KLASIFIKASI SISWA SMK BERPOTENSI PUTUS SEKOLAH MENGGUNAKAN ALGORITMA *DECISION TREE, SUPPORT VECTOR MACHINE* DAN *NAIVE BAYES*

Dwi Arum Ningtyas[1]; Mochamad Wahyudi[2]; N. Nurajijah[3]
Magister Ilmu Komputer, STMIK Nusa Mandiri[1]
Teknologi Informasi, Universitas Bina Sarana Informatika[2]
Sistem Informasi, STMIK Nusa Mandiri[3]
https://www.nusamandiri.ac.id[1][3], https://www.bsi.ac.id[2]
dwiarumnigtyas@gmail.com[1], wahyudi@bsi.ac.id[2], nurajijah.nja@nusamandiri.ac.id[3]

Abstract—*Dropping out of school in Vocational High School students is an educational problem that must be found out the causes, so that it does not happen again in the future. The purpose of this study is to classify student data so that it can be predicted that students who have the potential to drop out of school use the Decision Tree, Naive Bayes and Support Vector Machine algorithms. Then determine which algorithm is the best. The results showed that the Support Vector Machine algorithm was the best with an accuracy of 93.77% and Area Under the Curve of 0.990.*

Keywords*: data mining classification, dropout students, Decision Tree, Naive Bayes, Support Vector Machine.*

Intisari—Putus sekolah pada siswa Sekolah Menengah Kejuruan merupakan permasalahan pendidikan yang harus dicari tahu faktor penyebabnya, agar tidak terjadi kembali kedepannya. Tujuan penelitian ini adalah melakukan klasifikasi terhadap data siswa sehingga dapat diprediksi siswa yang berpotensi putus sekolah menggunakan algoritma *Decision Tree, Naive Bayes* dan *Support Vector Machine*. Kemudian menentukan algoritma mana yang terbaik. Hasil penelitian menunjukkan bahwa algoritma *Support Vector Machine* menjadi yang terbaik dengan akurasi sebesar 93,77% dan *Area Under the Curve* sebesar 0,990.

Kata Kunci*: klasifikasi data mining, siswa putus sekolah, Decision Tree, Naive Bayes, Support Vector Machine.*

PENDAHULUAN

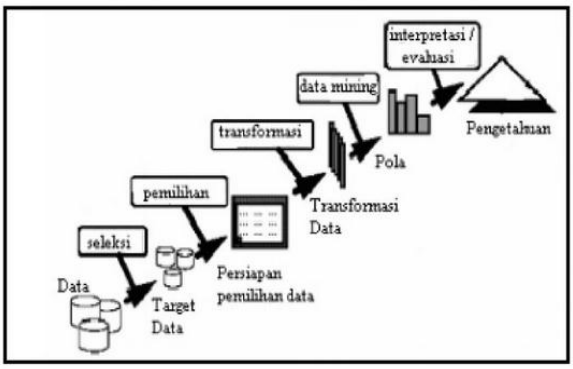
Putus sekolah merupakan permasalahan pendidikan yang perlu dicari tahu akar penyebabnya. Terlebih pada jenjang Sekolah Mengengah Kejuruan atau sederajat yang merupakan tingkat akhir program pemerintah yaitu wajib belajar 9 tahun, karena setelah itu dianggap sudah siap untuk terjun ke dunia kerja. Beberapa faktor penyebab diantaranya faktor ekonomi, kondisi geografis, dan keinginan siswa itu sendiri. Penelitian (Putri, 2017) menunjukkan bahwa faktor internal dianggap lebih dominan mempengaruhi terjadinya siswa putus sekolah, motivasi siswa dalam bersekolah serta kesadaran akan pentingnya pendidikan bagi para siswa rendah. Sedangkan faktor eksternal tidak mempengaruhi siswa putus sekolah . Faktor kerja adalah yang paling penting bagi siswa untuk tetap menghadiri kelas. Variabel paling penting kedua adalah status perkawinan. Akhirnya, usia juga dipengaruhi dalam keputusan drop out dari program (Villwock, Appio, & Andreta, 2015).

