Implementasi Orange Untuk Klasifikasi Kepribadian Siswa-Siswi SMAN 79 Jakarta Dengan Model *Random Forest* dan *Naive Bayes*

Septian Pratama
Teknik Informatika, Program Pascasarjana, Universitas Pamulang
e-mail: tian_yama@yahoo.com

***Abstrak*—Proses pemantauan dan evaluasi terhadap kepribadian siswa-siswi SMAN 79 Jakarta sangat diperlukan, karena akan sangat memudahkan sekolah dalam membentuk karakter peserta didik. Hal tersebut diimplementasikan dalam bentuk pelatihan analisis kepribadian. Hasil analisis inilah yang digunakan sekolah untuk mengklasifikasikan kepribadian peserta didik. *Data mining* bisa digunakan untuk proses klasifikasi tersebut. Penelitian ini bertujuan untuk menerapkan aplikasi *orange data mining* dengan menggunakan model *Random Forest* dan *Naive Bayes* yang selanjutnya akan dilakukan evaluasi akurasi dari masing-masing model tersebut. Hasil dari penelitian ini adalah model *Random Forest* memiliki nilai *accuracy* 95%, *precision* 95% dan *recall* 95% sedangkan Naive Bayes memiliki nilai *accuracy* 75%, *precision* 88% dan *recall* 75%. Dari kedua model tersebut menunjukkan bahwa model *Random Forest* lebih baik dibandingkan dengan model *Naive Bayes dalam melakukan klasifikasi analisis kepribadian siswa-siswi SMAN* SMAN 79 Jakarta.**

Kata Kunci*—Kepribadian; Klasifikasi; *Random Forest; Naive Bayes

I. PENDAHULUAN

Sekolah merupakan salah satu tempat yang efektif bagi pembentukan karakter seorang individu [1]. SMAN 79 Jakarta telah berupaya keras untuk membangun karakter siswa-siswinya. Salah satu cara yang dilakukan adalah dengan mengadakan pelatihan analisis kepribadian menggunakan konsep *FSQ Personalitree*. Konsep ini merupakan sebuah tes analisis kepribadian melalui struktur yang ada pada pohon mencakup akar, buah, cabang dan daun [2]. Untuk penerapannya siswa-siswi diarahkan mengisi *questionnaire* yang berisikan 28 pertanyaan. Jawaban dari pertanyaan tersebut, yang nantinya akan mengklasifikasikan kepribadian setiap peserta. Hasil dari kepribadian tersebut akan Penulis uji dalam pemodelan *machine learning* menggunakan aplikasi *orange*.

Penelitian ini bertujuan untuk ketepatan klasifikasi kepribadian siswa-siswi SMAN 79 Jakarta dengan menerapkan dua metode yaitu *Random Forest* dan *Naive Bayes*. Metode *Random Forest* adalah algoritma belajar pada mesin yang memiliki banyak pohon keputusan, dengan kombinasi dari model *bagging* dan *random sub spaces*, metode *Random Forest* ini telah membuktikan keberhasilannya dalam melakukan prediksi dan klasifikasi pada beberapa tahun terakhir dan menjadi salah satu algoritma *marchine learning* terbaik yang digunakan pada berbagai bidang [3]. Sedangkan *Metode Naïve bayes (Naïve Bayes Classifier*) salah satu metode untuk menentukan nilai probabilitas dalam memprediksi peluang dengan mengunakan data pengalaman sebelumnya [4]. Selanjutnya dari kedua metode tersebut, akan dilakukan analisis perbandingan dua model dengan menerapkan *Confusion Matrix* dan *ROC* untuk memastikan tingkat akurasi.

II. DASAR TEORI

Data mining merupakan bagian dari proses penemuan pengetahuan dari basis data *Knowledge Discovery in Databases* (Alkhairi & Windarto, 2019) [5]. Klasifikasi merupakan sebuah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep dan kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui [6]. Algoritma klasifikasi hendak menciptakan sekumpulan ketentuan yang disebut rule yang hendak digunakan selaku penanda buat bisa memprediksi kelas dari informasi yang mau diprediksi [7].

