PENERAPAN ALGORITMA C4.5 DALAM MEMPREDIKSI KETERLAMBATAN PEMBAYARAN UANG SEKOLAH MENGGUNAKAN PYTHON

Victor Saputra Ginting, Kusrini, Emha Taufiq Luthfi

Magister Teknik Informatika, Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Kabupaten Sleman, Daerah Istimewa Yogyakarta
victor.ginting@students.amikom.ac.id, kusrini@amikom.ac.id, emhataufiqluthfi@amikom.ac.id

Abstract – Payment of school fees is an important factor in carrying out education because payment of school fees is one of the important obligations in getting an education. Research conducted by Muqorobin, 2019 with the research title "Optimizing the Naive Bayes Method with Feature Selection Gain for Predicting Late School Payment" produces an accuracy rate of 80%, then optimization is carried out using information gain and Naive Bayes produces an accuracy rate of 90%. The research conducted will be a prediction of late payment of school fees using the C4.5 algorithm and then carried out into the form of programming using Python, to produce a prediction result. Information Prediction Results obtained from Python then will try to compare the level of accuracy with the dataset that has been collected using a confusion matrix. The dataset obtained from De Britto Yogyakarta High School using Python produces an accuracy rate of 73%. This research is expected to help the school in making decisions and minimize the level of late payment of school fees.
Keywords - Prediction, Algorithm C4.5, Python, and Data Mining.

Abstrak **–** Pembayaran uang sekolah merupakan faktor yang cukup penting dalam melangsungkan pendidikan karena pembayaran uang sekolah merupakan salah satu kewajiban penting dalam mendapatkan pendidikan. Penelitian yang telah dilakukan Muqorobin, 2019 dengan judul penelitian "Optimasi Metode Naive Bayes Dengan Feature Selection Gain Untuk Memprediksi Keterlambatan Pembayaran Uang Sekolah" menghasilkan tingkat akurasi sebesar 80%, kemudian dilakukan optimasi dengan menggunakan information gain dan Naive Bayes menghasilkan tingkat akurasi sebesar 90%. Penelitian yang dilakukan akan dilakukan prediksi keterlambatan pembayaran uang sekolah dengan menggunakan algoritma C4.5 dan kemudian dilakukan implementasi kedalam bentuk pemrograman dengan menggunakan Python, sehingga menghasilkan keterangan hasil prediksi. Keterangan Hasil Prediksi yang didapatkan dari Python, kemudian akan coba dilakukan perbandingan tingkat akurasi dengan dataset yang sudah dikumpulkan menggunakan confusion matrix. Dataset yang didapatkan dari Sekolah Menengah Atas De Britto Yogyakarta menggunakan Python menghasilkan tingkat akurasi sebesar 73%. Penelitian ini diharapkan dapat membantu pihak sekolah dalam mengambil keputusan dan meminimalisir tingkat keterlambatan pembayaran uang sekolah.
Kata kunci - Prediksi, Algoritma C4.5, Python dan Data Mining.

I. PENDAHULUAN

Pendidikan merupakan salah satu kebutuhan penting bagi setiap orang. Salah satu kewajiban untuk mendapatkan pendidikan di sekolah adalah melakukan pembayaran Sumbangan Pembangunan Pendidikan (SPP) Sekolah. Tidak dapat dipungkiri, SPP Sekolah merupakan salah satu faktor penting yang digunakan untuk mengalokasi biaya pembangunan sekolah, biaya untuk guru, karyawan, dan lain-lain. Biaya SPP Sekolah ini umumnya diterapkan oleh sekolah swasta yang dibebankan kepada siswanya, berbeda dengan sekolah negeri, yang biaya pengelolaan sekolah masih ada bantuan dari biaya pemerintah. Namun akan menjadi masalah yang cukup besar bagi instansi sekolah apabila keterlambatan pembayaran SPP Sekolah dilakukan oleh murid. Hal ini akan menjadi penghambat dalam mendapatkan pendidikan, khususnya di instansi sekolah swasta. Berdasarkan pemaparan diatas, perlu adanya sebuah penelitian untuk memprediksi keterlambatan pembayaran SPP Sekolah yang dilakukan murid. Penelitian yang berjudul "Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain untuk Prediksi Keterlambatan Pembayaran SPP

