Prediksi Kecocokan Jurusan Siswa SMK Dengan Support Vector Machine dan Random Forest

Dicka Y Kardono [1*], Yuliana Melita Pranoto [2], Endang Setyati [3]

[1,2,3] Program Magister Teknologi Informasi, Institut Sains dan Teknologi Terpadu Surabaya, Jawa Timur
Email: [1*] dyka.gitu@gmail.com, [2] ymp@stts.edu, [3] endang@stts.edu

(Naskah masuk: 7 Nov 2022, direvisi: 15 Des 2022, 20 Jan 2023, diterima: 26 Jan 2023)

Abstrak

SMK Antartika 1 Sidoarjo setiap tahunnya melakukan penerimaan siswa baru. Siswa SMP yang mendaftakan diri ke SMK Antartika 1 Sidoarjo rata-rata belum cukup mengetahui tentang minatnya pada jurusan yang ada di sekolah. Adapun jurusan yang ada di SMK Antartika 1 Sidoarjo adalah Teknik Pemesinan, Teknik Kendaraan Ringan, dan Rekayasa Perangkat Lunak. Dari permasalahan di atas, maka diperlukan sebuah sistem untuk memprediksi tentang pemilihan kecocokan jurusan pada siswa baru SMK Antartika 1 Sidoarjo. Dengan adanya sistem tersebut dapat membantu meningkatkan pelayanan terhadap siswa baru dalam memutuskan pemilihan jurusan yang terdiri dari 4 tingkatan, yaitu: sangat cocok, cocok, kurang cocok, dan sangat kurang cocok dengan siswa. Untuk mengetahui pola prediksi dari data siswa tersebut, menggunakan penerapan perbandingan menggunakan metode *Support Vector Machine* (SVM) dan *Random Forest* (RF). Fitur atribut yang digunakan ada 14 fitur yang terdiri dari: Jurusan, Pendidikan_Ayah, Penghasilan_Ayah, Pendidikan_Ibu, Penghasilan_Ibu, Transportasi_ke_Sekolah, NUS_MTK_SMP, NUS_BIND_SMP, NUS_BING_SMP, Disiplin, Tanggung_Jawab, Sikap, Komunikasi, dan Output_Kelas. Riset ini menggunakan *dataset* siswa SMK Antartika 1 Sidoarjo mulai tahun 2020 sampai 2022 dengan total 578 *record* data siswa. Berdasarkan hasil analisis dengan metode SVM dengan *kernel sigmoid* diperoleh tingkat akurasi sebesar 83%, sedangkan hasil analisa dengan dengan metode RF dengan jumlah *tree* 150 diperoleh tingkat akurasi sebesar 82%.

Kata Kunci : SVM, RF, Klasifikasi, *Educational Data Mining*

Prediction of Major Selection of Vocational School Students with Support Vector Machine and Random Forest

Abstract

SMK Antartika 1 Sidoarjo annually accepts new students. On average, junior high school students who register at SMK Antartika 1 Sidoarjo do not know enough about their interest in the majors at school. While the majors at SMK Antartika 1 Sidoarjo in Mechanical Engineering, Automotive Engineering, and Software Engineering. From the problems above, it is necessary to have a system to predict the selection of suitable majors for new students of SMK Antartika 1 Sidoarjo. With this system, it can help improve services to new students in deciding the selection of majors which consist of 4 levels, namely: very suitable, suitable, less suitable, and very less suitable for students. To find out the prediction pattern of the student data, the Support Vector Machine (SVM) and Random Forest (RF) methods were used. There are 14 features used, namely: Jurusan, Pendidikan_Ayah, Penghasilan_Ayah, Pendidikan_Ibu, Penghasilan_Ibu, Transportasi_ke_Sekolah, NUS_MTK_SMP, NUS_BIND_SMP, NUS_BING_SMP, Disiplin, Tanggung_Jawab, Sikap, Komunikasi, dan Output_Kelas. This research uses a dataset of students from SMK Antartika 1 Sidoarjo Vocational School from 2020 to 2022 with a total of 578 lines. Based on the results of the analysis with the SVM method using sigmoid kernel obtained an accuracy rate of 83%, while the results of the analysis with the RF method using number of trees 150 obtained an accuracy rate of 82%.