Data Mining adalah proses penggalian data dari tumpukan database yang berukuran besar yang digunakan untuk menemukan knowledge berupa informasi penting dan bermanfaat (Wahyuni, Saputra, & Perangin-angin, 2017). Salah satu algoritma data mining adalah klasifikasi serta banyak bidang ilmu yang menerapkan teknik klasifikasi dalam ilmu data mining untuk menyelesaikan masalah (Darmawan Wachid, 2017).

Pada penelitian sebelumnya, ID3 mampu menghasilkan sebuah pohon keputusan dari suatu kumpulan data yang sangat banyak. Pohon keputusan ini bisa dijadikan sebagai acuan prediksi kemungkinan terjadinya putus sekolah bagi siswa. Agar dapat menjadi acuan yang baik maka klasifikasi yang dihasilkan harus memiliki akurasi yang tinggi. PSO dapat meningkatkan akurasi berbagai macam klasifikasi data mining (Kurniawan & Rosadi, 2017). 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 (Noviriandini & Nurajijah, 2019). Menganalisis data menggunakan *decision tree*, data mining untuk memprediksi prestasi belajar siswa berdasarkan status sosial ekonomi orang tua, disiplin siswa dan prestasi

belajar siswa (Kuntoro & Sudarwanto, 2017). Melakukan perbandingan antara kinerja dari metode klasifikasi teknis naïve Bayes dan algoritma C4.5 (Marlina, lim, & Utama Siahaan, 2016). Penerapan metode Support Vector Machine mampu memisahkan siswa yang berpotensi baik dan yang berpeluang *drop out*. Metode ini mengolah kalkulasi data dari nilai akhir, prilaku, dan kehadiran (Fiska, 2017). Algoritma *Support vector Machine* menjadi algoritma yang menghasilkan akurasi terbaik dibanding *Naive Bayes* dan *Decision Tree* dalam proses klasifikasi (Nurajijah & Riana, 2019).

Berdasarkan penelitian sebelumnya algoritma klasifikasi yang menghasilkan tingkat akurasi tinggi yaitu *Support Vector Machine, Naive Bayes* dan *Decision Tree*. Pada penelitian sebelumnya belum dilakukan penerapan ketiga algoritma ini dalam satu penelitian. Penelitian ini bertujuan untuk melakukan klasifikasi siswa Sekolah Menengah Kejuruan menggunakan ketiga algoritma tersebut untuk kemudian dicari algortma mana yang menghasilkan akurasi terbaik. Manfaat penelitian ini adalah agar dapat diprediksi siswa yang berpotensi putus sekolah.



Sumber: Nofriansyah & Nurcahyo (2015)
Gambar 1. Proses *Knowledge Discovery in Database* (KDD)

BAHAN DAN METODE

Penelitian ini menggunakan data siswa Sekolah Menengah Kejuruan untuk dilakukan klasifikasi menggunakan algoritma *Decission Tree, Naive Bayes*, dan *Support Vector Machine* (SVM).

Decision Tree adalah pohon terstruktur dari sekumpulan atribut untuk diuji dengan tujuan meramalkan *output*-nya (Giovani, Mudjihartono, & Pranowo, 2011), di mana setiap simpul internal menunjukkan tes pada atribut, masing-masing cabang mewakili hasil dari tes, dan setiap node daun memegang label kelas. Node paling atas dalam sebuah pohon adalah simpul akar (Jiawei, Kamber, Han, Kamber, & Pei, 2012). Menentukan akar dari pohon dengan menghitung nilai gain yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelum menghitung nilai gain terlebih dahulu mencari nilai entropy dengan rumus pada persamaan 1 kemudian menghitung nilai gain pada persamaan 2 dengan keterangan S (himpunan kasus), n (jumlah partisi S), pi (jumlah kasus pada partisi ke-i), |Si| (jumlah kasus pada partisi ke-i) dan |S| (jumlah kasus dalam S).