Sekolah" [1] dengan objek penelitian di SMK Al-Islam Surakarta menghasilkan keterangan prediksi yang didapatkan melalui metode naive bayes, kemudian keterangan hasil dan keterangan hasil prediksi dilakukan perbandingan dan menghasilkan tingkat akurasi sebesar 90%. Penelitian yang direncakan menggunakan Dataset dari Sekolah Menengah Atas Kolese De Britto Yogyakarta. Dataset diambil secara acak dengan jumlah sebanyak 30 dataset, dengan menggunakan atribut atau variabel seperti Penghasilan Orangtua, Tanggungan Keluarga, Pendidikan Orangtua, Umur Orangtua. Penelitian dengan judul "Analisis Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada PT. Capella Dinamik Nusantara Cabang Muka Kuning" [2] dimana dalam penelitian tersebut menghasilkan rules yang didapat dari tiap-tiap atribut menggukana *tools Data Mining* Weka.

Penelitian yang berjudul "Penerapan Algoritma C4.5 Pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika" [3] dimana penelitian tersebut menggunakan atribut seperti asal daerah, IPK, TOEFL dan Lama Studi menghasilkan tingkat presisi sebesar 63.93%, recall 60,73%, dan akurasi sebesar 60,52%.

Penelitian dengan judul "Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal" [4] dengan pengujian data latih dengan variabel yang berbeda menghasilkan tingkat akurasi sebesar 50%. Penelitian dengan judul "Komparasi Kinerja Algoritma C4.5 dan Naive Bayes Untuk Prediksi Kegiatan Penerimaan Mahasiswa Baru Studi Kasus Universitas STIKUBANK Semarang"[5] menghasilkan tingkat akurasi sebesar 87,5% menggunakan Metode Naive Bayes.

A. Data Mining

Data Mining merupakan proses yang mengerjakan satu atau lebih teknik yang menggunakan pembelajaran tentang komputer (*Machine Learning*) untuk melakukan analisis dan melakukan ekstrasi pengetahuan (*Knowledge*) secara otomatis. Data Mining berisi tentang trend atau pola yang akan dilakukan dalam database dengan skala yang besar untuk membantu dalam pengambilan keputusan di masa depan. Pola-pola tersebut akan dikenali oleh perangkat tertentu yang akan memberikan sebuah analisa data dan berwawasan yang selanjutnya dapat diteliti, sehingga akan memberikan pendukung keputusan melalui sebuah perangkat. [6]

B. Algoritma C4.5

Algoritma C4.5 ialah algoritma yang sudah cukup terkenal dan merupakan salah satu algoritma dalam Data Mining. Salah satu kegunaanya adalah untuk membentuk sebuah pohon keputusan (*Decision Tree*). Pohon Keputusan sendiri merupakan metode klasifikasi dan prediksi yang digunakan untuk mencari data dan menemukan hubungan yang tersimpan dari variabel atau atribut yang digunakan dan sebuah variabel target yang biasa disebut class atau label. Algoritma yang sering digunakan untuk membuat sebuah pohon keputusan adalah ID3, algoritma C4.5, dan CART.

Algoritma C4.5 sendiri merupakan hasil dari pengembangan algoritma ID3, dimana proses pada algoritma C4.5 akan membentuk sebuah pohon keputusan, mengubah model pohon menjadi rules, kemudian rules tersebut disederhanakan kembali. Pembuatan decision tree dengan algoritma C4.5 digunakan untuk membangun sebuah pohon keputusan yang dimulai dari pemilihan variabel atau atribut sebagai akar, membangun cabang untuk tiap nilai, membagi kasus dalam cabang kemudian melakukan pengulangan proses untuk setiap cabang sampai seluruh kasus pada cabang mempunyai kelas yang sama [7]

C. Python

Python merupakan sebuah bahasa pemrograman yang cukup terkenal yang memiliki banyak manfaat untuk mendukung pemrograman yang berorientasi objek dan dapat berjalan diberbagai macam platform sistem operasi, seperti PCs, Macintosh, UNIX. Beberapa kelebihan dari bahasa pemrograman python diantara lain :

1. Pengembangan program dilakukan dengan cepat dan coding yang lebih sedikit
2. Mendukung multi *platform*
3. Memiliki sistem pengelolaan memori yang otomatis
4. Python bersifat *Object Oriented Programming* (OOD)

