

ANALISIS KINERJA ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRESTASI SISWA SEKOLAH MENENGAH KEJURUAN

Astrid Noviriandini¹; Nurajijah²

^{1,2}Ilmu Komputer

^{1,2}STMIK Nusa Mandiri

^{1,2}<http://www.nusamandiri.ac.id>

¹astridnovi81@gmail.com, ²nurazizah6878@gmail.com

Abstract— *This research informs students and teachers to anticipate early in following the learning period in order to get maximum learning outcomes. The method used is C4.5 decision tree algorithm and Naïve Bayes algorithm. The purpose of this study was to compare and evaluate the decision tree model C4.5 as the selected algorithm and Naïve Bayes to find out algorithms that have higher accuracy in predicting student achievement. Learning achievement can be measured by the value of report cards. After comparison of the two algorithms, the results of the learning achievement prediction are obtained. The results showed that the Naïve Bayes algorithm had an accuracy value of 95.67% and the AUC value of 0.999 was included in Excellent Clasification, for the C4.5 algorithm the accuracy value was 90.91% and the AUC value of 0.639 was included in the state of Poor Clasification. Thus the Naïve Bayes algorithm can better predict student achievement.*

Keywords: *Student Achievement, Naïve Bayes, Decision Tree C4.5.*

Intisari—Penelitian ini menginformasikan kepada siswa dan guru untukantisipasi dini dalam mengikuti masa belajar agar mendapatkan hasil belajar yang maksimal. Metode yang digunakan adalah algoritma *decision tree* C4.5 dan algoritma *Naïve Bayes*. Tujuan dari penelitian ini adalah melakukan komparasi dan evaluasi model pohon keputusan C4.5 sebagai algoritma terpilih dan *Naïve Bayes* untuk mengetahui algoritma yang memiliki keakuratan lebih tinggi dalam memprediksi prestasi siswa. Prestasi belajar dapat diukur dari nilai *raport*. Setelah dilakukan komparasi dari kedua algoritma tersebut, maka didapatkan hasil prediksi prestasi belajar. Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* memiliki nilai akurasi sebesar 95,67% dan nilai AUC sebesar 0,999 termasuk kedalam *Excellent Clasification*, untuk algoritma C4.5 memiliki nilai akurasi sebesar 90,91% dan nilai AUC sebesar 0,639 termasuk kedalam keadaan *Poor Clasification*. Dengan demikian algoritma

Naïve Bayes dapat memprediksi prestasi belajar siswa dengan lebih baik.

Kata Kunci: *Prestasi Siswa, Naïve Bayes, Decision Tree C4.5.*

PENDAHULUAN

Pendidikan merupakan usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran peserta didik secara aktif mengembangkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri, kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya, masyarakat, bangsa dan negara. Berdasarkan UU. No. 20 Tahun 2003 mengenai Sistem Pendidikan Nasional dalam pasal 3, bahwa tujuan pendidikan nasional adalah mengembangkan potensi peserta didik agar menjadi manusia yang beriman dan bertaqwa kepada Tuhan Yang Maha Esa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri dan menjadi warga negara yang demokratis serta bertanggung jawab. Berdasarkan tujuan pendidikan tersebut maka kualitas dan manajemen pembelajaran di sekolah atau lembaga pendidikan perlu ditingkatkan. Salah satu indikator kualitas dan manajemen sekolah atau lembaga pendidikan dapat dilihat dari prestasi belajar siswa.

Sekolah Menengah Kejuruan (SMK) merupakan bentuk suatu pendidikan yang orientasinya memberi bekal siswa untuk memasuki lapangan kerja tingkat menengah dan melanjutkan pendidikan ke jenjang yang sesuai dengan kekhususannya (kejuruannya). Pada pendidikan kejuruan memberikan suatu bentuk pengembangan bakat, pendidikan dasar keterampilan dan kebiasaan-kebiasaan yang mengarah pada dunia kerja yang dipandang sebagai latihan keterampilan. Siswa akan disiapkan untuk memasuki persaingan di dunia kerja. Kegiatan pembelajaran pun tidak hanya terjadi di sekolah, namun kegiatan praktik industri di dunia kerja nyata sangat ditekankan untuk mendapatkan dan meningkatkan pengalaman bekerja di persaingan dunia kerja.

