e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 Agustus 2022

Penerapan Klasifikasi *Random Forest* Terhadap Data Gangguan Spektrum Autisme (ASD) Pada Anak – Anak Menggunakan Seleksi Fitur *Principal Component Analysis*

Luthfiyah Amatullah[1], Yuni Widiastiwi[2], Nurul Chamidah[3]
Program Studi Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati , Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia
Email: luthfiyaha@upnvj.ac.id[1], widiastiwi@yahoo.com[2], nurul.chamidah@upnvj.ac.id[3]

Abstrak. *Autistic Spectrum Disorder* (ASD) merupakan gangguan perkembangan fungsi otak yang kompleks dan sangat bervariasi. Kelainan ini secara signifikan berpengaruh terhadap komunikasi verbal, non-verbal serta interaksi sosial. Gejala umum gangguan ini biasanya akan terlihat pada anak-anak mulai sejak usia dua tahun. Tujuan penelitian ini untuk menerapkan metode seleksi fitur *Principal Component Analysis* (PCA) pada data penelitian *Autistic Spectrum Disorder Screening Data for Children Data Set* yang diperoleh dari *University of California Irvine* (UCI) *Machine Learning Data Repository* dimana PCA berfungsi untuk mereduksi dimensi data dari dataset serta mengklasifikasikannya menggunakan pemodelan klasifikasi *Random Forest*. Selain itu, untuk mengetahui bagaimana hasil evaluasi (*confusion matrix*) serta bagaimana perbedaan yang terdapat pada hasil evaluasi tersebut terhadap dataset yang menggunakan metode seleksi fitur *Principal Component Analysis* (PCA) dengan yang tidak menggunakan metode tersebut. Evaluasi dari hasil penelitian yang melalui seleksi fitur PCA yaitu menghasilkan nilai akurasi sebesar 98%, *precision* sebesar 96%, *recall* sebesar 100% dan *specificity* sebesar 96%. Sedangkan hasil evaluasi yang tanpa melalui proses PCA menghasilkan nilai akurasi sebesar 91%, *precision* sebesar 92%, *recall* sebesar 84% dan *specificity* sebesar 100%.

Kata Kunci: ASD, Gangguan Spektrum Autisme, Klasifikasi, *Random Forest*, *Principal Component Analysis*.

1 Pendahuluan

Autisme atau biasa disebut ASD (*Autistic Spectrum Disorder*) merupakan gangguan perkembangan fungsi otak yang kompleks dan sangat bervariasi (spektrum), biasanya gangguan ini meliputi cara berkomunikasi, berinteraksi sosial dan kemampuan berimajinasi [4]. Autisme dapat didiagnosis pada usia berapa saja, namun gejala umumnya akan terlihat pada anak-anak mulai sejak usia dua tahun. Berdasarkan *National Center on Birth Defects and Developmental Disabilities Centers for Disease Control and Prevention* (2021), Para ilmuwan dari Pusat Pengendalian dan Pencegahan Penyakit (CDC) dan Administrasi Sumber Daya dan Layanan Kesehatan (HRSA) menemukan bahwa 17% anak-anak yang berusia 3-17 tahun memiliki cacat perkembangan dan yang persentase ini meningkat dari tahun ke tahun [3]. Skrining merupakan salah satu upaya deteksi dini untuk mengidentifikasi penyakit atau kelainan yang secara klinis belum jelas. Dari data hasil skrining tersebut, dibutuhkan solusi untuk prediksi dengan melakukan klasifikasi. Seleksi fitur juga dibutuhkan untuk menyederhanakan dan menghilangkan fitur-fitur yang tidak memiliki korelasi yang besar antara satu sama lain dan yang kurang relevan namun tanpa menghilangkan informasi penting dari dataset aslinya guna bertujuan untuk meningkatkan performa suatu klasifikasi [1]. Oleh sebab itu, penulis mengusulkan untuk melakukan penelitian dengan judul Penerapan Klasifikasi *Random Forest* Terhadap Data Gangguan Spektrum Autisme (ASD) pada Anak-anak Menggunakan Metode Seleksi Fitur *Principal Component Analysis* (PCA) dengan Klasifikasi *Random Forest*, dimana metode seleksi fitur PCA akan diterapkan untuk mereduksi dimensi data dari dataset yang digunakan dan kemudian hasil dari proses PCA nantinya akan diklasifikasikan pada pemodelan klasifikasi *Random Forest*. Selain itu, ingin mengetahui juga apakah terdapat perbedaan pada performa klasifikasi terhadap data yang dengan menerapkan proses seleksi fitur PCA dengan yang tanpa melalui proses tersebut. Data yang digunakan pada penelitian ini diperoleh dari *University of California Irvine* (UCI) *Machine Learning Data Repository* yaitu *Autistic Spectrum Disorder Screening Data for Children Data Set*. Jumlah data yang diperoleh sebanyak 292 *record data* dengan 20 fitur dan 1 atribut target (terdapat 2 label kelas di dalamnya) [7].