Keywords : *SVM, RF, Classification, Educational Data Mining*

I. PENDAHULUAN

Kemajuan teknologi informasi saat ini berkembang pesat disegala bidang kehidupan manusia, tidak terkecuali pada bidang pendidikan. Penerapan teknologi informasi pada sekolah menengah kejuruan menghasilkan data yang berlimpah mengenai siswa dan proses pembelajarannya. Apabila data siswa tersebut kita olah dengan baik, tentunya akan dapat menghasilkan informasi yang bermanfaat untuk kemajuan pendidikan sekolah. Salah satu informasi yang dapat dijadikan penelitian yang menarik adalah tentang kualitas pendidikan sekolah. Sekolah menengah kejuruan yang berhasil dalam menjalankan sistem pendidikannya adalah sekolah yang dapat mengarahkan siswanya dalam memilih jurusan sesuai minat dan dunia kerja. Kemampuan sekolah dalam memberikan arahan pada siswa dalam memilih jurusan yang tepat sejak awal dapat memberikan jaminan kesempatan untuk mendapat pekerjaan yang baik pada masa mendatang. Berdasarkan latar belakang di atas, penelitian ini membahas tentang prediksi kecocokan pemilihan jurusan pada siswa SMK Antartika 1 Sidoarjo agar dikemudian hari mendapatkan hasil belajar maksimal yang sesuai jurusannya. Salah satu alat yang digunakan untuk menyelesaikan masalah ini adalah dengan menggunakan *Educational data mining (EDM)*. *Educational data mining* (EDM) adalah aplikasi teknik *Data Mining* (DM) untuk data pendidikan, dan sebagainya, tujuannya adalah untuk menganalisis jenis data ini untuk menyelesaikan masalah penelitian pendidikan [1]. *Data mining* merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat yang tersimpan didalam *database* besar. Sedangkan untuk algoritma *data mining* yang untuk penelitian ini ada 2 yaitu, *Random Forest (RF)* dan S*upport Vector Machine (SVM)*. Beberapa penelitian sebelumnya yang sudah dilakukan tentang *Educational Data Mining*, seperti pada *paper* berikut. Penulis memprediksi dan menganalisis kemampuan dari siswa dengan menggunakan model MLP dan *random forest* [2]. Pada penulis lain melakukan studi tentang prediksi siswa *drop out* dengan menggunakan model *logistic regression* [3]. Penulis lain lainnya juga membahas tentang *Educational Data Mining* tentang prediksi kemampuan akademik siswa dengan menggunakan model *support vector machine* [4]. Penelitian lainnya yang juga membahas tentang *educational data mining* yang membahas tentang prediksi kemampuan akademik siswa berdasarkan dari *background* dan kegiatan sosial yang dilakukan siswa [5]. Penelitian lainnya yang membahas tentang analisis *educational data mining* dengan menggunakan model klasifikasi dan membandingkan peforma antar model klasifikasi yang digunakan [6]. Penelitian tentang *educational data mining* selanjutnya adalah tentang memprediksi keputusan menentukan jurusan pada SMK dengan metode *simple additive weight* [7]. Penelitian selanjutnya yang membahas tentang *educational data mining* yang membahas tentang prediksi tingkat kelulusan siswa dengan menggunakan model *support vector machine (SVM)* dan *decision tree* [8]. Penelitian berikutnya yang juga membahas *educational data mining* yang membahas tentang algoritma seleksi fitur yang akan digunakan pada proses *data mining* [9]. Penelitian selanjutnya penggunaan *support vector machine* pada *educational data mining* untuk mengklasifikasikan perhatian siswa dalam personalisasi pengembangan sistem pembelajaran [10]. Penelitian berikutnya yang membahas *educational data mining* tentang prediksi pekerjaan siswa di masa mendatang [11]. Penelitian selanjutnya adalah membahas *educational data mining* adalah tentang analisa perbandingan model *association rule* dengan *classification* [12]. Penelitian yang terakhir membahas sistem rekomendasi pemilihan jurusan pada universitas berdasarkan profil dan minat siswa dengan metode asosiasi [13]. Perbedaan penelitian ini dengan penelitian sebelumnya adalah penelitian ini mencoba mengkomparasikan kemampuan dua algoritma dalam melakukan klasifikasi data siswa berdasarkan kecocokan pemilihan jurusannya. Tujuan dari penelitian ini untuk mengetahui perbandingan kinerja kedua algoritma yaitu *random forest* dan *support vector machine* berdasarkan nilai akurasi, *recall*, presisi, AUC, dan *F1-Score*.