$$Entropy\ (S) = \sum_{i=0}^{n} -pi * log^2 pi \quad \text{.....(1)}$$

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

Algoritma *Naive Bayes* adalah salah satu algoritma yang terdapat pada teknik klasifikasi (Marlina et al., 2016) probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yang memprediksi peluang masa depan berdasarkan pengalaman sebelumnya dan dikenal sebagai *Teorema Bayes* dihitung dengan persamaan 3. P merupakan peluang, Variabel y adalah variabel kelas (lancar/macet), Variabel X mewakili parameter / fitur.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad \text{.....(3)}$$

Support Vector Machines (SVM) adalah seperangkat metode pembelajaran yang menganalisis data dan mengenali pola, digunakan untuk klasifikasi dan analisis regresi, tidak hanya itu teknik ini dapat melakukan prediksi dan penilaian tentang sebuah system (Maulana, Setyanto, & Kurniawan, 2018).

Metode penelitian yang digunakan adalah *Knowledge Discovery in Database* (Nofriansyah, 2017) dengan tahapan (Gambar 1) seperti berikut:

1. Seleksi

Dataset yang diperoleh dari salah satu Sekolah Menengah Kejuruan di Tangerang sebanyak 224 *record* yang terdiri dari 181 *record* tidak berpotensi putus sekolah dan 43 *record* berpotensi putus sekolah. Data siswa yang telah dikumpulkan kemudian dilakukan penyeleksian kriteria menjadi 16 kriteria dideskripsikan pada Tabel 1 yaitu jenis kelamin (A), nilai (B), hadir (C), sakit(D), ijin (E), alpha (F), jarak (G), pekerjaan ayah (H), pekerjaan ibu (I), penghasilan ayah (J), penghasilan ibu (K), pendidikan ayah (L), pendidikan ibu (M), jumlah tanggungan (N), pelanggaran (O), dan status berisi berpotensi putus (0) atau tidak berpotensi butus (1). Penyeleksian kriteria ini dilakukan untuk dijadikan kriteria penentu dalam klasifikasi data.

Tabel 1. Dataset Penelitian

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Status
L	60-73	< 100	> 3	tidak ada	> 3	dekat	wiraswasta	karyawan swasta	1000000-1999999	500000-999999	sma / sederajat	sma / sederajat	> 4	< 2	1
P	74-87	< 100	1-3	1-3	> 3	sedang	karyawan swasta	tidak bekerja	2000000-4999999	< 500000	sma / sederajat	sma / sederajat	2-4	< 2	1
P	74-87	< 100	> 3	1-3	tidak ada	dekat	wiraswasta	tidak bekerja	2000000-4999999	tidak berpenghasilan	sma / sederajat	sd / sederajat	> 4	< 2	1
P	74-87	< 100	1-3	1-3	tidak ada	sedang	wiraswasta	tidak bekerja	1000000-1999999	tidak berpenghasilan	sma / sederajat	sma / sederajat	2-4	< 2	0
P	74-87	< 100	1-3	1-3	tidak ada	sedang	karyawan swasta	karyawan swasta	2000000-4999999	2000000-4999999	d3	d3	2-4	< 2	0
P	74-87	< 100	> 3	tidak ada	1-3	dekat	buruh	guru	1000000-1999999	2000000-4999999	sma / sederajat	s1	2-4	< 2	1
P	74-87	lengkap	tidak ada	tidak ada	tidak ada	dekat	wiraswasta	tidak bekerja	1000000-1999999	tidak berpenghasilan	sma / sederajat	sma / sederajat	2-4	< 2	1
P	74-87	< 100	1-3	tidak ada	tidak ada	dekat	karyawan swasta	karyawan swasta	2000000-4999999	2000000-4999999	smp / sederajat	smp / sederajat	2-4	< 2	0
P	74-87	< 10	1-3	tidak ada	tidak ada	dekat	karyawan swasta	tidak bekerja	2000000-4999999	tidak berpenghasilan	s1	sma / sederajat	< 2	< 2	1

2. Pemilihan

Membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaikin kesalahan pada data, seperti kesalahan penulisan (tipografi). Juga dilakukan proses *encrihment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformasi

Mentransformasi bentuk data yang belum memiliki entitas yang jelas kedalam bentuk data yang valid atau siap untuk dilakukan proses data mining.