Menurut Peraturan Pemerintah No. 29 Tahun 1990 pasal 2 ayat (1) menyebutkan bahwa sekolah kejuruan bertujuan untuk meningkatkan pengetahuan siswa dalam pengembangan diri dan untuk meningkatkan kemampuan siswa sebagai anggota masyarakat. Menurut Peraturan Pemerintah No. 29 Tahun 1990 Pasal 3 ayat (2) disebutkan bahwa sekolah kejuruan bertujuan untuk menyiapkan siswa dalam memenuhi lapangan kerja, menyiapkan siswa agar mampu memiliki karir dan menyiapkan tamatan agar menjadi warga Negara yang produktif, adatif dan normatif (Susanto & sudiyatno, 2014).

Beberapa faktor yang mempengaruhi prestasi belajar siswa Sekolah Menengah Kejuruan antara lain sosial ekonomi orang tua, fasilitas belajar sekolah, motivasi, kedisiplinan siswa dan lingkungannya (Susanto & sudiyatno, 2014). Faktor sosial ekonomi orang tua yang mewujudkan pada kemampuan finansialnya. Kemampuan finansial yang berbeda-beda akan berpengaruh terhadap prestasi belajar siswa (Susanto & sudiyatno, 2014).

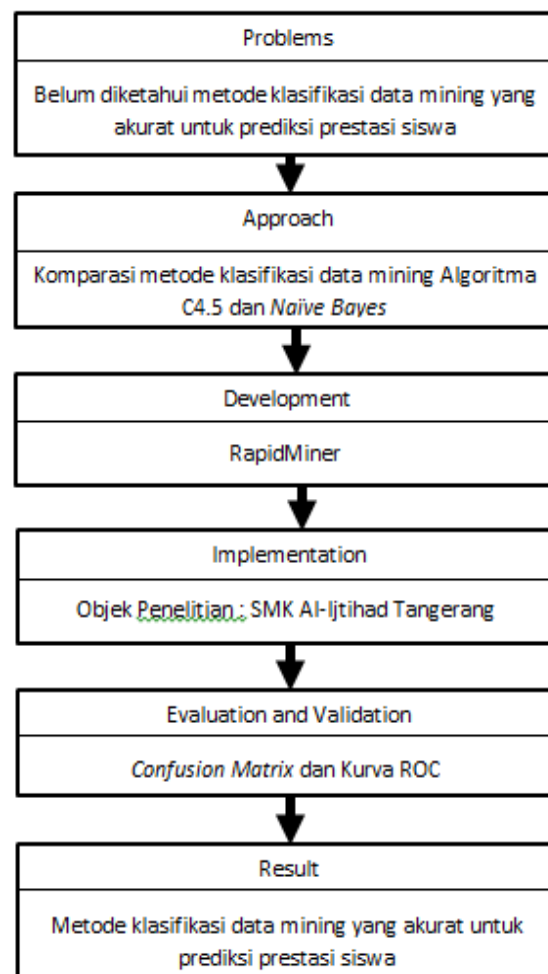
Berdasarkan Undang-Undang Sistem Pendidikan Nasional No. 20 Tahun 2003 Bab III Pasal 45 tentang sarana dan prasarana pendidikan, dinyatakan bahwa (1) setiap satuan pendidikan formal maupun non formal menyediakan sarana dan prasarana yang memenuhi keperluan pendidikan sesuai dengan pertumbuhan dan perkembangan secara fisik, kecerdasan intelektual sosial, emosional dan kejiwaan peserta didik; (2) ketentuan mengenai penyediaan sarana dan prasarana pendidikan pada semua satuan pendidikan sebagaimana dimaksud pada ayat (1) diatur lebih lanjut dengan peraturan pemerintah. Dari kedua ayat tersebut dimaksudkan agar setiap sekolah menyediakan sarana dan prasarana pembelajaran yang memadai semua keperluan pendidikan agar siswa dapat memanfaatkannya sebagai penunjang belajar siswa.