2 Landasan Teori

2.1 Gangguan Spektrum Autisme (ASD)

Autistic Spectrum Disorder (ASD) merupakan gangguan perkembangan fungsi otak yang kompleks dan bervariasi. Biasanya gangguan ini meliputi cara berkomunikasi, berinteraksi sosial dan kemampuan berimajinasi (Pangestu & Fibriana, 2017).

2.2 *Principal Component Analysis* (PCA)

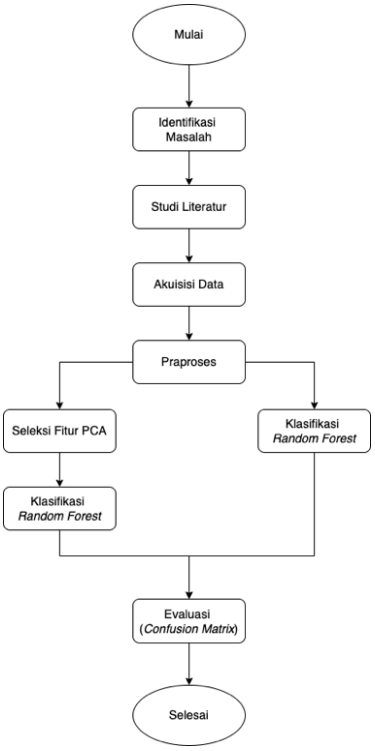
Principal Component Analysis (PCA) bertujuan untuk menyederhanakan dan menghilangkan fitur yang kurang relevan namun tanpa menghilangkan informasi penting dari data set aslinya (Adinugroho, 2018).

2.3 Klasifikasi *Random Forest*

Random Forest adalah salah satu teknik atau metode dalam kategori ensemble learning pada machine learning yang dapat bekerja dengan baik untuk mengatasi permasalahan regresi dan klasifikasi. Metode ini dapat mengurangi dimensi, mengatasi missing value, nilai outlier dan mampu mengeksplorasi tahap penting lainnya (Tahyudin dkk, 2021).

3 Metode Penelitian

Gambar 1 menunjukkan tahapan alur dari awal sampai akhir penelitian penelitian:



Gambar 1 Tahapan Penelitian

3.1 Akuisisi Data

Data yang digunakan pada penelitian ini yaitu sebanyak 292 *record data* dengan 20 fitur dan 1 atribut target. Dalam dataset terdapat sepuluh fitur perilaku (*A1_Score – A2_Score* Anak) ditambah dengan sepuluh fitur karakteristik individu yang terbukti efektif dalam mendeteksi kasus ASD dari kontrol dalam ilmu perilaku, dengan satu atribut target (terdapat dua label kelas yaitu 'YES' dan 'NO' di dalamnya) [7].