A. Random Forest

Random Forest adalah pengembangan dari metode *Decision Tree* yang menggunakan beberapa *Decision Tree*, dimana setiap *Decision Tree* telah dilakukan pelatihan menggunakan sampel individu dan setiap atribut dipecah pada pohon yang dipilih antara atribut subset yang bersifat acak. *Random Forest* memiliki beberapa kelebihan, yaitu dapat meningkatkan hasil akurasi jika terdapat data yang hilang dan untuk *resisting outliers*, serta efisien untuk penyimpanan sebuah data. Selain itu, *Random Forest* mempunyai proses seleksi fitur dimana mampu mengambil fitur terbaik sehingga dapat meningkatkan performa terhadap model klasifikasi [8].

B. Naive Bayes

Naive Bayes ialah tata cara pengklasifikasian probabilistik simpel. Model ini hendak menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi serta campuran nilai dari dataset yang diberikan [7]. Keuntungan pemakaian Naive Bayes ialah hanya membutuhkan beberapa kecil informasi data latih, memastikan parameter mean, serta varians dari variabel yang dibutuhkan untuk klasifikasi. *Teorema Bayes* mempunyai wujud universal selaku berikut [9].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Keterangan :

X : Data dengan class yang belum diketahui
H : Hipotesis data merupakan suatu class spesifik
P(H|X) : Probabilitas hiotesis H berdasar kondisi X (Posteriori Probabilitas)
P(H) : Probabilitas Hipotesis H (Prior Probabilitas)
P(X|H) : Probabilitas X berdasarkan kondisi hipotesis H
P(X) : Probabilitas X

C. Elemen Pengukuran

Confusion Matrix merupakan pengukuran performa untuk permasalahan klasifikasi *machine learning* dimana luaran bisa berbentuk 2 kelas ataupun lebih. *Confusion Matrix* merupakan tabel dengan 4 campuran berbeda dari nilai prediksi serta nilai aktual. Berdasarkan tabel 1 dapat dijelaskan bahwa terdapat 4 sebutan representasi hasil proses klasifikasi pada *confusion matrix* ialah *True Positif (TP), True Negatif (TN), False Positif (FP),* serta *False Negatif (FN)* [7], [10].

Tabel 1.
Confusion Matrix

Kelas Sebenarnya	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Pengukuran dalam evaluasi metode yang digunakan antara lain *accuracy*, *precision* dan *recall*. *Accuracy* direpresentasikan dengan *confusion matrix* untuk membandingkan antara kelas prediksi dengan kelas sebenarnya. Sementara *precision* menunjukkan perbandingan antara peristiwa sebenarnya.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

III. METODE PENELITIAN

Penelitian ini bertujuan untuk melakukan analisa perbadingan metode *Random Forest* dan *Naive Bayes* yang digunakan untuk mengklasifikasi kepribadian siwa-siswi SMAN 79 Jakarta. Aplikasi yang digunakan untuk simulasi adalah *Orange Data Mining* yaitu aplikasi *data mining open source* yang terbukti mampu membantu peneliti untuk menganalisa datanya. Tahapan proses pada penelitian ini bisa dilihat pada Gambar 1.

Gambar 1.
Tahapan Penelitian

Sesuai Gambar 1 tersebut, langkah pertama adalah identifikasi masalah, perumusan dan kajian pustaka hal ini dilakukan untuk menyusun tujuan riset dan kontribusi riset. Kedua adalah proses pengumpulan data dilanjutkan dengan menyusun data latih dan data uji sebagai sumber dari klasifikasi data. Ketiga adalah proses perancangan model menggunakan aplikasi *orange data mining* untuk proses klasifikasi kepribadian siswa-siswi dan perbandingan metode. Keempat adalah proses klasifikasi kepribadian siswa-siswi SMAN 79 Jakarta mengggunakan *Random Forest* dan *Naive Bayes*. Kelima adalah proses evaluasi kinerja metode klasifikasi dan menganalisa hasil perbandingan dari metode yang digunakan.