Belajar sudah menjadi kewajiban yang harus dilakukan oleh seorang siswa. Kadang siswa jenuh dan akhirnya memilih untuk tidak belajar. Untuk itulah mereka perlu diberi dorongan baik dari segala pihak. Pengaruh motivasi terhadap prestasi belajar siswa sangat besar karena itu orang-orang disekitar hendaknya memberi motivasi yang tepat. Faktor kedisiplinan merupakan suatu hal yang sangat penting. Tanpa adanya kesadaran akan keharusan melaksanakan aturan yang sudah ditentukan sebelumnya pengajaran tidak mungkin mencapai target yang maksimal. Sikap disiplin yang timbul dari kesadarannya sendiri akan dapat lebih memacu dan tahan lama dibandingkan dengan sikap disiplin yang timbul karena adanya pengawasan dari orang lain. Disiplin dapat tumbuh dan dibina melalui latihan, pendidikan atau

peneneman kebiasaan yang harus dimulai sejak dalam lingkungan keluarga. Tujuan dari penelitian ini adalah melakukan komparasi dan evaluasi model pohon keputusan C4.5 sebagai algoritma terpilih dan *Naïve Bayes* untuk mengetahui algoritma yang memiliki keakuratan lebih tinggi dalam memprediksi prestasi siswa.

BAHAN DAN METODE

Penelitian ini terdiri dari beberapa tahapan (Gambar1). Dua pendekatan utama dalam penelitian yaitu pendekatan kualitatif dan pendekatan kuantitatif. Dalam penelitian ini metode yang digunakan yaitu metode penelitian kuantitatif. Jenis penelitian yang dilakukan pada penelitian ini adalah jenis penelitian *experiment*, yaitu penelitian yang dilakukan dengan cara menguji kebenaran sebuah hipotesis dengan statistik yang melibatkan penyelidikan beberapa variabel dengan menggunakan tes tertentu dan menghubungkannya dengan masalah penelitian. Penelitian ini menggunakan data *primer* yang diperoleh dari arsip data siswa SMK Al-Ijtihad.



Gambar 1. Tahapan Penelitian

Langkah-langkah untuk melakukan penelitian (Wahyudi, Indrajit, & Muh. Fauzi, 2017) yaitu:

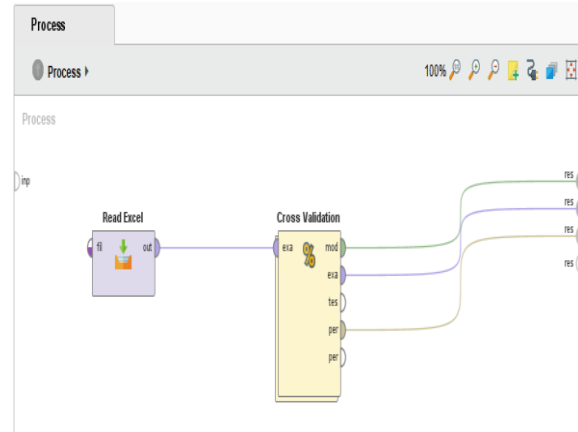
1. Pengumpulan data, pada bagian ini dijelaskan tentang bagaimana dan darimana data dalam penelitian ini didapatkan, meliputi data sekunder dan data primer.
2. Pengolahan data awal, pengolahan data awal meliputi proses input data ke format yang dibutuhkan, pembersihan data, pengelompokan dan penentuan atribut data, serta pemecahan data (*split*) untuk digunakan dalam proses pembelajaran (*training*) dan pengujian (*testing*).
3. Metode yang diusulkan, Setelah proses pengolahan data awal dilakukan, maka dibuatkan model yang sesuai dengan jenis data. Pembagian data ke dalam data pelatihan (*training dataset*) dan data pengujian (*testing dataset*) juga diperlukan untuk pembuatan model. Metode yang diusulkan untuk digunakan pada prediksi prestasi siswa dengan menggunakan pemilihan atribut-atribut melalui uji coba.
4. Eksperimen dan pengujian model, pada bagian ini dijelaskan tentang langkah-langkah eksperimen meliputi cara pemilihan arsitektur yang tepat dari model atau metode yang diusulkan sehingga didapatkan hasil yang dapat membuktikan bahwa metode yang digunakan adalah tepat.
5. Evaluasi hasil, pada bagian ini dilakukan pengujian terhadap model-model untuk mendapatkan informasi model yang akurat. Evaluasi menggunakan metode *Confusion Matrix* dan *Curva ROC*.