D. Machine Learning

Python merupakan sebuah bahasa pemrograman yang cukup terkenal yang memiliki banyak manfaat untuk mendukung pemrograman yang berorientasi objek dan dapat berjalan diberbagai macam platform sistem operasi, seperti PCs, Macintosh, UNIX. Beberapa kelebihan dari bahasa pemrograman python diantara lain :

5. Pengembangan program dilakukan dengan cepat dan coding yang lebih sedikit
6. Mendukung multi *platform*
7. Memiliki sistem pengelolaan memori yang otomatis
8. Python bersifat *Object Oriented Programming* (OOD)

Bahasa Pemrograman Python memiliki efisiensi yang cukup tinggi, permrograman berorientasi dengan objek yang lebih sederhana namun cukup efektif, dan dapat digabungkan dengan bahasa pemrograman yang lain [8]

Bahasa Pemrograman Python memiliki efisiensi yang cukup tinggi, permrograman yang berorientasi

dengan objek lebih sederhana, namun efektif dan dapat digabungkan dengan bahasa pemrograman lain nya [8]

E. Confusion Matrix

Confusion Matrix merupakan sebuah metode yang berguna untuk melakukan perbandingan tikat akurasi. Evaluasi yang dilakukan dalam Confusion Matrix akan menghasilkan nilai akurasi, presisi, dan recall. Akurasi dalam klasifikasi adalah persentase dalam ketepatan pada record data yang telah diklasifikasi dengan benar setelah dilakukan pengujian pada hasil klasifikasi [9].

Presisi atau confidence merupakan proporsi kasus yang diprediksi positif, data yang sebenarnya juga positif. Recall atau sensitivity merupakan proporsi kasus positif yang diprediksi secara benar [10].

Pengukuran akurasi dilakukan dengan metode pengujian confusion matrix dapat dilihat pada Tabel 1.

Tabel 1. Confusion Matrix

Correct Classification	Classification	
	Positif	Negatif
Positif	TP	TN
Negatif	FP	FN

II. METODE PENELITIAN

Penelitian ini dilakukan dengan penelitian induktif yakni mencari dan mengumpulkan data yang didapat dari Sekolah Menengah Atas Kolese De Britto Yogyakarta. Data yang diperoleh didapatkan dari bagian administrasi sekolah dan dipilih sebanyak 30 data secara acak.

1. Mengumpulkan Data
   Penelitian dilakukan dengan mengumpulkan data, yang terdiri dari penelitian perpustakaan (*library research*) dan penelitian lapangan (*field research*)
2. *Cleansing Data*
   Pada tahap ini dilakukan *Cleansing Data*. Yakni mengubah, mengoreksi atau menghapus data-data yang dianggap tidak perlu dalam penelitian atau data yang dianggap tidak lengkap (*Missing Values*)
3. *Data Integration*
   Setelah tahap *Cleansing Data* dilakukan, selanjutnya dilakukan penggabungan data dan menentukan variabel atau atribut yang selanjutnya akan dilakukan untuk memprediksi
4. *Data Transformation*
   Pada bagian ini disijakan *Transformation Data*, yakni mensinambungkan atribut yang digunakan dan mengubahnya kedalam bentuk konsep hierarki, yakni mengganti konsep level rendah seperti numerik untuk usia, dan diubah ke konsep yang lebih tinggi seperti muda, dewasa, dan manula.
5. Implementasi Pemrograman Pyhton
   Pada tahap ini dilakukan implementasi kedalam bentuk pemrograman dengan menggunakan bahasa pemrograman python. Sebelum melakukan pemrograman, terlebih dahulu melakukan *import library* yang dibutuhkan untuk memprediksi. Library yang dipakai dalam pengujian ini seperti Chefboost, pandas, numpy dan sklearn. Untuk alur penelitian dapat dilihat pada Gambar 1.