A. Atribut Penelitian

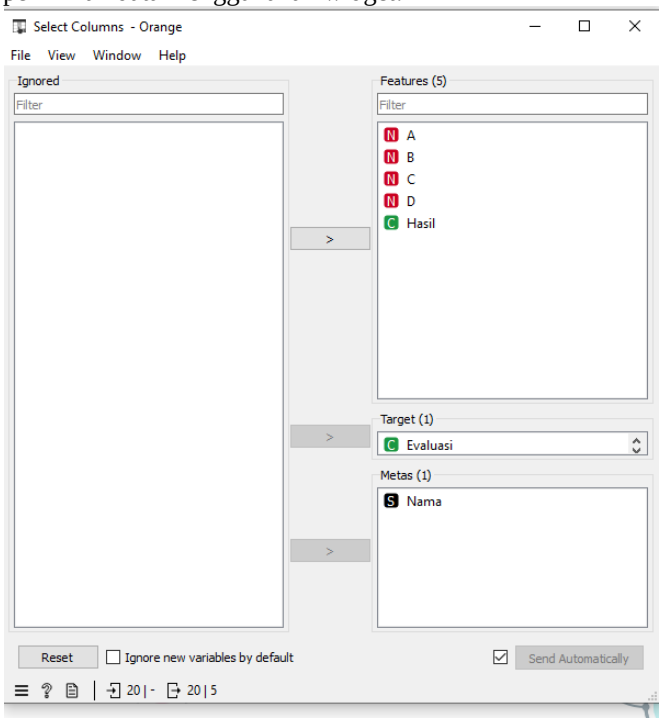
Kumpulan data siswa-siswi SMAN 79 Jakarta berupa data primer diperoleh melalui wawancara melalui *Character Building Training* dengan banyak data 100 orang. Data *training* sebanyak 80 orang dan data *testing* sebanyak 20 orang. Berikut ini adalah deskripsi dataset dijelaskan pada Tabel 2.

Tabel 2.
Attribute Data Siswa-Siswi

No	*Attribute*	*Type*	*Description*
1	Nama	Text	Nama Siswa/i
2	A	Numeric	Nilai Kepribadian Akar
3	B	Numeric	Nilai Kepribadian Buah
4	C	Numeric	Nilai Kepribadian Cabang
5	D	Numeric	Nilai Kepribadian Daun
6	Hasil	Categorical	Nilai Tertinggi Dari Keempat Kepribadian
7	Evaluasi	Categorical	Benar / Salah dari hasil prediksi metode

B. Data Selection Process / Prepocessing

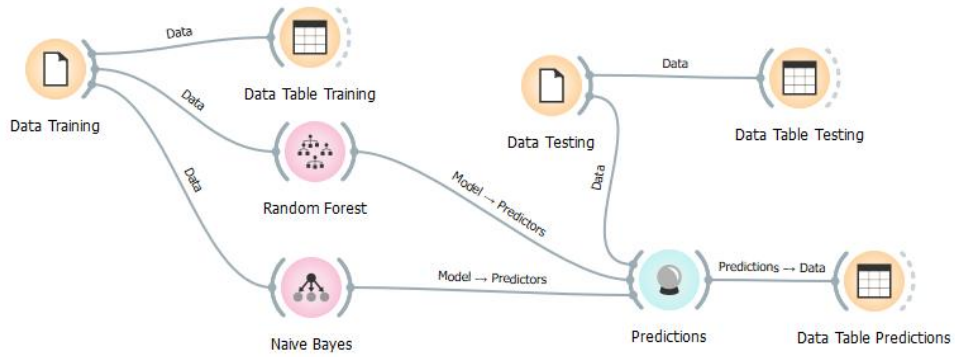
Pada saat proses *prepocessing* dataset siswa-siswi SMAN 79 ini diambil data yang memiliki satu dominan nilai kepribadian tertinggi. Berikut ini adalah proses pemilihan data menggunakan widget.