HASIL DAN PEMBAHASAN

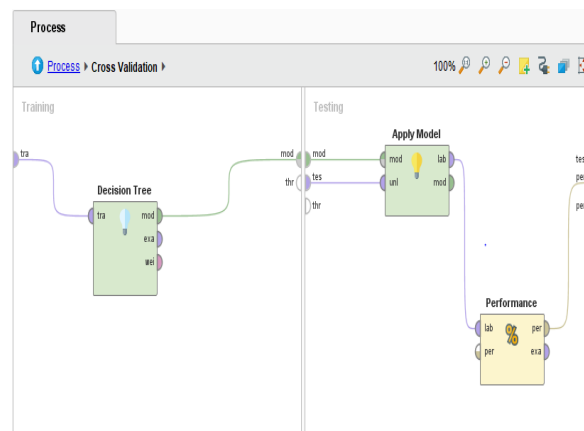
A. Eksperimen dan pengujian Model C4.5

Pembuatan model C4.5 dilakukan pada *dataset* yang terdiri dari 21 atribut yang merupakan atribut dari prediksi prestasi siswa dan *class* yang merupakan hasil akhir prediksi. Data kemudian di validasi agar proses pelatihan dapat berjalan dengan cepat dan mampu digunakan untuk melakukan pelatihan.

Tahap ini dibuatkan model pengolahan dengan menggunakan perangkat lunak aplikasi Rapidminer seperti Gambar 2. Model dari algoritma C4.5 yaitu berupa pohon keputusan, untuk dapat membuat pohon keputusan, langkah pertama adalah menghitung jumlah *class* yang berprestasi dan yang tidak berprestasi dari masing-masing *class* berdasarkan atribut yang telah ditentukan dengan menggunakan data *training*. Kemudian menghitung *Entropy* (Total) dengan menggunakan persamaan.



Gambar 2. Model Proses DT



Gambar 3. Model Validasi DT

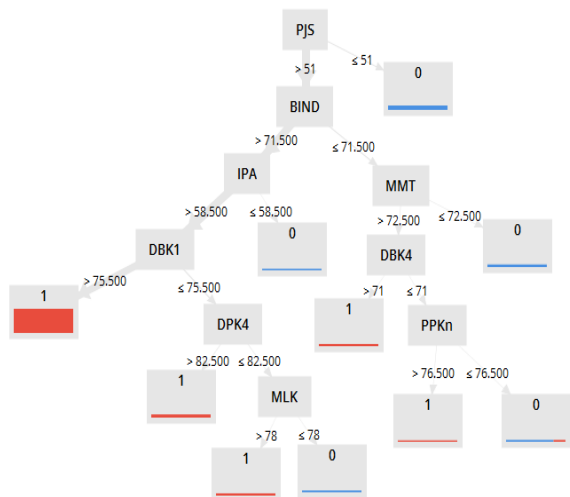
Untuk mendapat *gain* tiap atribut, maka harus dihitung *entropy* atribut berdasarkan tiap-tiap kasus (Persamaan 1,2 dan 3). Setelah *entropy* dari atribut sudah didapat maka langkah berikutnya menghitung *gain* pada Persamaan 4.

$$\begin{aligned} \text{Entropy (Total)} &= \\ &= \left(-\frac{188}{232} \log_2 \left(\frac{188}{232} \right) \right) + \left(-\frac{44}{232} \log_2 \left(\frac{44}{232} \right) \right) \quad (1) \\ \text{Entropy (Total)} &= 0,70074955 \end{aligned}$$

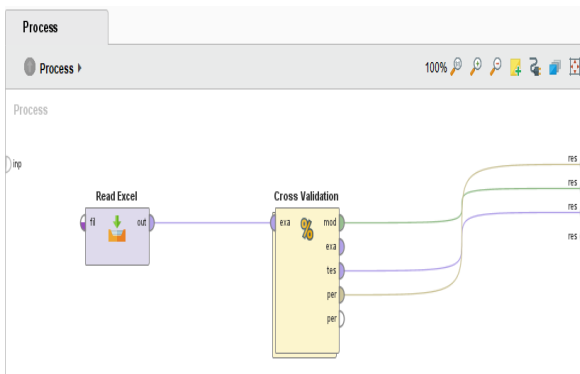
$$\begin{aligned} \text{Entropy (JK, 0)} &= \\ &= \left(-\frac{81}{113} \log_2 \left(\frac{81}{113} \right) \right) + \left(-\frac{32}{113} \log_2 \left(\frac{32}{113} \right) \right) \quad (2) \\ \text{Entropy (JK, 0)} &= 0,859755509 \end{aligned}$$

$$\begin{aligned} \text{Entropy (JK, 1)} &= \\ &= \left(-\frac{115}{119} \log_2 \left(\frac{115}{119} \right) \right) + \left(-\frac{4}{119} \log_2 \left(\frac{4}{119} \right) \right) \quad (3) \\ \text{Entropy (JK, 1)} &= 0,212201328 \end{aligned}$$