3.2 Pra-proses Data

Tahap pertama dalam pra-proses data ialah melakukan pengecekan *missing value* untuk memastikan bahwa benar adanya terdapat *missing value* pada dataset, fitur mana saja yang mengandung *missing value* serta berapa jumlah data *missing value* dari masing – masing fitur tersebut. Kemudian terhadap beberapa fitur yang mengandung *missing value* dapat ditangani dengan imputasi modus. Berikut tahapan proses dalam menangani data *missing value* menggunakan teknik imputasi dengan modus.

a) Mencari nilai atau data yang paling banyak muncul (modus) dari fitur *age* dan *ethnicity*.
b) Mengisikan *missing value* ('NaN') pada fitur *age* dan *ethnicity* dengan nilai modus yang didapat.
c) Mengecek kembali apakah masih terdapat *missing value* pada fitur *age* dan *ethnicity* atau tidak.

Selain menggunakan imputasi modus, penanganan *missing value* juga dilakukan dengan cara menghapus fitur yang tidak bekontribusi atau berkorelasi banyak diantara satu sama lain.

3.3 Seleksi Fitur

Pada dataset yang digunakan dalam penelitian tidak seluruh fitur atau kolomnya sudah bertipe data numerik. Terdapat beberapa fitur lain yang masih dalam berbentuk kata – kata (*categorical*). Sebelum masuk pada tahap seleksi fitur PCA, harus dilakukan *label encoding* untuk mengonversi label kata menjadi angka dikarenakan pada proses PCA hanya memperhitungkan atau memproses data yang bertipe data *int.* Setelah *label encoding* dan seluruh fitur atau kolom sudah bertipe data *int*, tahap selanjutnya dapat diproses pada seleksi fitur PCA, berikut merupakan tahapan seleksi fitur *Principal Component Analysis* (PCA).

a. Standard Scaling / Z-score Normalization. Standarisasi data dilakukan untuk menyeragamkan nilai – nilai data agar tidak ada lagi data yang memiliki rentang atau skala yang berbeda-beda. Hal ini dilakukan untuk mencegah dari terciptanya bias. Dimana agar fitur – fitur pada data dapat terkontribusi dengan merata pada proses perhitungan matriks kovarian selanjutnya, dan untuk melakukan perhitungan *standard scaler* ini bisa dengan menggunakan rumus pada persamaan berikut.

$$\boldsymbol{X(new)} = \frac{\boldsymbol{Xi - \mu}}{\boldsymbol{\sigma}} \quad (1)$$

Keterangan:
X(new) = data atau hasil nilai yang diperoleh dari perhitungan
Xi = suatu data atau nilai observasi ke-i pada dataset
μ = rata – rata (mean) dari masing – masing fitur
σ = standar deviasi

Untuk mencari nilai standar deviasinya, dapat dengan menggunakan persamaan berikut.

$$\boldsymbol{\sigma^2} = \frac{\sum(\boldsymbol{X - \mu})^{\boldsymbol{2}}}{\boldsymbol{N}} \quad (2)$$

Keterangan:
σ = hasil nilai standar deviasi didapat dari akar dari total perhitungan
Σ = jumlah atau total nilai dari perhitungan
X = suatu data atau nilai observasi ke-i pada dataset
μ = rata – rata (mean) dari masing – masing fitur
N = total data pada dataset

b. Menghitung matriks varians kovarian dari data penelitian. Varians dihitung untuk menemukan penyebaran data dalam dataset untuk menentukan penyimpangan data dalam sample dataset. Kovarians adalah ukuran bagaimana perubahan dalam satu variabel dikaitkan dengan perubahan variabel kedua atau dengan kata lain biasa digunakan untuk mengukur besarnya hubungan antara dua buah variabel atau fitur. Matriks kovarians adalah matriks bujur sangkar yang menunjukkan kovarians antara banyak variabel yang berbeda dan nantinya akan digunakan sebagai nilai masukan untuk mendapatkan *eigenvalues* dan *eigenvector*. Untuk menghitung nilai varians dapat menggunakan persamaan berikut.

$$\boldsymbol{Var(x)} = \boldsymbol{\sigma^2} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{Z_{ij} - \mu_j})^2 \quad (3)$$

Dan untuk menghitung nilai kovarians dapat dengan menggunakan persamaan sebagai berikut.