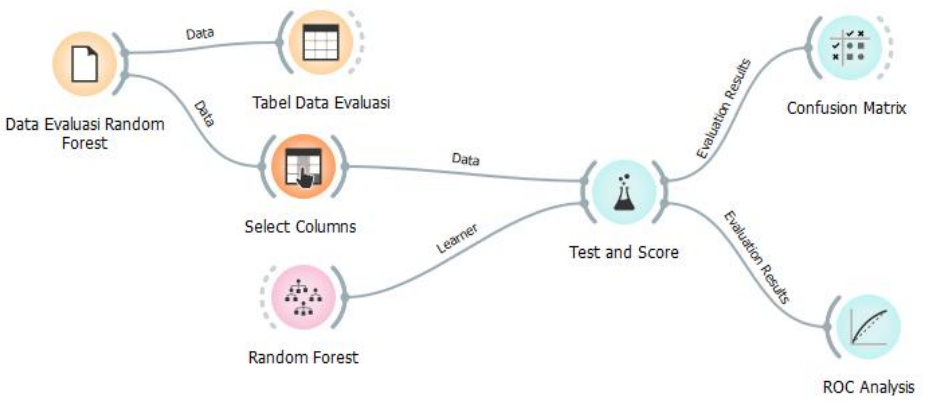
Gambar 2.
Selection Process

C. Data Mining Process

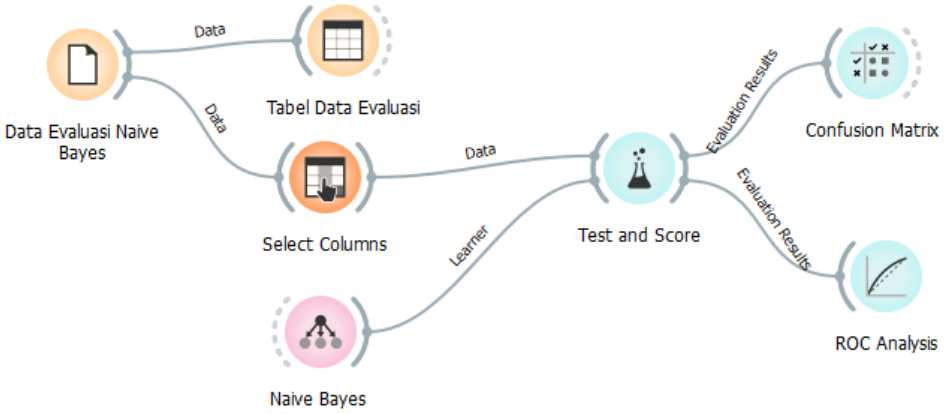
Dalam menganalisa performa beberapa model klasifikasi menggunakan orange, dilakukan perbandingan dari metode yang digunakan untuk memilih metode yang terbaik dengan akurasi yang tinggi, dalam mengklasifikasi dataset kepribadian siswa-siswi. Sebagaimana terlihat pada Gambar 3, Gambar 4 dan Gambar 5.



Gambar 3.
Prediksi Klasifikasi Kepribadian



Gambar 4.
Evaluasi Metode *Random Forest*



Gambar 5.
Evaluasi Metode Naive Bayes

IV. HASIL DAN PEMBAHASAN

A. Hasil Simulasi Model

Hasil simulasi model klasifikasi dilakukan dengan menggunakan kumpulan data uji dengan 1 atribut sebagai target, 5 atribute numeric yaitu A, B, C, D dan hasil analisis. Sehingga diperoleh hasil *test score* seperti terlihat pada Gambar 6 dan Gambar 7.