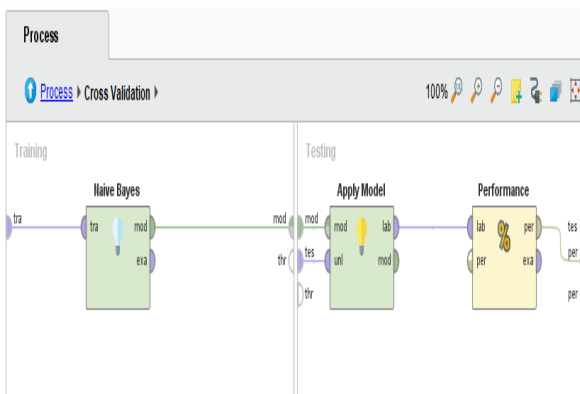
$$\begin{aligned} \text{Gain (Total, JK)} &= \\ &= 0,70074955 - \left(\left(\frac{113}{232} \right) * 0,859755509 \right) + \\ &= \left(\left(\frac{119}{232} \right) * 0,212201328 \right) \quad (4) \\ \text{Gain (Total, JK)} &= 0,173144677 \end{aligned}$$



Gambar 4. Pohon Keputusan Klasifikasi Prediksi Prestasi Siswa Menggunakan Algoritma C4.5



Gambar 5. Model Proses Naive Bayes



Gambar 6. Model Validasi Naive Bayes

Dari hasil perhitungan manual *entropy* dan *gain* keseluruhan atribut, diketahui atribut PJS memiliki nilai paling tinggi yaitu 0,422618468, oleh karena itu atribut PJS menjadi akar atau node yang pertama dari pohon keputusan yang terbentuk. Setelah didapatkan hasil perhitungan *entropy* dan *gain*, maka pohon keputusan yang terbentuk dapat dilihat seperti gambar 4.

B. Evaluasi dengan Confusion Matrix Model C4.5

Model *confusion matrix* akan membentuk matrix yang terdiri dari *true positive* atau tupel positif dan *true negative* atau tupel negatif, kemudian masukan data *testing* yang sudah disiapkan ke dalam *confusion matrix* sehingga didapatkan hasil pada Tabel 1. Berdasarkan Tabel 1 data *testing* terdapat rincian jumlah *True Positive* (TP) 178, *False Negative* (FN) 33, *False Positive* (FP) 11, dan *True Negative* (TN) 10. Dari data tersebut maka dapat dihitung nilai *accuracy*, *sensitivity*, *specificity*, *PPV*, dan *NPV*. Data hasil olahan dapat dilihat pada Tabel 2.

C. Eksperimen dan Pengujian Model Naive Bayes

Pembuatan model *Naive Bayes* dilakukan pada *dataset* yang terdiri dari 21 atribut yang merupakan atribut dari prediksi prestasi siswa dan *class* yang merupakan hasil akhir prediksi. Data kemudian di validasi agar proses pelatihan dapat berjalan dengan cepat dan mampu digunakan untuk melakukan pelatihan. Tahap ini dibuatkan model pengolahan dengan menggunakan perangkat lunak aplikasi Rapidminer seperti Gambar 5.

D. Evaluasi Dengan Confusion Matrix Model Naive Bayes

Model *confusion matrix* akan membentuk *matrix* yang terdiri dari *true positive* atau tupel positif dan *true negative* atau tupel negatif, kemudian masukan data *testing* yang sudah disiapkan ke dalam *confusion matrix* sehingga didapatkan hasil pada Tabel 3.

Tabel 1. Confusion Matrix Algoritma Klasifikasi C4.5 pada data Testing

	True 0	True 1	Class Precision
Pred. 0	33	10	76.74%
Pred. 1	11	178	94.18%
Class	75.00%	94.68%	
Recall			

Tabel 2. Nilai Accuracy, Sensitivity, Specificity, PPV dan NPV C4.5

	Nilai
Accuracy	0,9095
Sensitivity	0,8436
Specificity	0,4762
PPV	0,9418
NPV	0,2326

Tabel 3. *Confusion Matrix* Algoritma *Naïve Bayes* Pada data Testing

	True 0	True 1	Class Precision
Pred. 0	44	10	81.48%
Pred. 1	0	178	100.00%
Class recall	100.00%	94.68%	