$$Cov(x, y) = \frac{z\,1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \quad (4)$$

Keterangan:
μx dan μy = rata – rata (*mean*) sampel dari variabel x dan y
xi dan yi = nilai observasi ke-i dari variabel x dan y.

c. Hitung *eigenvalues* dan *eigenvector* menggunakan matriks kovarian yang telah didapat sebelumnya. Rumus perhitungan untuk mencari nilai eigen dapat menggunakan persamaan berikut.

$$Det\ (A - \lambda I) = 0 \quad (5)$$

Keterangan:
Det = determinan
A = matriks n x n
λ = nilai eigen
I = matriks identitas

d. Kemudian nilai eigen yang telah didapat pada proses sebelumnya akan diurutkan dalam urutan menurun (yaitu urut dari yang memiliki nilai paling besar ke yang paling kecil).

e. Menentukan nilai "K" atau jumlah *optimal Principal Component* (PC) dengan menggunakan metode *Kaiser's Stopping Rule* yaitu dengan cara cukup memilih semua PC dengan nilai eigennya yang lebih besar dari 1 (satu).

f. Berdasarkan nilai "K" atau jumlah PC yang telah ditentukan, maka jumlah "K" vektor eigen yang ditempatkan bersama akan menghasilkan matriks proyeksi (*Projection Matrix* / PM).

g. Langkah terakhir yaitu mentransformasi dataset asli atau dataset yang digunakan pada awal masuk pemrosesan PCA, dimana dataset itu akan ditransformasi ke ruang fitur yang baru / matriks proyeksi. Ini akan dilakukan dengan menggunakan matriks proyeksi (PM) yang diperoleh pada proses sebelumnya, transformasi data dapat menggunakan persamaan berikut.
Rumus = X*(PM) (6)
Transformasi data = matriks fitur (atau dataset asli) * K vektor eigen teratas

Setelah seluruh tahapan selesai dilakukan, interpretasi hasil seleksi fitur menggunakan *Principal Component Analysis* (PCA) yaitu dengan diperolehnya dataset baru yang telah direduksi dimensinya dan ditransformasi data, serta fitur yang diperoleh dengan memilih fitur (nilai K atau jumlah PC) mana saja yang memiliki bobot nilai *eigenvector* paling tinggi dari sejumlah *principal component* yang memiliki korelasi cukup besar.

3.4 Klasifikasi *Random Forest*

Selanjutnya data dari hasil seleksi fitur yang telah dilakukan sebelumnya akan digunakan untuk pemodelan klasifikasi *Random Forest*. Dengan melakukan pembagian data terlebih dahulu menjadi 2 bagian yaitu training data (data latih) dan testing data (data uji). Sebanyak 80% akan digunakan sebagai data latih pada model klasifikasi *random forest* sebanyak 80% sedangkan sebanyak 20% data uji akan digunakan sebagai evaluasi model klasifikasi *Random Forest*. Pemilihan persentase dalam pembagian data pada penelitian ini didasari oleh Prinsip Pareto, dimana prinsip ini juga dikenal sebagai aturan 80/20. Ini pada dasarnya adalah teori yang menetapkan bahwa 80% dari output atau hasil berasal dari 20% efek atau aliansi yang tidak proporsional antara input dengan output (The Pareto Principal, 2019) [6]. Setelah pembagian data, masuk pada proses pemodelan menggunakan klasifikasi *random forest*. Pada metode klasifikasi ini menerapkan *decision tree* dimana *random forest* terbentuk dari sekumpulan *decision tree* (pohon keputusan). Tahapan algoritma dalam membangun tree menggunakan *Random Forest* yaitu sebagai berikut.

a. Buat subset data dari dataset menggunakan *bagging + features randomness*
b. Gunakan subset data untuk membangun pohon ke i (i= 1,2,3 ….. *k*)
c. Ulangi langkah kesatu dan kedua sebanyak *k*

3.5 Evaluasi

Pada evaluasi akan dilakukan pengumpulan informasi yang berkaitan dengan kinerja atau performa yang dihasilkan yang dapat digunakan untuk menentukan alternatif terbaik dalam membuat keputusan. Membangun model *machine learning* saja tidaklah cukup, dan perlu mengetahui seberapa baik model tersebut bekerja. Evaluasi pada penelitian ini menggunakan *Confusion Matrix* yang ditunjukkan pada **Tabel 1** yang berfungsi untuk mengukur kinerja algoritma klasifikasi atau suatu model klasifikasi.