II. METODE PENELITIAN

A. Pengertian CRISP-DM

CRISP-DM [14] memiliki kepanjangan *Cross-Industry Standard Process for Data Mining* adalah sebuah metode *data mining* yang dikembangkan bersama antara Daimler-Chrysler, SPSS, dan NCR dimana dari namanya merupakan sebuah metode netral dan dapat digunakan dalam segala lini bisnis dan berbagai *tool*. Sebagai sebuah metodologi, CRISP-DM menggambarkan fase dari beberapa tahapan dalam sebuah proyek, pekerjaan yang terkait dalam tiap fase dan penjabaran terkait hubungan antar pekerjaan tersebut serta memberikan sebuah gambaran siklus hidup (*life-cycle*) dari *Data Mining* sebagai model proses. CRISP-DM ini memiliki 6 tahapan model seperti pada Gambar 1 dalam keseluruhan proses *data mining* yaitu: *Bussiness Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment*.

Gambar 1. Proses CRISP-DM *Lifecycle*

B. *Support Vector Machine*

SVM [15] sudah ada sejak tahun 1992 ketika ada kebutuhan akan metode klasifikasi dan regresi berdasarkan beberapa prediksi. SVM diperkenalkan oleh Vapnik, Guyon, dan Boser di COLT-92. Untuk memisahkan data apa pun, SVM mendefinisikan kelas-kelas tertentu dan tergantung pada kompleksitas *dataset*.

SVM mendefinisikannya sebagai klasifikasi linear atau klasifikasi *nonlinier*. SVM hanya dapat didefinisikan sebagai metode prediksi dengan mencari garis tertentu atau batas keputusan yang disebut *hyperplane* yang dengan mudah memisahkan kumpulan data atau kelas untuk menghindari *overfit* ekstra ke data. Penggunaan kernel bertujuan untuk mentransformasikan data ke ruang berdimensi tinggi, dengan menjadikan data *non linier* terpisah secara *linier*.

Ada beberapa pilihan fungsi kernel yang dipakai pada sebuah aplikasi untuk mengatasi masalah pada metode *Support Vector Machine* (SVM) yaitu:

1. *Linier Kernel*

$$K(x_i, x) = x_i^T x \quad (1)$$

Dengan x_i merupakan data latih (*training*), x adalah data uji

2. *Polynomial Kernel*

$$K(x_i, x) = \left(\gamma(x_i^T x) + r\right)^p \quad (2)$$

Dengan x merupakan data latih (*training*), adalah data uji, p adalah derajat polinomial.

3. *Radial Basis Function* (RBF)

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right) \quad (3)$$

4. *Sigmoid Kernel*

$$K(x_i, x) = tanh(\gamma(x_i^T x) + r) \quad (4)$$

Dengan x merupakan data latih (*training*), adalah data uji, r adalah koefisien.

C. *Random Forest*

Random Forest [16] adalah algoritma *supervised learning* yang dikeluarkan oleh Breiman pada tahun 2001. Random Forest biasa digunakan untuk menyelesaikan masalah yang berhubungan dengan klasifikasi, regresi, dan sebagainya.

Algoritma ini berupa kombinasi dari beberapa *tree predictors* atau bisa disebut *decision trees* dimana setiap *tree* bergantung pada nilai *random vector* yang dijadikan sampel secara bebas dan merata pada semua *tree* dalam *forest* tersebut. Hasil prediksi dari *Random Forest* didapatkan melalui hasil terbanyak dari setiap individual *decision tree* (*voting* untuk klasifikasi dan rata-rata untuk regresi). Untuk RF yang terdiri dari N *trees* dirumuskan sebagai:

$$l(y) = argmax_c\left(\sum_{n=1}^{N} I_{h_n(y)=c}\right) \quad (5)$$

Dimana I adalah fungsi indikator dan *hn* adalah *tree* ke-n dari *Random Forest.*

D. *Bussiness Understanding*

Tujuan dari penelitian ini adalah mendapatkan pengetahuan dan informasi dari data siswa yang merasa sangat cocok, cocok, kurang cocok, dan sangat kurang cocok dengan jurusan yang dipilih oleh siswa. Tujuan dari pengolahan data tersebut adalah membuat prediksi yang mampu membantu pemangku kebijakan di SMK Antartika 1 Sidoarjo untuk mengevaluasi dan meningkatkan pelayanan pada siswa baru. Sistem ini diharapkan mampu dalam merekomendasikan pemilihan jurusan yang cocok sesuai bakat dan minat siswanya.