4. Implementasi Model *Data Mining*

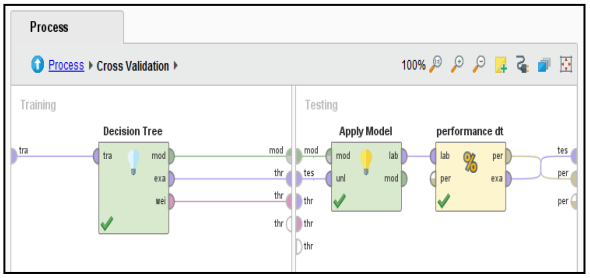
Pemodelan dilakukan dengan menggunakan rapid miner yang digunakan untuk membandingkan tingkat akurasi dengan menggunakan algortima *Decission Tree, Naive Bayes* dan *Support Vector Machine* (SVM).

5. Evaluasi

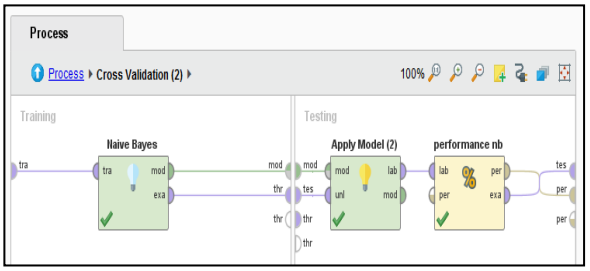
Tahap ini merupakan hasil dari pengolahan pada proses validasi untuk mengetahui akurasi dari ketiga metode yang digunakan yaitu algoritma *Decission Tree, Naive Bayes* dan Support Vector Machine (SVM). Pada bagian ini dilakukan pengujian terhadap model-model untuk mendapatkan informasi model yang akurat. Evaluasi menggunakan metode *Confusion Matrix, Kurva ROC* dan pengujian *T-Test*.

HASIL DAN PEMBAHASAN

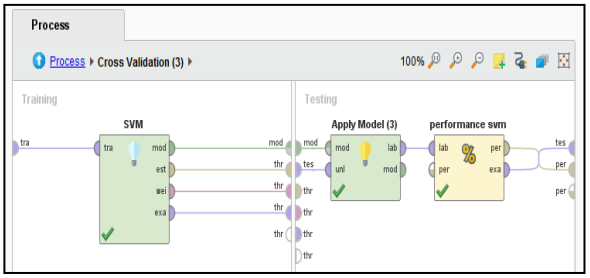
Proses validasi algoritma *Decision Tree* , *Naive Bayes* dan *Support Vector Machine* pada klasifikasi data siswa Sekolah Menengah Kejuruan (Tabel 1) yang dijelaskan pada Gambar 2, 3 dan 4 menggunakan RapidMiner dengan *K-Fold Cross Validation* untuk mengetahui Confusion Matrix dan *Accuracy, Precision, Recall*, dan AUC. *Cross Validation* merupakan teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi (Puspita & Wahyudi, 2015). K-Fold *Cross Validation* merupakan salah satu dari variasi teknik pengujian cross validation, dilakukan dengan membagi training set dan *test* set (Pattipeilohy, Wibowo, & Utari, 2017).



