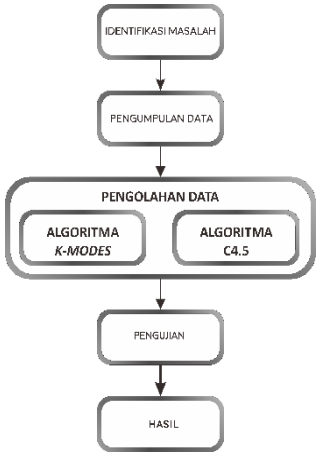
Correct Classification	*Classified as*	
	Preditected positive	*Preditected negative*
Actual positive	*True positive*	*False negative*
Actual negative	*False positive*	*True negative*

Berdasarkan tabel confusion matrix diatas maka dapat di jelaskan bahwa :

a) True positive merupakan jumlah record data positif yang di klasifikasikan sebagai nilai positif
b) False positive merupakan jumlah record data negatif yang di klasifikasikan sebagai nilai positif
c) False negative merupakan jumlah record data positif yang di klasifikasikan sebagai nilai negatif
d) True negative merupakan jumlah record data negatif yang di klasifikasikan sebagai nilai negative

Metode Penelitian

Metode dari penelitian ini adalah metode eksperimen, dan adapun tahapan-tahapan dari penelitian ini dapat dilihat pada gambar



Gambar 4. Tahapan Penelitian

1. Identifikasi Masalah
   Pada tahap ini, mengidentifikasikan permasalahan yang muncul dan ditempuh dengan melakukan kegiatan survei secara langsung pada SMA Islam Ta'allumul Huda Bumiayu, survei tersebut di

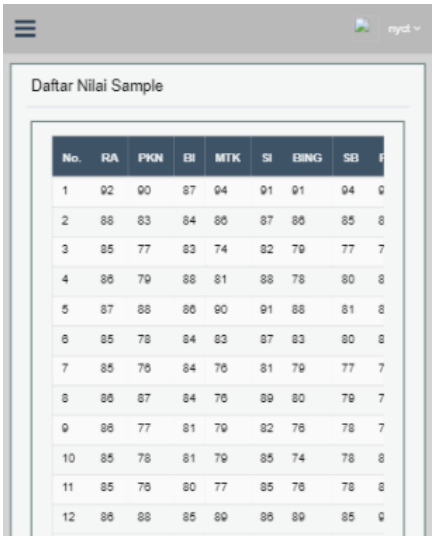
laksanakan untuk mengetahui permasalahan yang ada, dari kegiatan survei tersebut menghasilkan sebuah informasi bahwa hasil dari belajar mereka selama di sekolah termasuk nilai akademik masih belum bisa untuk menentukan pilihan mereka didalam memilih jurusan pada jenjang perkuliahan.

2. Pengumpulan Data
   Mengumpulkan data-data yang berkaitan dengan penelitian dalam hal ini penelitian tentang hal-hal yang mempengaruhi siswa di dalam masuk suatu jurusan pada jenjang perguruan tinggi, data yang telah diperoleh dan yang akan digunakan dalam penelitian ini adalah data nilai siswa kelas XII angkatan 2018/2019 serta data jurusan yang ada di SMA Islam Ta'allumul Huda Bumiayu
3. Pengolahan Data
   Pada tahapan ini, setelah data data terkumpul selanjutnya Pengolahan data dilakukan dengan menerapkan KDD (*knowledge data discovery*).
4. Pengujian
   dari *dataset* yang telah ada kemudian di olah dan di lakukan pengujian maka akan menghasilkan tingkat akurasi menggunakan *confussion matrix*, serta di hitung pengukuran besarnya tingkat kesalahan atau *error* menggunakan *root mean square error*.
5. Hasil
   Dari hasil seluruh tahapan diatas yang telah dilakukan sehingga mendapatkan hasil dari model yang di inginkan dan selanjutnya *rule* ataupun aturan yang dihasilkan.