E. *Data Understanding*

Data siswa SMK Antartika 1 Sidoarjo dibagi menjadi 2 dua bagian yaitu: Data Primer dan Data Sekunder. Data primer siswa diperoleh dari proses penyebaran kuesioner minat pada responden yang terdiri dari 578 siswa siswi SMK Antartika 1 Sidoarjo dengan menggunakan skala *likert* 1-10 dalam melakukan perhitungan dan pengamatan tentang minat siswa terhadap jurusannya. Sedangkan data sekunder siswa didapatkan dengan mengumpulkan informasi tentang latar belakang siswa seperti: jurusan, pendidikan orang tua, penghasilan orang tua, kendaraan ke sekolah, nilai ujian sekolah saat SMP, nilai karakter kedisiplinan, nilai karakter tanggung tanggung jawab, nilai karakter sikap, dan nilai karakter komunikasi.

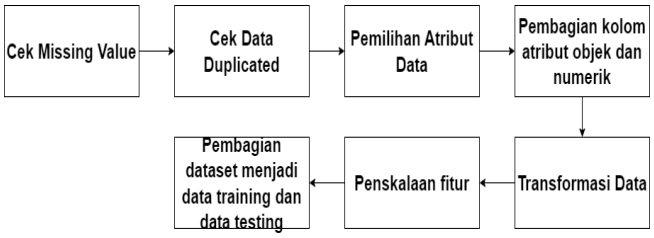
Tabel 1. *Dataset* Siswa SMK Antartika 1 Sidoarjo

Nama Atribut	Deskripsi	Tipe Data	Nilai
Nama Siswa	Nama siswa SMK Antartika 1 Sidoarjo	*object*	-
Kelas	Kelas siswa di SMK Antartika 1 Sidoarjo	*object*	X RPL, XI RPL, XII RPL, X TKR, XI TKR, XII TKR, X TPM, XI TPM, XII TPM
Jurusan	Jurusan siswa SMK Antartika 1 Sidoarjo	*object*	Teknik Pemesinan (TPm),

			Teknik Kendaraan Ringan (TKR), Rekayasa Perangkat Lunak(RPL)
Jenis Kelamin	Jenis kelamin siswa	*object*	Pria | Wanita
Pendidikan Ayah	Pendidikan terakhir ayah dari siswa	*object*	0#TidakTamat SD 1 # SD/MI 2# SMP/MTs 3#SMA/SMK/MAK 4 #Diploma 5# Sarjana 6#Magister/ Doktoral
Penghasilan Ayah	Penghasilan ayah siswa dalam satu bulan	*object*	0# < 1.000.000 1# 1.000.000 – 2.500.000 2# 2.500.000 – 3.500.000 3# 3.500.000 – 4.500.000 4# > 4.500.000
Pendidikan Ibu	Pendidikan terakhir ibu dari siswa	*object*	0#TidakTamat SD 1 # SD/MI 2# SMP/MTs 3#SMA/SMK 4 #Diploma 5# Sarjana 6#Magister/ Doktoral
Transportasi Sekolah	Alat transportasi yang digunakan siswa ke sekolah	*object*	0#Jalan Kaki 1#Sepeda 2#SepedaMotor 3#Mobil Pribadi 4#Kendaraan Umum
NUS_MTK_SMP	Nilai matematika siswa saat ujian sekolah di SMP	*numerik*	1-100
NUS_BIND_SMP	Nilai Bahasa Indonesia siswa saat ujian sekolah di SMP	*numerik*	1-100
NUS_BING_SMP	Nilai Bahasa Inggris siswa saat ujian sekolah di SMP	*numerik*	1-100
Disiplin	Nilai karakter kedisiplinan siswa	*numerik*	1-100
Tanggung jawab	Nilai karakter tanggung jawab siswa	*numerik*	1-100
Sikap	Nilai karakter sikap siswa	*numerik*	1-100
Komunikasi	Nilai karakter komunikasi siswa	*numerik*	1-100
P1 – P15	Nilai kuesioner kecocokan pemilihan jurusan siswa	*numerik*	1-10
Output kelas	Hasil kecocokan pemilihan jurusan siswa	*object*	0#Sangat KurangCocok 1#KurangCocok 2#Cocok 3#SangatCocok