Evaluation results for target (None, show average over classes)

Model	AUC	CA	Prec	Recall
Random Forest	0.980	0.950	0.953	0.950

Gambar 6.
Test Score Ramdom Forest

Evaluation results for target (None, show average over classes)

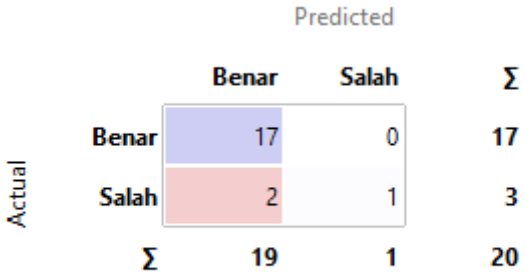
Model	AUC	CA	Prec	Recall
Naive Bayes	0.930	0.750	0.889	0.750

Gambar 7.
Test Score Naive Bayes

Berdasarkan Gambar 6 dan Gambar 7 maka dapat dilihat model *Random Forest* mempunyai nilai *Classification Accuracy* (*CA*) sebesar 95%, *Precision (Prec)* sebesar 95%, *Recall* sebesar 95% dan *Area Under ROC Curve (AUC)* sebesar 98%. Sementara pada model *Naive Bayes* mempunyai nilai *Classification Accuracy* (*CA*) sebesar 75%, *Precision (Prec)* sebesar 88%, *Recall* sebesar 75% dan *Area Under ROC Curve (AUC)* sebesar 93%. Dari hasil tersebut model *Random Forest* lebih unggul dari pada model *Naive Bayes. AUC* digunakan untuk mengukur kinerja diskriminatif dengan memperkirakan probabilitas output dari ilustrasi yang diseleksi secara acak pada populasi positif ataupun negatif. Semakin besar *AUC*, maka semakin baik hasil klasifikasi yang digunakan [7].

B. Hasil Evaluasi dengan Confussion Matrix

Confusion Matrix merupakan pengukuran performa untuk permasalahan klasifikasi *machine learning* dimana keluaran bisa berbentuk dua kelas ataupun lebih. *Confusion Matrix* merupakan tabel dengan empat campuran berbeda dari nilai prediksi serta nilai actual [7]. Berikut hasil *Confusion Matrix* dari model *Random Forest* dan *Naive Bayes.*



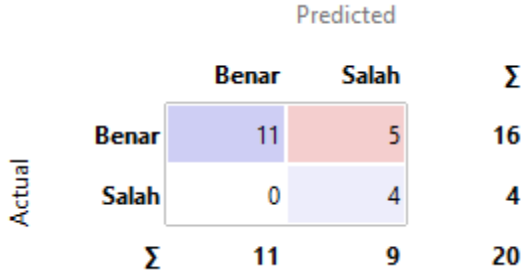
Gambar 8.
Confusion Matrix Model *Random Forest*

Pada Gambar 8 menunjukkan bahwa nilai dari *True Positif (TP)* adalah 17, *True Negatif (TN)* adalah 1, *False Positif (FP)* adalah 2, dan *False Negatif (FN)* adalah 0. Maka nilai *Accuracy*, *Precision* dan *Recall* dari metode *Random Forest* adalah sebagai berikut:

$$Accuracy = \frac{17+1}{17+1+2+0} \times 100\%, maka\ nilai\ accuracy\ 90\%$$

$$Precision = \frac{17}{17+2} \times 100\%, maka\ nilai\ precision\ 89\%$$

$$Recall = \frac{17}{17+0} \times 100\%, maka\ nilai\ recall\ 100\%$$



Gambar 9.
Confusion Matrix Model *Naive Bayes*

Pada Gambar 9 menunjukkan bahwa nilai dari *True Positif (TP)* adalah 11, *True Negatif (TN)* adalah 4, *False Positif (FP)* adalah 0, dan *False Negatif (FN)* adalah 5. Maka nilai *Accuracy*, *Precision* dan *Recall* dari metode *Naive Bayes* adalah sebagai berikut:

$$Accuracy = \frac{11+4}{11+4+0+5} \times 100\%, maka\ nilai\ accuracy\ 75\%$$

$$Precision = \frac{11}{11+0} \times 100\%, maka\ nilai\ precision\ 100\%$$

$$Recall = \frac{17}{17+5} \times 100\%, maka\ nilai\ recall\ 77\%$$

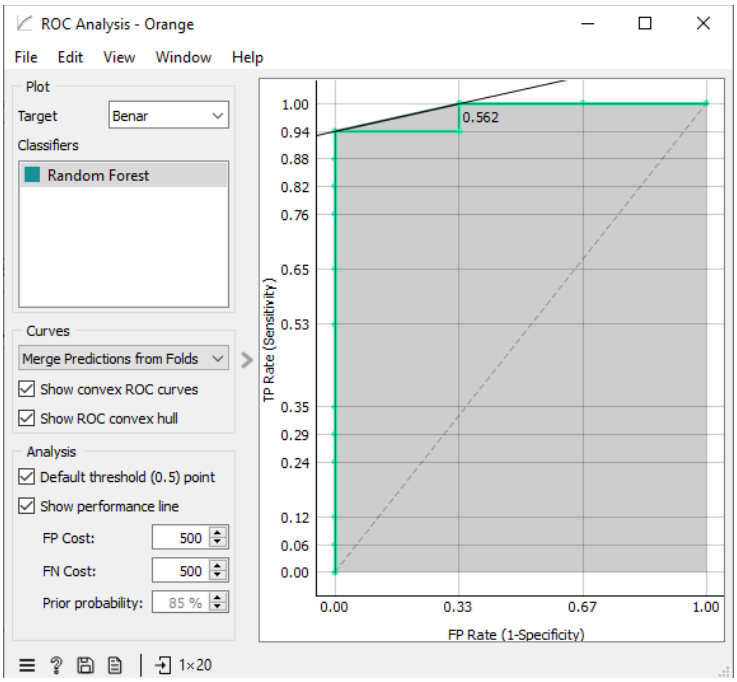
Tabel 3.
Perbandingan Kinerja Model

Metode	Accuracy	Precision	Recall
Random Forest	90%	89%	100%
Naive Bayes	75%	100%	77%

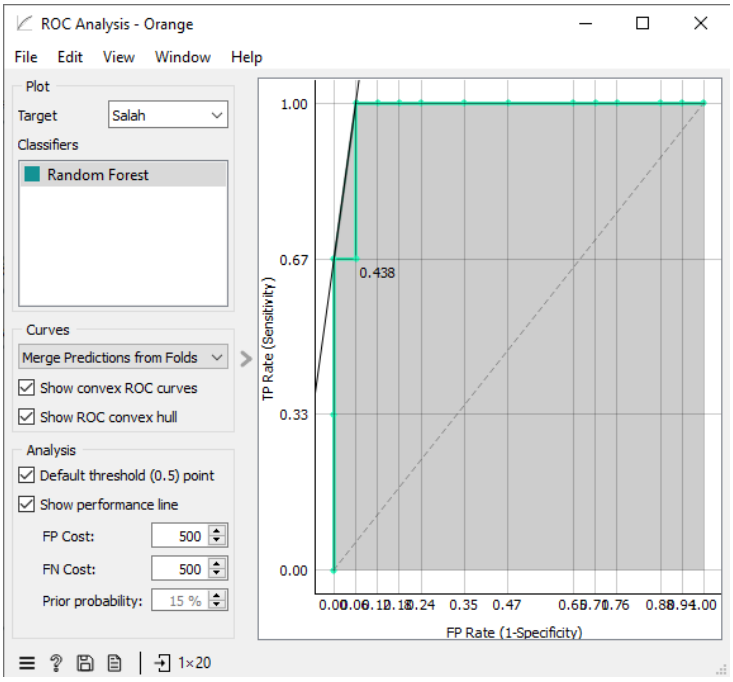
Berdasarkan Tabel 3 di atas meskipun nilai *Precision Naive Bayes* lebih tinggi dari *Random Forest*, namun secara keseluruhan dapat diketahui bahwa kinerja dari model *Random Forest* lebih baik dari model *Naive Bayes.*