Tabel 4. Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NPV* *Naïve Bayes*

	Nilai
Accuracy	0,9569
Sensitivity	0,8018
Specificity	1,0000
PPV	1,0000
NPV	0,1852

Tabel 5. Pengujian Algoritma Klasifikasi C4.5 dan *Naïve Bayes*

	Accuracy	AUC
C4.5	90,95%	0,639
<i>Naïve Bayes</i>	95,69%	0,999

Berdasarkan Tabel 3 dari data *testing* terdapat rincian jumlah *true Positive* (TP) 178, *False Negative* (FN) 44, *False Positive* (FP) 0, dan *True Negative* (TN) 10. Dari data tersebut maka dapat dihitung nilai *Accuracy*, *Sensitivity*, *Specificity* *PPV*, dan *NPV*. Data hasil olahan dapat dilihat pada Tabel 4.

E. Komparasi Model Algoritma C4.5 dengan Algoritma *Naïve Bayes*

Hasil pengujian model C4.5 dibandingkan dengan model *Naïve Bayes* dapat dilihat pada Tabel 5. Perbedaan akurasi dan AUC Algoritma C4.5 dan Algoritma *Naïve Bayes*, dapat dilihat pada Gambar 7 dan Gambar 8.

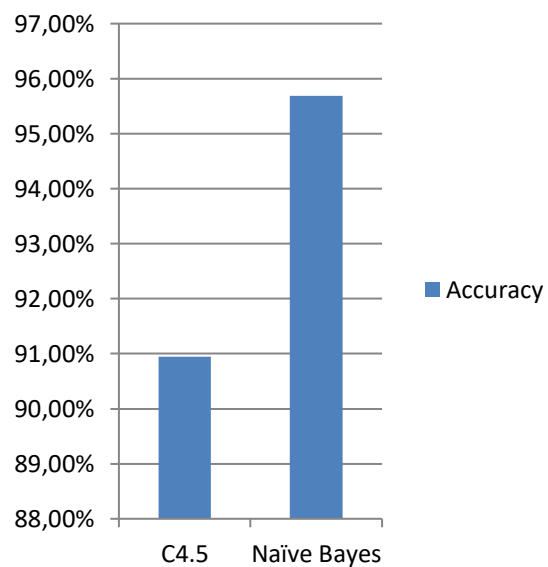
Dari hasil pengujian, dengan dilakukan evaluasi baik secara *confusion matrix* terbukti bahwa maupun ROC terbukti bahwa pengujian yang dilakukan oleh algoritma *Naïve Bayes* memiliki nilai akurasi yang lebih tinggi dibanding C4.5. Nilai akurasi untuk model algoritma *Naïve Bayes* sebesar 95,69% dengan nilai AUC 0,999 dan nilai akurasi model algoritma C4.5 sebesar 90,95% dengan nilai AUC 0,639. Berdasarkan nilai tersebut diperoleh selisih akurasi sebesar 4,74% dan selisih AUC sebesar 0,36.

KESIMPULAN

Dalam penelitian ini dilakukan analisis dan komparasi dua metode klasifikasi data mining yang memiliki karakteristik yang berbeda. Algoritma C4.5 atau *Decision Tree* yaitu metode yang mengubah data menjadi pohon keputusan dan menghasilkan *rule*. Dari hasil pengujian dengan mengukur kinerja kedua metode tersebut

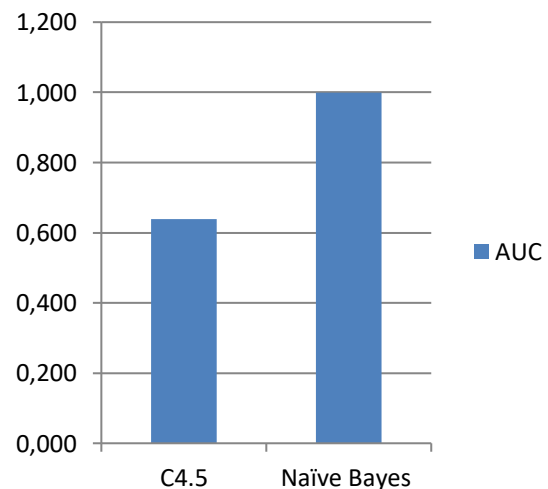
menggunakan *confusion matrix*, kurva ROC dan t-Test pada dataset SMK Al-Ijtihad Tangerang diketahui bahwa *Naïve Bayes* memiliki nilai akurasi 95,69%, nilai AUC 0,999 dan signifikan terhadap algoritma C4.5 memiliki nilai akurasi 90,95%, nilai AUC 0,639. Adapun model yang telah terbentuk selanjutnya dapat dikembangkan dan dapat diimplementasikan ke dalam sebuah aplikasi sehingga dapat membantu dan memudahkan bagi para pemegang kepentingan dalam pengambilan sebuah keputusan untuk memprediksi prestasi siswa.