Gambar 2. Validasi Algoritma *Decision Tree*



Gambar 3. Validasi Algoritma *Naive Bayes*

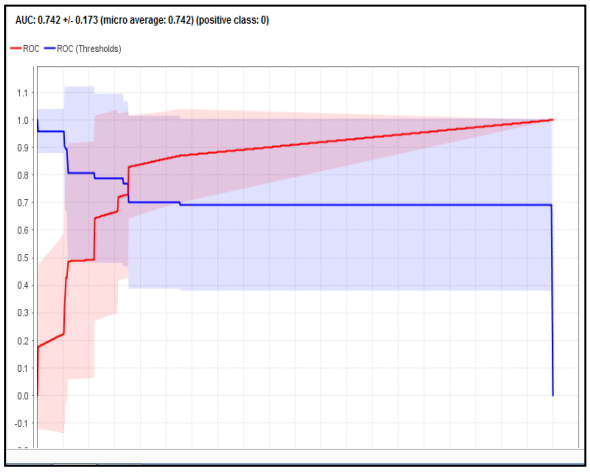


Gambar 4. Validasi Algoritma SVM

accuracy: 87.91% +/- 7.27% (micro average: 87.95%)

	true 1	true 0	class precision
pred. 1	164	10	94.25%
pred. 0	17	33	66.00%
class recall	90.61%	76.74%	

Gambar 5. *Confusion Matrix* dan Akurasi Algoritma *Decision Tree*

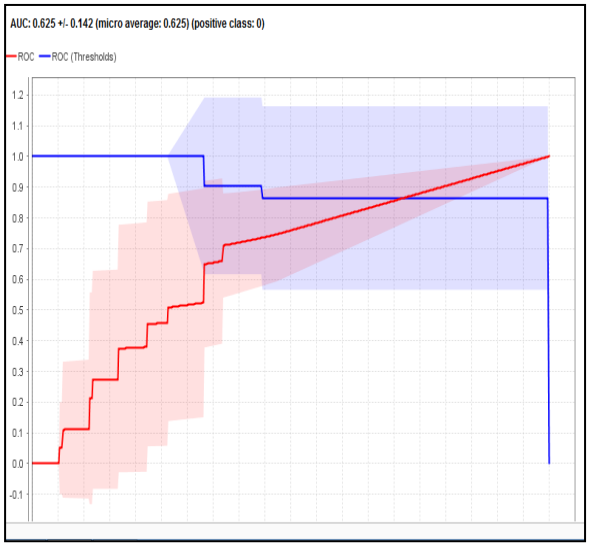


Gambar 6. Hasil AUC Algoritma *Desion Tree*

accuracy: 70.12% +/- 10.25% (micro average: 70.09%)

	true 1	true 0	class precision
pred. 1	128	14	90.14%
pred. 0	53	29	35.37%
class recall	70.72%	67.44%	

Gambar 7. *Confusion Matrix* dan Akurasi Algoritma *Naive Bayes*



Gambar 8. Hasil AUC Algoritma *Naive Bayes*

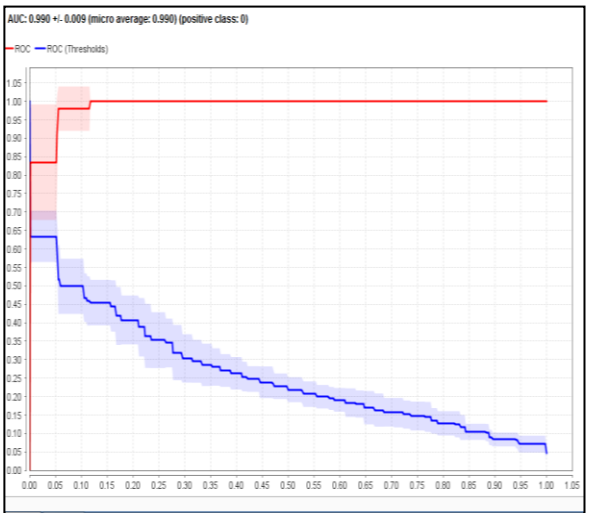
Hasil pengujian menggunakan Rapidminer digambarkan melalui *Confussion Matrix* untuk mengetahui nilai akurasi, recall dan presisi. Klasifikasi menggunakan algoritma *Decission Tree* menghasilkan nilai akurasi sebesar 87,91% (gambar 5) dan nilai *Area Under the Curve* dihitung untuk mengukur perbedaan performansi, dengan hasil sebesar 0,742 yang termasuk dalam kategori sedang atau *Fair Classification* (gambar 6). Nilai ini tidak sebesar pada penelitian sebelumnya yang menggunakan algoritma *Decision Tree* dengan nilai akurasi sebesar 91,84% (Villwock et al., 2015) dan 98,14% (Abu-Oda & El-Halees, 2015).