Tabel 1 *Confusion Matrix*

Two-Class Prediction		***Actual Values***	
		Positif (1)	**Negatif (0)**
Predicted Values	**Positif (1)**	*True Positive (TP)*	*False Positif (FP)*
	Negatif (0)	*False Negatif (FN)*	*True Negatif (TN)*

Keterangan tabel:
True Positive (TP) : Data dengan nilai aktual positif dan diprediksi positif
False Positive (FP) : Data dengan nilai aktual negatif dan diprediksi positif
False Negative (FN) : Data dengan nilai aktual positif dan diprediksi negatif
True Negative (TN) : Data dengan nilai aktual negatif dan diprediksi negatif

Confusion Matrix menampilkan beberapa informasi seperti *True Negative, True Positive, False Negative* dan *False Positive*, dimana dari informasi tersebut dapat diketahui nilai *accuracy* (akurasi), *precision, recall* dan *specificity*. Untuk menghitung masing – masing nilai tersebut sebagai bahan evaluasi dapat menggunakan persamaan berikut.

Nilai akurasi dapat dihitung melalui **persamaan** sebagai berikut.

$$\boldsymbol{Akurasi} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+FN+FP+TN}} \quad (6)$$

Nilai *precision* dapat dihitung melalui **persamaan** sebagai berikut.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TN}}{\boldsymbol{TP+FP}} \quad (7)$$

Nilai *recall* dapat dihitung melalui **persamaan** sebagai berikut.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (8)$$

Nilai *specificity* dapat dihitung melalui **persamaan** sebagai berikut.

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \quad (9)$$

4 Hasil dan Pembahasan

4.1 Akuisisi Data

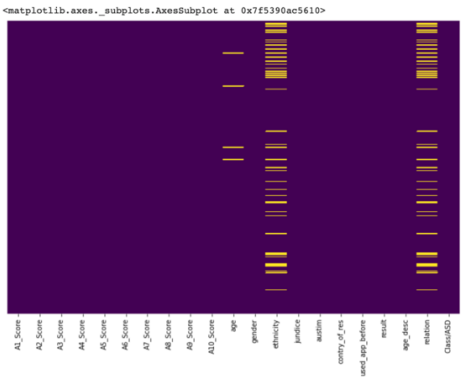
Data yang digunakan dalam penelitian ini diperoleh dan diakuisisi dari *University of California Irvine* (UCI) *Machine Learning Data Repository* yaitu *Autistic Spectrum Disorder Screening Data for Children Data Set*. **Tabel 2** di bawah menampilkan informasi mengenai *dataframe* dari *dataset* yang digunakan, seperti jumlah baris data, jumlah kolom (fitur) data beserta masing-masing namanya, jumlah data dan tipe datanya.

Tabel 2 Informasi *dataframe*

Range Index: 292 entri, 0 sampai 291			
Data Columns (total ada 22 kolom)			
#	***Column***	***Non-Null Count***	***Dtype***
0	id	292 non-null	int64
1	A1_Score	292 non-null	int64
2	A2_Score	292 non-null	int64
3	A3_Score	292 non-null	int64
4	A4_Score	292 non-null	int64
5	A5_Score	292 non-null	int64
6	A6_Score	292 non-null	int64
7	A7_Score	292 non-null	int64
8	A8_Score	292 non-null	int64
9	A9_Score	292 non-null	int64
10	A10_Score	292 non-null	int64
11	age	288 non-null	float64
12	gender	292 non-null	object
13	ethnicity	249 non-null	object
14	jundice	292 non-null	object
15	austim	292 non-null	object
16	contry_of_res	292 non-null	object
17	used_app_before	292 non-null	object
18	result	292 non-null	int64
19	age_desc	292 non-null	object
20	relation	249 non-null	object
21	Class/ASD	292 non-null	object