Gambar 1. Alur penelitian

III. HASIL DAN PEMBAHASAN

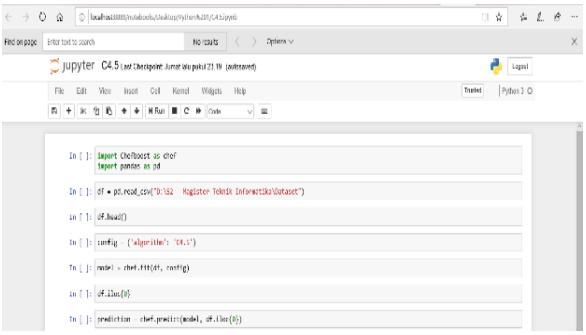
Dari dataset yang sudah diambil dari Sekolah Menengah Atas Kolese De Britto Yogyakarta meliputi 30 data yang diambil secara random dengan atribut yang digunakan adalah Pendapatang Orangtua, Tanggungan Keluarga, Pendidikan Orangtua, Umur Orang tua dapat dilihat pada Tabel.2 berikut :

Tabel.2

No		Penghasilan Orang Tua	Tanggungan Keluarga	Pendidikan Ayah	Umur Ayah	Pendidikan Ibu	Umur Ibu	Ket
1	Christine	2 - 4 Juta	Cukup	SD	Lansia Awal	SMP	Dewasa Awal	Tepat
2	Adrian	< 1 Juta	Banyak	SD	Lansia Awal	SMP	Lansia Awal	Terlambat
3	Yosafat	1 – 2 Juta	Sedikit	S1	Lansia Awal	D3	Lansia Awal	Tepat
4	Devi Siswani	< 1 Juta	Cukup	SMP	Lansia Akhir	SD	Lansia Akhir	Terlambat
5	Abdiel	1 – 2 Juta	Sedikit	SMA	Dewasa Awal	SMP	Dewasa Akhir	Terlambat
6	Abraham	< 1 Juta	Banyak	SMA	Lansia Akhir	SD	Manula	Tepat
7	Christian	1 – 2 Juta	Sedikit	D3	Dewasa Awal	SMA	Dewasa Akhir	Tepat
8	Alfredo	2- 4 Juta	Cukup	SD	Lansia Akhir	D3	Lansia Awal	Terlambat
9	Christopher	1 – 2 Juta	Sedikit	SMP	Manula	SD	Manula	Terlambat
10	Carlos	< 1 Juta	Cukup	SD	Manula	SMA	Lansia Akhir	Tepat
....								
30	Edward	>4 Juta	Banyak	S1	Dewasa Awal	D3	Dewasa Akhir	Tepat

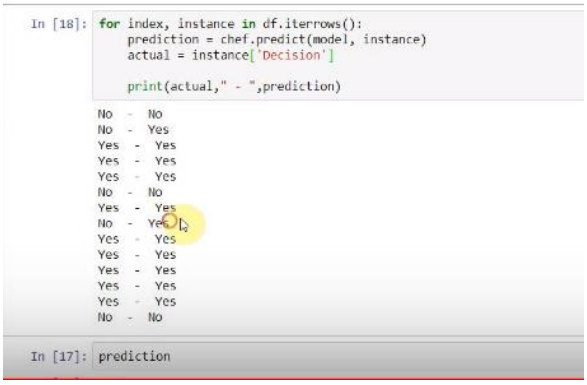
Dataset yang sudah dikumpulkan kemudian diimplementasikan kedalam bentuk pemrograman menggunakan Python. Dataset yang tersimpan dalam bentuk excel selanjutnya diubah kedalam bentuk CSV (*Comma Delimeted*) agar dapat dibaca dalam pemrograman. Dataset yang sudah dalam bentuk CSV kemudian diimport kedalam *Framework* Jupyter untuk nantinya dieksekusi untuk mendapatkan keterang hasil prediksi.

Python yang dogunakan oleh penulis adalah Python dengan versi 3.8.1. kemudian selanjutnya melakukan *import* library yang dibutuhkan dalam proses prediksi keterangan hasil dari Python. Penulis menggunakan library seperti numpy, pandas, Chefboost, dan sklearn. Proses coding dapat dilihat pada Gambar 2.



Gambar 2. Implementasi Python

Proses dataset yang sudah di import kedalam Jupyter Framework, selanjutnya menginputkan persamaan dari Algoritma C4.5 kemudian program di *run.* Hasil keterangan prediksi menggunakan python dapat dilihat pada Gambar 3.



Gambar 3. Keterangan Hasil Prediksi Python

Untuk perbandingan keterangan hasil dataset dan keterangan hasil prediksi python dari dataset yang berjumlah 30 dapat dilihat pada Tabel 3.

Tabel 3.