Dataset siswa yang digunakan dalam penelitian ini adalah data siswa aktif pada tahun ajaran 2021-2022, 2022-2023 yang memiliki ukuran 578 *record* data siswa.

F. *Data Preparation*



Gambar 2. *Workflow* Proses *Data Preparation*

Pada Gambar 2, tahapan *data preparation* dibagi menjadi beberapa tahap yaitu, cek *missing value* diperlukan untuk memastikan *dataset* bersih dari data kosong/*missing*. Jika masih ada *missing value*, maka harus dilakukan *impute missing value* agar tidak menganggu proses selanjutnya. Proses selanjutnya adalah mengecek data *duplicate* pada

dataset. Setelah itu adalah memilah *dataset* dengan atribut data yang dipakai dan tidak dipakai. Dalam *dataset* siswa terdapat 31 atribut siswa yang terdiri dari: Nama Siswa, Jurusan, Kelas, Jenis Kelamin, Pendidikan Ayah, Penghasilan Ayah, Pendidikan Ibu, Transportasi ke Sekolah, NUS_MTK_SMP, NUS_BIND_SMP, NUS_BING_SMP, DISIPLIN, TANGGUNG JAWAB, SIKAP,KOMUNIKASI, P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, dan Output_Kelas.

Dari 31 atribut di atas akan dipilih atribut yang sesuai untuk digunakan dalam pemilihan fitur, maka dipilih 13 atribut yang akan digunakan untuk membuat model *data mining* untuk memprediksi kecocokan siswa dalam pemilihan jurusan di SMK Antartika 1 Sidoarjo yang sesuai dengan minatnya. Atribut yang digunakan adalah sebagai berikut: Jurusan, Pendidikan_Ayah, Penghasilan_Ayah, Pendidikan_Ibu,Transportasi_ke_Sekolah,NUS_MTK_SMP, NUS_BIND_SMP, NUS_BING_SMP, Disiplin, Tanggung_Jawab, Sikap, Komunikasi, dan Output_Kelas.

Setelah atribut yang akan digunakan untuk *modeling data* ditentukan, maka selanjutnya dilakukan proses pembagian kolom data kategoris dan kolom data numerik, karena saat pra proses data tersebut akan ditangani secara terpisah.

Adapun kolom data kategoris terdiri dari, Jurusan, Jenis Kelamin, Pendidikan Ayah, Pendidikan Ibu, Penghasilan Ayah, Pendidikan Ibu, Transportasi ke Sekolah, dan Output Kelas, sedangkan data numerik terdiri dari kolom, NUS_MTK_SMP, NUS_BIND_SMP, NUS_BING_SMP, Disiplin, Tanggung_Jawab, Sikap, dan Komunikasi.

Setelah kolom data kategoris dan numerik dibagi maka pra proses data akan dilanjutkan pada proses transformasi data dengan mengunakan *label encoding* pada atribut yang bersifat kategoris. Tujuan diubahnya data kategoris menjadi menjadi angka agar data kategoris dapat diproses oleh algoritma *machine learning.*