4.2 Hasil Pra-proses Data

Hasil pengecekan *missing value* terhadap *dataset* menunjukkan bahwa dari seluruh fitur yang ada, terdapat 3 buah fitur yang mengandung *missing value* dengan jumlah (*total*) seluruhnya sebanyak 90 data *missing value*, yaitu *age* sebanyak 4 buah data, *ethnicity* dan *relation* masing-masing sebanyak 43 buah data. Hasil pengecekan *missing value* dalam bentuk subplot ditunjukkan dengan **Gambar 2** berikut.



Gambar 2 Hasil pengecekan *missing value* dalam bentuk subplot

Penanganan *missing value* dengan imputasi modus diterapkan pada fitur *age* dan *ethnicity*. Sedangkan pada fitur *relation*, penanganan *missing value* dilakukan dengan menghapus fitur tersebut, tidak hanya itu, fitur *used_app_before*, *age_desc*, dan *result* juga akan dihapus dari *dataframe*. Kemudian *dataframe* akan diduplikat terlebih dahulu dan kemudian target atau kelas *Class/ASD* baru dihapus pada *dataframe* yang baru. Menduplikasi *dataframe* sengaja dilakukan sebagai persiapan untuk proses *Label Encoding* dan PCA. *Dataframe* yang baru akan digunakan sebagai wadah untuk mengotak-atik jalannya seluruh proses *Label Encoding* hingga PCA, sehingga *dataframe* asli tidak terikut campur oleh proses apapun yang dilakukan setelah penduplikasian *dataframe* sebelumnya. *Dataframe* asli sengaja dibiarkan untuk menyimpan dan menjaga keutuhan data dari kelas atau target Class/ASD sebagai *y* yang akan digunakan pada proses pembagian data (*splitting data*). Kemudian *label encoding* dilakukan terhadap fitur – fitur yang masih bukan bertipe data *int*, karena pada proses PCA hanya memperhitungkan dan memproses data yang bertipe *int*. **Tabel 3** di bawah menunjukkan informasi dari *dataframe* yang siap digunakan pada proses selanjutnya, yaitu *Principal Component Analysis* (PCA).

Tabel 3 Informasi *dataframe* siap digunakan pada proses PCA

#	*Column*	*Non-Null Count*	*Dtype*
1	A1_Score	292 non-null	int64
2	A2_Score	292 non-null	int64
3	A3_Score	292 non-null	int64
4	A4_Score	292 non-null	int64
5	A5_Score	292 non-null	int64
6	A6_Score	292 non-null	int64
7	A7_Score	292 non-null	int64
8	A8_Score	292 non-null	int64
9	A9_Score	292 non-null	int64
10	A10_Score	292 non-null	int64
11	age	292 non-null	int64
12	gender_label	292 non-null	int64
13	ethnicity_label	292 non-null	int64
14	jundice_label	292 non-null	int64
15	austim_label	292 non-null	int64
16	contry_of_res_label	292 non-null	int64

4.3 Hasil Seleksi Fitur *Principal Component Analysis* (PCA)

Berikut merupakan 10 indeks baris pertama dari hasil proses seleksi fitur *Principal Component Analysis* (PCA) yang ditunjukkan pada **Tabel 4** di bawah.