No	Keterangan Hasil Dataset	Keterangan Hasil Prediksi Python
1	Tepat	Tepat
2	Terlambat	Terlambat
3	Tepat	Terlambat
4	Terlambat	Terlambat
5	Terlambat	Terlambat
6	Tepat	Terlambat
7	Tepat	Tepat
8	Terlambat	Tepat
9	Terlambat	Terlambat
10	Tepat	Terlambat
	...	
30	Tepat	Tepat

Confusion Matrix

Confusion Matrix digunakan untuk melakukan perbandingan tingkat akurasi yang diperoleh dari keterangan hasil dataset dengan keterangan hasil prediksi menggunakan python. Tabel confusion matrix dapat dilihat pada Tabel 4.

Tabel 4. Pengujian Confusion Matrix

Correct Classification	Classification	
	Positif	Negatif
Positif	10	4
Negatif	4	12

Keterangan :

1. Classification Positif dengan Positif = 10 karena jumlah data positif terklasifikasi dengan benar oleh sistem
2. Classification Negatif dengan Positif = 4 karena jumlah data negatif namun terklasifikasi dengan benar oleh sistem
3. Classification Positif dengan Negatif = 4 karena jumlah data positif namun terklasifikasi salah oleh sistem
4. Classification Negatif dengan Negatif = 12 karena jumlah data negatif dan terklasifikasi salah oleh sistem
5. Hasil pengujian yang didapatkan adalah :
   Akurasi = 10+12/(10+4+4+12) * 100% = 73%
   Presisi = 10/(10+4) * 100% = 71%
   Recall = 10/(10+4) * 100% = 71%

IV. KESIMPULAN

Berdasarkan penelitian, implementasi dan pengujian dari pemaparam diatas maka dapat diambil kesimpulan sebagai berikut :

1. Penerapan Algoritma C4.5 kedalam bentuk pemrograman python untuk memprediksi keterlambatan pembayaran SPP dapat dilakukann.
2. Tingkat akurasi yang diperoleh dari keterangan hasil dataset dengan keterangan hasil prediksi python menghasilkan akurasi sebesar 73%, recall sebesar 71%, dan presisi sebesar 71%.

Saran yang dapat diberikan untuk penelitian selanjutnya yaitu :

1. Menggunakan metode Data Mining lain seperti Metode KNN, Metode CNN, Metode Naive Bayes untuk mendapatkan tingkat akurasi yang lebih tinggi.
2. Mencoba melakukan penelitian yang lebih mendalam pada atribut yang akan digunakan sehingga mengoptimalkan tingkat akurasi yang didapatkan

DAFTAR PUSTAKA

[1] M. Muqorobin, K. Kusrini, and E. T. Luthfi, "Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah," *J. Ilm. SINUS*, vol. 17, no. 1, p. 1, 2019, doi: 10.30646/sinus.v17i1.378.

[2] N. Azwanti, "Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada Pt. Capella Dinamik Nusantara Cabang Muka Kuning," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, 2018, doi: 10.30872/jim.v13i1.629.

[3] R. P. S. Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, 2018, doi: 10.23917/khif.v4i1.5975.

[4] R. H. Pambudi, B. D. Setiawan, and Indriati, "Penerapan Algoritma C4 . 5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, 2018.

[5] N. Yahya and A. Jananto, "KOMPARASI KINERJA ALGORITMA C.45 DAN NAIVE

BAYES UNTUK PREDIKSI KEGIATAN PENERIMAANMAHASISWA BARU (STUDI KASUS : UNIVERSITAS STIKUBANK SEMARANG)," *Pros. SENDI*, 2019.

[6] Turban, "Mechine Learning untuk Mengesktraksi dan Mengidentifikasi Informasi yang bermanfaat," *Mechine Learn.*, 2005.

[7] R. W. Abdullah, K. Kusrini, and E. T. Luthfi, "PREDIKSI KETERLAMBATAN PEMBAYARAN SPP SEKOLAH DENGAN METODE K-NEAREST NEIGHBOR (STUDI KASUS SMK AL-ISLAM SURAKARTA)," *Pemodelan Arsit. Sist. Inf. Perizinan Menggunakan Kerangka Kerja Togaf Adm*, 2019.

[8] J. A. A. Imam Adli, HarunMukhtar, "Perancangan dan pembuatan visual novel sejarah kh. ahmad dahlan sebagai media pembelajaran berbasis android," *RABIT (Jurnal Teknol. dan Sist. Inf. Univrab)*, 2018.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[10] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, doi: 10.1109/CVPR.2019.01150.