Penerapan algoritma *Naive Bayes* pada klasifikasi data siswa berpotensi putus sekolah menghasilkan nilai akurasi sebesar 70.12% ynag dieskripsikan pada Gambar 7 dan nilai *Area Under the Curve* sebesar 0.625 termasuk dalam kategori *performace* klasifikasi buruk atau *poor classification* (gambar 8), hasil ini dibawah dari akurasi penelitian (Darmawan Wachid, 2017) menggunakan *Naive Bayes* dengan 92,72%.

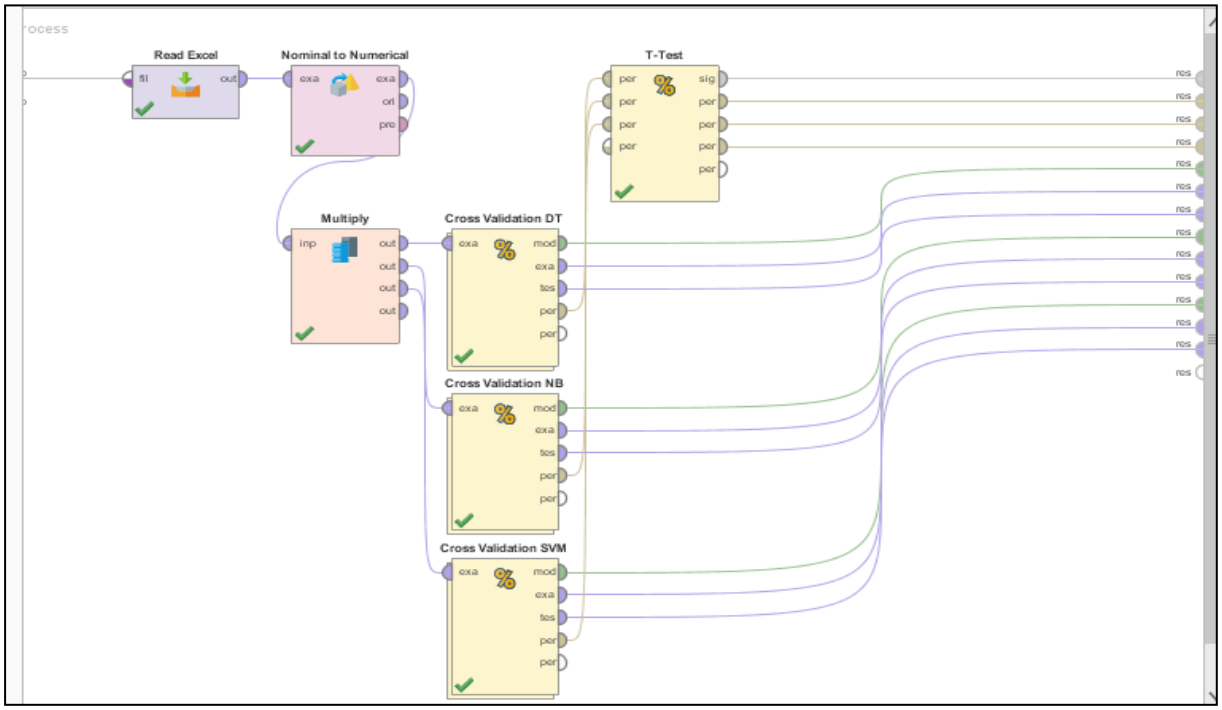
accuracy: 93.77% +/- 4.06% (micro average: 93.75%)

	true 1	true 0	class precision
pred. 1	169	2	98.83%
pred. 0	12	41	77.36%
class recall	93.37%	95.35%	