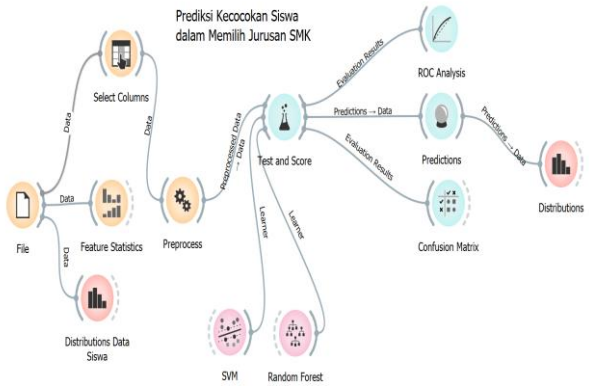
Selanjutnya setelah proses transformasi data kolom kategoris dilakukan, maka akan dilakukan proses penskalaan fitur/*feature selection*. Proses ini digunakan untuk menempatkan fitur pada skala yang sama, penskalaan fitur digunakan algoritma *machine learning* untuk menghitung jarak antar data, jika tidak diskalakan fitur dengan rentang nilai yang lebih tinggi akan mendominasi dalam penghitungan jarak. Metode *Support Vector Machine* membutuhkan penskalaan fitur terlebih dulu, sedangkan metode *Random Forest* tidak memerlukan penskalaan fitur.

Proses selanjutnya adalah membuang atribut Output_Kelas yang akan menjadi kelas target pada prediksi kecocokan jurusan siswa SMK, setelah atribut kelas target dipilih maka dilakukan pembagian *dataset* menjadi dua *set*: satu set untuk pelatihan dan satu *set* untuk pengujian.

Adapun pengaturan untuk pembagian *dataset* menjadi 80% data pelatihan dan 20% data pengujian.

G. *Modeling*

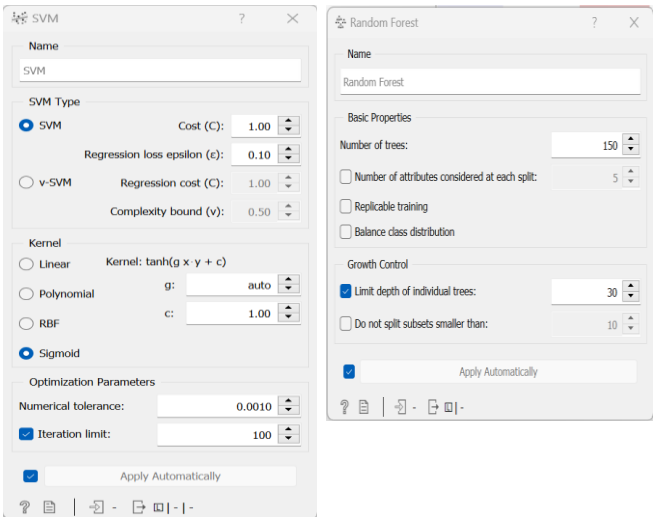
Pada tahap keempat ini, menggunakan perbandingan dua algoritma berbeda untuk melakukan prediksi kecocokan pemilihan jurusan di SMK Antartika 1 Sidoarjo, Adapun *workflow* dari modeling data pada algoritma *Random Forest* dan *Support Vector Machine* dengan menggunakan *software Orange* dapat dilihat pada Gambar 3:



Gambar 3. *Workflow Modeling Data*

Pengujian model menggunakan *k-fold cross validation* dengan jumlah *fold* sebanyak 10 fold. Parameter uji coba yang digunakan pada riset adalah *Cost* (C) = 1, *Regression loss epsilon* (Ɛ) = 0,10. *Kernel* SVM yang digunakan untuk penelitian ini adalah menggunakan kernel *sigmoid* (4) dan perulangannya dilakukan sebanyak 100 kali, sedangkan parameter uji coba pada algoritma *random forest* yang digunakan pada penelitian ini adalah *Number of trees = 150 (5), Limit depth of individual trees = 30.*

Parameter yang digunakan di dalam algoritma *Support Vector Machine* dan *Random Forest* dapat dilihat pada Gambar 4.



Gambar 4. Pengaturan Parameter pada *Support Vector Machine* dan *Random Forest*

Untuk menghitung hasil dari model SVM dan RF menggunakan nilai *Accuracy*, *F1 Score*, *Precision*, dan *Recall*. *Formula Accuracy* dapat dilihat pada (6), formula untuk *F1 Score* pada (7), formula untuk *Precision* pada (8), dan formula untuk *Recall* pada (9)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ x } 100 \quad (6)$$

$$F1\text{-}Score = 2 \text{ x } (Recall*Precission) / (Recall + Precision) \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \text{ x } 100 \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \text{ x } 100 \quad (9)$$