C. Hasil Evaluasi dengan ROC Curve

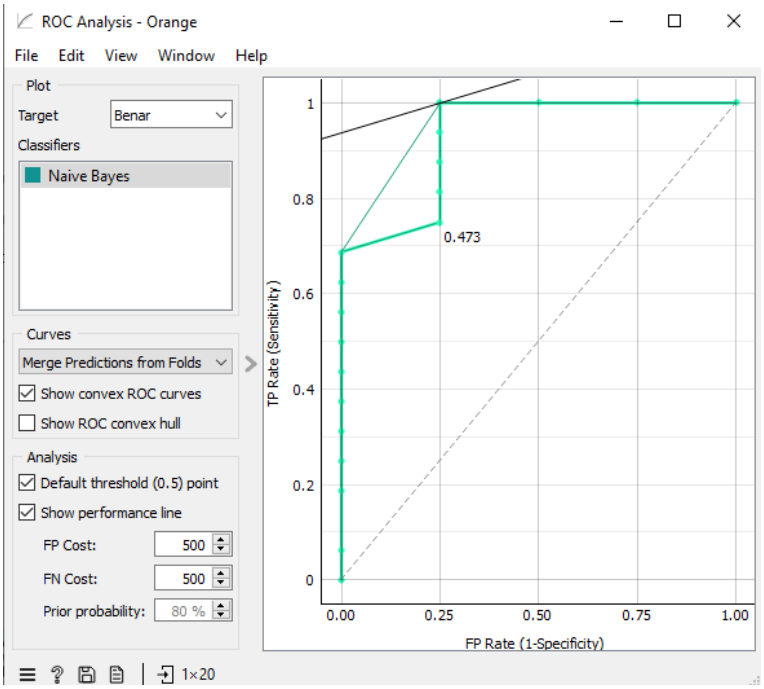
Nilai akurasi secara manual bisa dilakukan dengan melihat perbandingan *curve ROC* yang divisulisasi dari *Confusion Matrix* [7]. Model kurva *ROC* merupakan cara yang paling mudah terlihat untuk membandingkan nilai akurasi masing-masing model klasifikasi secara grafis. Semakin dekat kurva mengikuti batas kiri dan batas atas ruang *ROC*, maka semakin akurat *classifier* tersebut. Berikut adalah gambar kurva dari model *Random Forest dan Naive Bayes.*



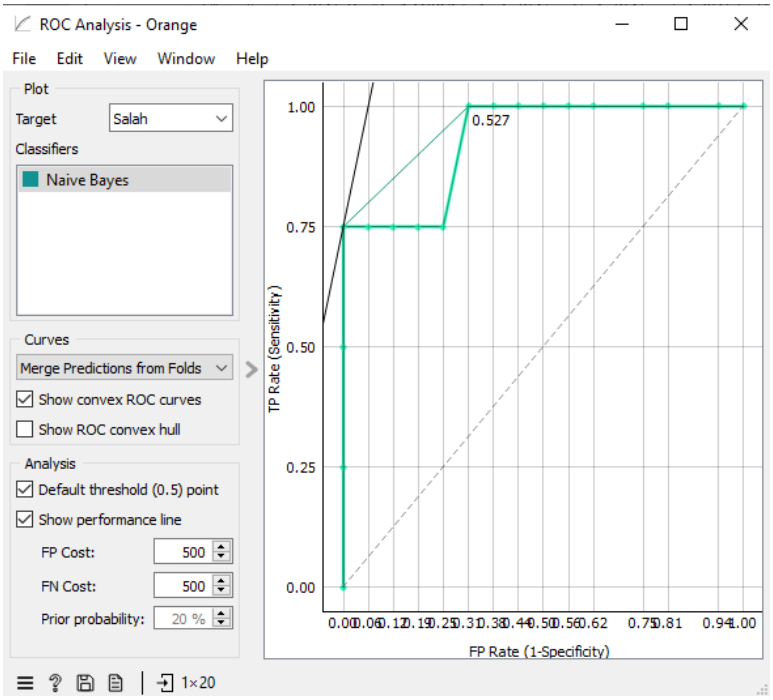
Gambar 10.
Analisis *ROC Random Forest* dengan target Benar