Accuracy



Gambar 7. Perbedaan Akurasi Algoritma C4.5 dan Algoritma *Naïve Bayes*

AUC



Gambar 8. Perbedaan AUC Algoritma C4.5 dan Algoritma *Naïve Bayes*

REFERENSI

- Andini, T. I., Witanti, W., & Renaldi, F. (2016). Prediksi Potensi Pemasaran Produk Baru dengan Metode Naïve Bayes Classifier dan Regresi Linear. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 27-32.
- Arora, R. K. (2013). Evaluating Student ' s Performance Using k-Means Clustering, 8491.
- Defiyanti Sofi, M. K. (2013). Analisis dan Prediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining. *Syntax*, 2, 1-10.
- Mariño, M. A., Rezende, C. A., & Tasic, L. (2018). A multistep mild process for preparation of nanocellulose from orange bagasse. *Cellulose*, 25(10),5739-5750.
<https://doi.org/10.1007/s10570-018-1977-y>
- Nelson Butarbutar, Agus Perdana Windarto, Dedi Hartama, S. (2016). Komparasi Kinerja Algoritma Fuzzy C-Means Dan K-Means Dalam Pengelompokan Data Siswa Berdasarkan Prestasi Nilaiakademik Siswa. *JURASIK (Jurnal Riset Sistem Informasi & Teknik Informatika)*, 1(2012), 46-55.
<https://doi.org/10.30645/jurasik.v1i1.8>
- Novandya, Adhika, Oktria, I. (2017). Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi. *Jurnal Format*, 6(2),98-106.
<https://doi.org/10.1016/j.surfcoat.2005.02.204>
- Pagnotta, F. (2016). Using Data Mining To Predict Secondary Using Data Mining To Predict Secondary School, 2014(September), 0-9.
<https://doi.org/10.13140/RG.2.1.1465.8328>
- Purba, W., Tamba, S., & Saragih, J. (2018). The effect of mining data k-means clustering toward students profile model drop out potential. *Journal of Physics: Conference Series*,1007(1).
<https://doi.org/10.1088/1742-6596/1007/1/012049>
- Puspita, A., & Wahyudi, M. (2015). Algoritma C4.5 Berbasis Decision Tree untuk Prediksi Kelahiran Bayi Prematur. *Konferensi Nasional Ilmu Pengetahuan Dan Teknologi (KNIT)*, 1(1), 97-102. Retrieved from <http://konferensi.nusamandiri.ac.id/proceeding/index.php/KNIT/article/view/175>
- Rima Ramadhani, D. (2014). Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi. *Industrial Marketing Management*, 1(1), 1-9.
<https://doi.org/10.1016/j.indmarman.2016.05.016>
- Shovon, H. I., & Haque, M. (2012). An Approach of Improving Student ' s Academic Performance by using K-means clustering algorithm and Decision tree. *International Journal of Advanced Computer Science and Applications*, 3(8), 146-149.
- Sri Rahayu, Dodon T. Nugrahadi, F. I. (2014). Clustering Penentuan Potensi Kejahatan Daerah Di Kota Banjarbaru Dengan Metode K-Means. *Kumpulan jurnal Ilmu Komputer (KLIK)*, 1(1), 33-45.
- Susanto, Heri, and sudiyatno. 2014. "Data Mining Untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan Dan Prestasi Masa Lalu." *Jurnal Pendidikan Vokasi* 4(2): 222-31.
<http://journal.uny.ac.id/index.php/jpv/article/view/2547>.
- Widodo, & Wahyuni, D. (2017). Implementasi algoritma k-means clustering untuk mengetahui bidang skripsi mahasiswa multimedia pendidikan teknik informatika dan komputer universitas negeri jakarta. *Pinter*,1(2),157-156.
<https://doi.org/10.21009/pinter.1.2.10>