III. HASIL DAN ANALISIS

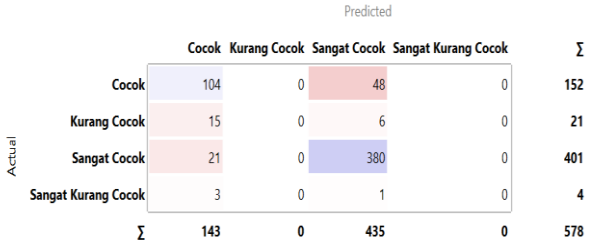
A. Evaluation

Eksperimen yang dilakukan pada penelitian ini menggunakan pengukuran keberhasilan algoritma klasifikasi dengan *confusion matrix*. Hasil pengujian model pada penelitian ini dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pengujian Model

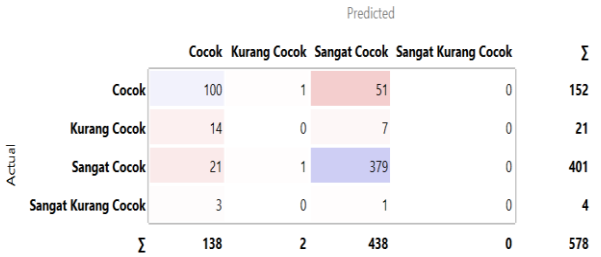
Model	*Accuracy*	F1	*Precision*	*Recall*
SVM	83	81	79	83
Random Forest	82	80	79	82

Dari Gambar 5, diketahui bahwa hasil prediksi kecocokan pemilihan jurusan siswa yang memanfaatkan metode *support vector machine* dengan kernel *sigmoid* menghasilkan nilai *TP(True Positive)* sebesar 484 *record* dari 578 *record* keseluruhan.



Gambar 5. *Confusion Matrix Support Vector Machine*

Dari Gambar 6, diketahui bahwa hasil prediksi kecocokan pemilihan jurusan siswa yang memanfaatkan metode *random forest* dengan jumlah pohon sebanyak 100 pohon keputusan, sehingga menghasilkan nilai *TP (True Positive)* sebesar 479 *record* dari total 578 *record* keseluruhan.



Gambar 6. *Confusion Matrix Random Forest*

IV. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan seperti yang telah terlihat, memprediksi kecocokan pemilihan jurusan pada siswa SMK Antartika 1 Sidoarjo menjadi tugas yang sulit untuk dilakukan apabila dilakukan dengan cara konvensional dengan melihat banyaknya parameter yang digunakan. Dengan pemanfaatan pengetahuan teknik data mining dalam dunia pendidikan atau disebut *Educational data mining* dapat digunakan sebagai langkah awal untuk memprediksi kecocokan jurusan siswa SMK Antartika 1 Sidoarjo sedini mungkin. Secara umum, kesimpulan utamanya adalah sebagai berikut:

1. Model prediksi untuk kecocokan pemilihan jurusan siswa SMK Antartika 1 Sidoarjo berhasil dibangun dengan metode *Random Forest* dan *Support Vector Machine* dengan beberapa *input* fitur yang kemudian menghasilkan prediksi kecocokan pemilihan jurusan siswa yang sesuai.
2. Penelitian ini membandingkan dua metode yang digunakan yaitu, *Random Forest dan Support Vector Machine.* Dari dua metode tersebut diperoleh hasil prediksi yang terbaik yaitu dengan metode *Support Vector Machine* dengan hasil *Accuracy* adalah 83%, *Precision* adalah 79%, *Recall* 83%, dan *F1-Score* 81% dan AUC 83%.
3. Dalam penelitian ini, metode *Random Forest* dan *Support Vector Machine* menggunakan metode *sampling* dengan *k-Fold Cross Validation* dengan jumlah *fold* adalah 10 *fold*, sehingga menghasilkan *Accuracy* pada *Support Vector Machine (SVM)* dan *Random Forest (RF)* masing-masing 83% dan 82%.
4. Dalam ketepatan prediksi dari 578 *record*, jumlah *record* siswa yang berhasil diprediksi menggunakan metode *Random Forest* dengan *True Positive (TP)* yaitu 479 *record*, sedangkan jumlah *record* siswa yang berhasil diprediksi menggunakan metode *Support Vector Machine* dengan *True Positive (TP)* yaitu 484 *record*. Dan artinya sistem yang dibangun dengan metode *Support Vector Machine (SVM)* dan *Random Forest (RF)* sudah sangat baik dalam memberikan saran terhadap kecocokan pemilihan jurusan pada siswa SMK Antartika 1 Sidoarjo.