Tabel 4 Hasil akhir proses seleksi fitur PCA

index	PC1	PC2	PC3	PC4	PC5	PC6
0	-0.956404	1.029613	0.083464	-0.475670	0.117402	-0.078828
1	-0.962773	1.175267	0.098017	-0.512876	0.129099	-0.011156
2	-1.215777	-0.082864	-0.250572	-1.331137	1.168771	0.059799
3	-1.594654	-0.946446	1.810095	0.702401	-1.356767	1.890485
4	2.403483	-1.071391	1.365161	0.063402	0.728800	-0.177096

5	-0.572618	1.504444	0.368479	2.196544	-0.620507	-1.512352
6	1.040040	0.545760	0.052723	0.295978	-0.143660	-1.724172
7	0.748498	0.982862	-0.365109	-0.238631	-0.923630	0.762955
8	0.262691	0.155094	-1.485136	-0.992185	-1.218355	1.698403
9	-0.875972	-0.233985	-2.784859	0.853375	-0.222739	-0.138975

4.4 Klasifikasi *Random Forest*

Sebelum masuk pada tahap pembuatan model klasifikasi *random forest*, data akan dibagi terlebih dahulu. Pembagian data mengacu berdasarkan prinsip pareto dengan menerapkan 80/20, yang berarti sebesar 80% dari data untuk data training (data latih) dan 20% dari data untuk data testing (data uji). Pada penelitian ini maka diperoleh sebanyak 233 data sebagai data latih dan 59 data sebagai data uji. Tahap selanjutnya yaitu pembuatan model. Pada penelitian ini, jumlah pohon (*trees*) yang dibangun ialah sebanyak 100 buah. Dalam proses pengklasifikasiannya, menerapkan *bagging* dan *features randomness*, dimana untuk setiap bag yang dihasilkan itu akan mengadopsi sejumlah fitur yang diambil secara acak dari training set. Hal inilah *yang* menyebabkan bahwa setiap model yang dihasilkan akan ditraining dengan masing – masing *subset bag* yang beragam, tidak hanya beragam pada baris data saja, melainkan juga dalam hal fitur yang diikutsertakan. Oleh sebab itu, setiap bag yang dihasilkan (*bag 1, bag 2*, dst.) mengusung baris data serta fitur yang berbeda – beda (beragam) yang masing – masing dipilih secara acak dari training set sumbernya. Oleh karena setiap model *decision tree* ditraining dengan menggunakan dataset yang berbeda – beda, maka akan menghasilkan trained model yang berbeda – beda (beragam) juga. Setiap *trained model* yang dihasilkan itu nantinya akan digunakan untuk melakukan prediksi, dan prediksi yang dihasilkan nantinya dapat disatukan dengan melalui proses *majority voting* (pengambilan suara terbanyak) untuk menghasilkan prediksi final.

4.5 Evaluasi

4.5.1 Evaluasi dengan menggunakan PCA

Evaluasi dilakukan pada data uji menggunakan model klasifikasi yang telah dibuat dan direpresentasikan dengan *confusion matrix*. Berikut merupakan hasil *confusion matrix* terhadap data yang dengan melalui proses seleksi fitur *Principal Componen Analysis* (PCA) yaitu ditunjukkan pada **Tabel 5** berikut.

Tabel 5. *Confusion Matrix* dengan menggunakan PCA

Two-Class Prediction		***Actual Values***	
		Negatif (0)	**Positif (1)**
Predicted Values	**Negatif (0)**	31	0
	Positif (1)	1	27

Keterangan:
True Positive (TP) = 27
False Positive (FP) = 1
False Negative (FN) = 0
True Negative (TN) = 31

Berdasarkan hasil dari *confusion matrix*, dapat dihitung nilai akurasi, *precision, recall* dan *specificity* secara manual yaitu sebagai berikut.

1) Nilai akurasi :
$$\boldsymbol{Akurasi = \frac{27+31}{27+0+1+31} = \frac{58}{59} = 0.9830508475}$$

2) Nilai *precision* :
$$\boldsymbol{Precision = \frac{27}{27+1} = 0.9642857143}$$

3) Nilai *recall* :
$$\boldsymbol{Recall = \frac{27}{27+0} = 1.00}$$

4) Nilai *specificity* :

$$\boldsymbol{Specificity} = \frac{31}{31+1} = \mathbf{0.96875}$$

4.5.1 Evaluasi yang tidak menggunakan PCA

Berikut merupakan hasil *confusion matrix* terhadap data yang tanpa melalui proses seleksi fitur *Principal Componen Analysis* (PCA) yaitu ditunjukkan pada **Tabel 6** berikut.