Gambar 11.
Analisis *ROC Random Forest* dengan target Salah



Gambar 12.
Analisis *ROC Naive Bayes* dengan target Benar



Gambar 13.
Analisis *ROC Naive Bayes* dengan target Salah

Dari gambar 10 dan 11 dapat dilihat kurva lebih dekat mengikuti batas kiri dan batas atas ruang *ROC* dibandingan dengan gambar 12 dan 13. Dengan demikian klasifikasi analalis kepribadian dengan model *Random Forest* lebih baik dibandingkan dengan model *Naive Bayes.*

V. KESIMPULAN

Setelah menggunakan model *Random Forest* dan *Naive Bayes* untuk mengklasifikasi tipe kepribadian siswa-siswi SMAN 79 Jakarta diperoleh hasil *accuracy*, *precision* dan *recall* model *Random Forest* lebih unggul dari pada akurasi *Naive Bayes*. Dapat dilihat dari 20 data uji yang digunakan, *Random Forest* memiliki nilai accuracy 95%, *precision* 95% dan *recall* 95% sedangkan Naive Bayes memiliki nilai accuracy 75% , *precision* 88% dan *recall 75%*. Kontribusi riset ini bisa digunakan oleh guru bimbingan konseling untuk memudahkan dalam klasifikasi kepribadian siswa-siswi SMAN 79 Jakarta.

DAFTAR PUSTAKA

[1] Supraptiningrum, & Agustini. (2015). Membangun Karakter Siswa Melalui Budaya Sekolah Di Sekolah Dasar. Jurnal Pendidikan Karakter (2), 219–228. https://journal.uny.ac.id/index.php/jpka/article/view/8625/7118
[2] P. Septian, S. Agung Ferdinan dan A. Aulia Ar Rakhman, "Perancangan Sistem Analisis Karakter Siswa-Siswi SMAN 79 Jakarta Berbasis Java", JUPTI Vol 1 No. 3 September (2022)570–578. https://ejurnal.stie-trianandra.ac.id/index.php/jupti/article/view/597/428
[3] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," Sistemasi, Vol. 10, No. 1, p. 163, 2021, doi: 10.32520/stmsi.v10i1.1129.
[4] D. Tuhenay, "Perbandingan Klasifikasi Bahasa Menggunakan Metode Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM)," JIKO (Jurnal Inform. dan Komputer), Vol. 4, No. 2, pp. 105–111, 2021, doi: 10.33387/jiko.v4i2.2958.
[5] Alkhairi, P., & Windarto, A. P. (2019). Penerapan K-Means Cluster pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara. Seminar Nasional Teknologi Komputer & Sains, 762–767.
[6] B. Irwan, Muliadi & R. Retma, "Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi", Jurnal Jupiter Vol. 7 No. 1 ,April 2015, 39-50.
[7] Hoziri, Anwari & Alim Syariful, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes", Jurnal Ilmiah NERO Vol. 6 No. 2, 133-144.
[8] S. Riki, G. Windu, A. Fauzi & M. Nurlaelatul, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah", Vol.13, No.2, Desember 2020, pp. 67 – 75.
[9] Bustami, "Penerapan Algoritma Naive Bayes Untuk Nasabah Asuransi," J. Inform., vol. 8, no. 1, pp. 884–898, 2014.
[10] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naive Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," J. Inform. Univ. Pamulang, vol.5, no. 4, p. 646, 2021.