REFERENSI

[1] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007, doi: 10.1016/j.eswa.2006.04.005.

[2] A. Jain, K. Shah, P. Chaturvedi, and A. Tambe, "Prediction and Analysis of Student Performance using Hybrid Model of Multilayer Perceptron and Random Forest," *2018 Int. Conf. Adv. Comput. Telecommun. ICACAT 2018*, pp. 1–7, 2018, doi: 10.1109/ICACAT.2018.8933580.

[3] D. De La Peña, J. A. Lara, D. Lizcano, M. A. Martínez, C. Burgos, and M. L. Campanario, "Mining activity grades to model students' performance," *Proc. - 2017 Int. Conf. Eng. MIS, ICEMIS 2017*, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/ICEMIS.2017.8272963.

[4] I. Burman and S. Som, "Predicting Students Academic Performance Using Support Vector Machine," *Proc. - 2019 Amity Int. Conf. Artif. Intell. AICAI 2019*, pp. 756–759, 2019, doi: 10.1109/AICAI.2019.8701260.

[5] C. C. Kiu, "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities," *Proc. - 2018 4th Int. Conf. Adv. Comput. Commun. Autom. ICACCA 2018*, pp. 1–5, 2018, doi: 10.1109/ICACCAF.2018.8776809.

[6] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 243–247, 2019, doi: 10.1109/COMITCon.2019.8862214.

[7] Rusdiansyah, "Analisis Keputusan Menentukan Jurusan Pada Sekolah Menengah dengan Netode Simple Additive Weighting," *J. Techno Nusa Mandiri*, vol. XIV, no. 1, pp. 49–56, 2017.

[8] X. Ma and Z. Zhou, "Student pass rates prediction using optimized support vector machine and decision tree," *2018 IEEE 8th Annu. Comput. Commun. Work. Conf. CCWC 2018*, vol. 2018-Janua, pp. 209–215, 2018, doi: 10.1109/CCWC.2018.8301756.

[9] M. Zaffar, M. A. Hashmani, and K. S. Savita, "Performance analysis of feature selection algorithm for educational data mining," *2017 IEEE Conf. Big Data Anal. ICBDA 2017*, vol. 2018-Janua, pp. 7–12, 2018, doi: 10.1109/ICBDAA.2017.8284099.

[10] M. Ross, C. A. Graves, J. W. Campbell, and J. H. Kim, "Using support vector machines to classify student attentiveness for the development of personalized learning systems," *Proc. - 2013 12th Int. Conf. Mach. Learn. Appl. ICMLA 2013*, vol. 1, pp. 325–328, 2013, doi: 10.1109/ICMLA.2013.66.

[11] M. Y. Arafath, M. Saifuzzaman, S. Ahmed, and S. A. Hossain, "Predicting career using data mining," *2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018*, pp. 889–894, 2019, doi: 10.1109/GUCON.2018.8674995.

[12] P. Rojanavasu, "Educational data analytics using association rule mining and classification," *ECTI DAMT-NCON 2019 - 4th Int. Conf. Digit. Arts, Media Technol. 2nd ECTI North. Sect. Conf. Electr. Electron. Comput. Telecommun. Eng.*, pp. 142–145, 2019, doi: 10.1109/ECTI-NCON.2019.8692274.

[13] D. P. Kusumaningrum, N. A. Setiyanto, E. Y. Hidayat, and K. Hastuti, "Recommendation System for Major University Determination Based on Student's Profile and Interest," *J. Appl. Intell. Syst.*, vol. 2, no. 1, pp. 21–28, 2017, doi: 10.33633/jais.v2i1.1389.

[14] T. K. Spss and T. R. Daimlerchrysler, "Crisp-dm 1.0," pp. 1–78.

[15] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," *Proc. - 2nd Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2016*, 2017, doi: 10.1109/ICCUBEA.2016.7860040.

[16] S. Jeganathan, P. M. A. Kumar, S. Parthasarathy, and A. R. Lakshminarayanan, "Predicting the Post Graduate Admissions using Classification Techniques," pp. 346–350, 2021.