Gambar 9. *Confusion Matrix* dan Akurasi Algoritma SVM



Gambar 10. Hasil AUC Algoritma SVM



Gambar 11. Desain Model *T-Test*

Tabel 2. Perbandingan Hasil Pengujian

	Decision Tree	*Naive Bayes*	(SVM)
Accuracy	87.91%	70.12%	93.77%
Precision	70.63%	38.26%	79.50%
Recall	77.00%	67.00%	95.50%
AUC	0.742	0.625	0.990

Tabel 3. Hasil Pengujian *T-Test*

	Decision Tree	**Naive Bayes**	**SVM**
	0.879+/- 0.073	0.701+/- 0.102	0.938+/- 0.041
0.879+/- 0.073		*0.000*	*0.039*
0.701+/- 0.102			*0.000*
0.938+/- 0.041			

Nilai akurasi klasifikasi menggunakan algoritma *Support Vector Machine* mencapai 93.77% (gambar 9) dan nilai *Area Under the Curve* sebesar 0.990 merupakan nilai tertinggi yang didapat, termasuk dalam kategori sangat baik dalam melakukan klasifikasi atau Excelent (gambar 10). Nilai ini lebih baik dibanding penelitian (Fiska, 2017) dengan hasil akurasi 43.33% menggunakan *Support Vector Machine*.

Tabel 1 mendeskripsikan hasil pengujian dari seluruh algoritma yang digunakan pada penelitian ini dalam melakukan klasifikasi dta siswa Sekolah mengah Kejuruan yang berpotensi putus sekolah. Algoritma *Support Vector Machine* menunjukkan hasil akurasi dan nilai *Area Under the Curve* terbaik dibanding Algoritma *Decision Tree* dan *Naive Bayes*.

Pengujian *T-Test* dilakukan menggunakan Rapidminer untuk mengetahui tingkat signifikan dari algoritma yang digunakan. Desain model T-test dapat dilihat pada gambar 11. Hasil pengujian *T-test* pada tabel 2 menunjukan terdapat perbedaan yang signifikan antara *algoritma Naive Bayes* dengan Decision Tree *dan Support Vector Machine* (SVM). Algoritma *Support Vector Machine* memiliki akurasi terbaik dan dapat dipertimbangkan dalam klasifikasi data siswa berpotensi putus sekolah.

KESIMPULAN

Penelitian ini telah melakukan pengujian model menggunakan *Decission Tree, Naive Bayes* dan *Support Vector Machine* (SVM) dalam klasifikasi untuk data siswa Sekolah Menengah Kejuruan yang berpotensi putus sekolah. Beberapa eksperimen dilakukan untuk melihat hasil terbaik. Hasil penelitian ini dengan

menggunakan model algoritma *Decision Tree* menghasilkan akurasi sebesar 87,91% dan AUC sebesar 0,742, algoitma *Naive Bayes* menghasilkan akurasi sebesar 70,12% dan AUC sebesar 0625, dan algoritma *Support Vector Machine* (SVM) dengan hasil terbaik yaitu akurasi sebesar 93,77% dan *AUC* sebesar 0,990. Penelitian ini membuktikan bahwa algoritma *Support Vector Machine* dapat digunakan dengan baik untuk klasifikasi data siswa berpotensi putus sekolah.

REFERENSI

Abu-Oda, G. S., & El-Halees, A. M. (2015). Data Mining In Production Management And Manufacturing. *Annals Of Daaam And Proceedings Of The International Daaam Symposium*, *5*(1), 827–828. Https://Doi.Org/10.2507/Daaam.Scibook.2009.11

Darmawan Wachid. (2017). Algoritma Klasifikasi Data Mining Untuk Prediksi Status Mhasiswa Menggunakan Algortima Decision Tree C4.5. *Ic-Tech*, *Xii*(1), 15–22.