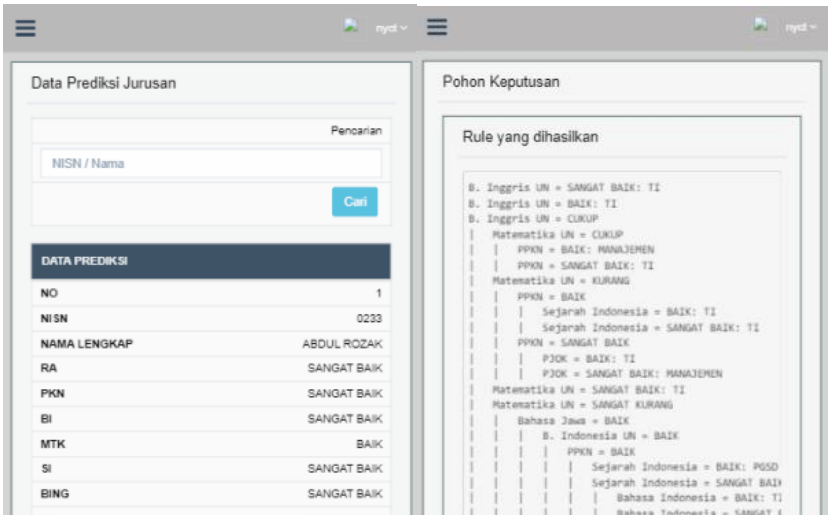
Hasil dan Pembahasan

Proses penggalian informasi atau pengolahan terhadap data yang telah terkumpul dari semua nilai siswa selanjutnya data di olah dengan menerapkan KDD (*knowledge data discovery*) untuk mencari pola maupun informasi dari suatu data tersebut, Berdasarkan penjabaran pada bab 2 tentang tahapan KDD dalam Data Mining maka langkah-langkah pada tahapan ini adalah *Data Selection, Data Cleaning, Transformasion,* proses *Data Mining* yang terdiri dari penerapan algoritma K-Modes dan Algoritma C 4.5, *Interpretation* serta Pengujian. Hasil dari penelitian ini adalah aplikasi yang dapat memprediksi jurusan kuliah siswa-siswi SMA Islam Ta'allumul Huda Bumiayu berdasarkan nilai akademik.

Gambar 5. Tampilan menu utama



Gambar 6. Tampilan daftar nilai sample



Gambar 7. Tampilan data prediksi jurusan dan pohon keputusan

Kesimpulan dan Saran

Kesimpulan

Berdasarkan hasil dari penelitian yang telah dilakukan oleh penulis, maka dapat disimpulkan bahwa jurusan kuliah siswa-siswi SMA Islam Ta'allumul Huda yang akan melanjutkan ke jenjang perguruan tinggi dapat diprediksi dan dievaluasi menggunakan teknik *data mining* menggunakan algoritma *K-modes* dan C4.5 untuk prediksi pemilihan jurusan di Universitas Peradaban pada siswa SMA (Studi kasus : SMA Islam Ta'allumul Huda Bumiayu). Pengolahan data yang telah dilakukan pada 158 *record* data dari masing-masing nilai siwa-siswi SMA Islam Ta'allumul Huda Bumiayu tahun ajaran 2018/2019 menggunakan Algoritma *K-Modes* dan Algoritma C4.5 diperoleh akurasi sebesar 87,50% dengan predikat *good classification.*

Saran

Saran yang dapat disampaikan sebagai bahan dalam pertimbangan kepada pihak-pihak yang berkepentingan dalam mengembangkan serta menyempurnakan lebih lanjut mengenai penerapan Algoritma K-modes dan C4.5 untuk prediksi pemilihan jurusan di universitas peradaban pada siswa

SMA (Studi kasus: SMA Islam Ta'allumul Huda Bumiayu) adalah sebagai berikut

1. Mengembangkan penelitian ini dengan menambahkan metode optimasi agar dapat meningkatkan nilai akurasi yang lebih besar dalam memprediksi jurusan siswa-siswi SMA Islam Ta'allumul Huda Bumiayu.
2. Diharapkan nantinya aplikasi yang telah terbentuk dapat berjalan di semua platform seperti pada sistem operasi berbasis Desktop maupun IOS (Apple).

Referensi

[1] Kusrini dan Emha Taufiq, "Proses Data Mining," Data Min., pp. 1–143, 2015.

[2] Eko Prasetyo, DATA MINING Mengolah Data Menjadi Informasi Menggunakan Matlab. yogyakarta: CV.ANDI OFFSET, 2014.

[3] Choirul Anam and Harry Budi Santoso, "Perbandingan Kinerja Algoritma C4 . 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," J. Ilm. Ilmu-Ilmu Tek., vol. 8, no. 1, pp. 13–19, 2018.

[4] Mohammed J. Zaki dan Wagner Meira Jr, DATA MINING Fundamental Concepts and Algorithms, vol. 35, no. 6. 2015.

[5] Alfannisa Annurullah Fajrin dan Algifanri Maulana, "Penerapan Data Mining Untuk Analisis Pola Pembelian Konsumen Dengan Algoritma Fp-Growth Pada Data Transaksi Penjualan Spare Part Motor," Klik - Kumpul. J. Ilmu Komput., vol. 5, no. 1, p. 27, 2018, doi: 10.20527/klik.v5i1.100.

[6] Siti Sundari, Irfan Sudahri Damanik, Agus Perdana Windarto, Heru Satria Tambunan, Jalaluddin, and A. Wanto, "Analisis K-Medoids Clustering Dalam Pengelompokkan Data Imunisasi Campak Balita di Indonesia," Pros. Semin. Nas. Ris. Inf. Sci., vol. 1, no. September, p. 687, 2019, doi: 10.30645/senaris.v1i0.75.

[7] Nurbaiti, "Kriteria Nasabah Non Muslim Menabung (Penggalian Data Menggunakan Klasifikasi Algoritma C4.5 Studi Kasus Di Pt.Bank Bri Syariah Kantor Cabang Medan)," J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2019, doi: 10.1017/CBO9781107415324.004.

[8] N. Anwar, A. Pranolo, and R. Kurnaiwan, "Grouping the community health center patients based on the disease characteristics using C4.5 decision tree," IOP Conf. Ser. Mater. Sci. Eng., vol. 403, no. 1, 2018, doi: 10.1088/1757-899X/403/1/012084.

[9] Faid Ari Prasetya, "Penerapan Algoritma C4.5 Untuk Prediksi Jurusan Siswa Sman 3 Rembang," vol. 01, pp. 1–8, 2019.

[10] M. I. D. dan D. A. M.Fadly Rahman, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)," J. Inform., vol. 11, no. 1, p. 36, 2017, doi: 10.26555/jifo.v11i1.a5452.