Fiska, R. R. (2017). Penerapan Teknik Data Mining Dengan Metode Support Vector Machine (Svm) Untuk Memprediksi Siswa Yang Berpeluang Drop Out (Studi Kasudi Smkn 1 Sutera). *Jaringan Sistem Informasi Robotik*, *1*(01), 42–51.

Giovani, R. A., Mudjihartono, P., & Pranowo, P. (2011). Sistem Pendukung Keputusan Prediksi Kecepatan Studi Mahasiswa Menggunakan Metode Id3. *Jurnal Buana Informatika*, *2*(2), 102–108. Https://Doi.Org/10.24002/Jbi.V2i2.313

Jiawei, H., Kamber, M., Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts And Techniques. In *San Francisco, Ca, Itd: Morgan Kaufmann*. Https://Doi.Org/10.1016/B978-0-12-381479-1.00001-0

Kuntoro, R. K., & Sudarwanto, R. (2017). *Prediction Of Student Performance Using Decision Tree C 4 . 5 Algorithm*. 214–219.

Kurniawan, M. Y., & Rosadi, M. E. (2017). Optimasi Decision Tree Menggunakan Particle Swarm Optimization Pada Data Siswa Putus Sekolah. *Jtiulm*, *2*(1), 15–22.

Marlina, L., Lim, M., & Utama Siahaan, A. P. (2016). Data Mining Classification Comparison (Naïve Bayes And C4.5 Algorithms). *International Journal Of Engineering Trends And Technology*, *38*(7), 380–383. Https://Doi.Org/10.14445/22315381/Ijett-V38p268

Maulana, M. A., Setyanto, A., & Kurniawan, M. P. (2018). *Analisis Sentimen Media Sosial Universitas Amikom*. 7–12.

Nofriansyah, D. (2017). *Algoritma Data Mining Dan Pengujian*. Yogyakarta: Deepublish.

Noviriandini, A., & Nurajijah, N. (2019). Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Prestasi Siswa Sekolah Menengah Kejuruan. *Jitk (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *5*(1), 23–28. Https://Doi.Org/10.33480/Jitk.V5i1.607

Nurajijah, N., & Riana, D. (2019). Algoritma Naïve Bayes, Decision Tree, Dan Svm Untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah. *Jurnal Teknologi Dan Sistem Komputer*, *7*(2), 77. Https://Doi.Org/S10.14710/Jtsiskom.7.2.2019.77-82

Pattipeilohy, W. F., Wibowo, A., & Utari, D. R. (2017). Pemodelan Dan Prototipe Sistem Informasi Untuk Prediksi Pembaharuan Polis Asuransi Mobil Menggunakan Algoritma C.45. *Seminar Nasional Teknologi Dan Informatika 2017 (Snatif)*, (October), 791–799. Retrieved From Https://Www.Neliti.Com/Id/Publications/173500/Pemodelan-Dan-Prototipe-Sistem-Informasi-Untuk-Prediksi-Pembaharuan-Polis-Asuran

Puspita, A., & Wahyudi, M. (2015). Algoritma C4.5 Berbasis Decision Tree Untuk Prediksi Kelahiran Bayi Prematur. *Konferensi Nasinal Ilmu Pengetahuan Dan Teknologi (Knit)*, *1*(1), 97–102. Retrieved From Http://Konferensi.Nusamandiri.Ac.Id/Proceeding/Index.Php/Knit/Article/View/175

Putri, R. T. N. (2017). Faktor Penyebab Siswa Putus Sekolah Di Sekolah Menengah. *Jurnal Hanata Widya*, *6*(8), 70–82.

Villwock, R., Appio, A., & Andreta, A. A. (2015). Educational Data Mining With Focus On Dropout Rates. *International Journal Of Computer Science And Network Security*, *15*(3), 17–23.

Wahyuni, S., Saputra, K., & Perangin-Angin, M. I. (2017). *Implementasi Rapidminer Dalam Menganalisa Data Mahasiswa Drop Out*. *10*, 2013–2016. Https://Doi.Org/10.3969/J.Issn.1002-5006.2017.06.012