Tabel 6 *Confusion Matrix* tanpa menggunakan PCA

Two-Class Prediction		***Actual Values***	
		Negatif (0)	**Positif (1)**
Predicted Values	**Negatif (0)**	26	5
	Positif (1)	0	28

Keterangan:
True Positive (TP) = 26
False Positive (FP) = 5
False Negative (FN) = 0
True Negative (TN) = 28

Berdasarkan hasil dari *confusion matrix*, dapat dihitung nilai akurasi, presisi, *recall* dan *specificity* secara manual yaitu sebagai berikut.

1) Nilai akurasi :

$$\boldsymbol{Akurasi} = \frac{28+26}{28+0+5+26} = \frac{54}{59} = \mathbf{0.9152542373}$$

2) Nilai *precision* :

$$\boldsymbol{Precision} = \frac{26}{28+0} = \mathbf{0.9285714286}$$

3) Nilai *recall* :

$$\boldsymbol{Recall} = \frac{28}{28+5} = \mathbf{0.848484845}$$

4) Nilai *specificity* :

$$\boldsymbol{Specificity} = \frac{26}{26+0} = \mathbf{1.00}$$

Tabel 7 di bawah ini menunjukkan perbandingan kedua hasil evaluasi antara yang menggunakan PCA dan yang tidak menggunakan metode tersebut.

Tabel 7 Perbandingan hasil evaluasi

	Evaluasi dengan menggunakan PCA	**Evaluasi tanpa menggunakan PCA**
Nilai Akurasi	98%	91%
Nilai *Precision*	96%	92%
Nilai *Recall*	100%	84%
Nilai *Specificity*	96%	100%

5 Penutup

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, dapat ditarik kesimpulan sebagai berikut:

1. Penelitian ini telah berhasil menerapkan metode seleksi fitur *Principal Component Analysis* (PCA) untuk mereduksi dimensi data dan transformasi data yaitu melalui beberapa tahapan seperti standarisasi data, menghitung matriks kovarian, mencari nilai *eigenvectors* dan *eigenvalues*, menentukan nilai K atau jumlah PC dengan metode *Kaiser's Stopping Rule* kemudian akan menghasilkan matriks proyeksi dan terakhir mentransformasi data, kemudian hasil

data dari proses PCA tersebut akan dilakukan klasifikasi *random forest* dengan membagi data menjadi data latih dan data uji, data latih akan digunakan untuk membuat model, kemudian data uji akan dilakukan uji coba pada pemodelan yang telah dibuat.

2. Pada penelitian ini, hasil evaluasi terhadap data yang melalui proses PCA menghasilkan nilai akurasi sebesar 98%, dengan presisi sebesar 96%, *recall* sebesar 100% dan s*pecifi*city sebesar 96% dari data penelitian Gangguan Spektrum Autisme (ASD) pada Anak-anak.

3. Terdapat perbedaan terhadap data yang tidak menggunakan atau tanpa melalui proses *Principal Component Analysis* (PCA) dengan yang menggunakan metode tersebut. Setelah diuji coba pada model, data yang tanpa melalui proses PCA didapat hasil evaluasi yaitu nilai akurasi sebesar 91%, *precision* sebesar 92%, *recall* sebesar 84% dan *specificity* sebesar 100%. Berdasarkan hasil tersebut, diketahui bahwa penerapan metode *Principal Component Analysis* (PCA) terhadap data yang digunakan pada penelitian ini dapat meningkatkan performa klasifikasi *random forest*.

5.2 Saran

Penelitian ini memiliki banyak keterbatasan serta masih banyak yang belum dieksplorasi dan dikembangkan lagi. Pada penelitian selanjutnya dengan topik serupa dapat mempertimbangkan beberapa saran berikut:

1. Pada praproses data, untuk menangani missing value pada data dapat diterapkan metode imputasi lainnya seperti mean atau median.
2. Pada penelitian selanjutnya mungkin dapat menerapkan jumlah pohon yang berbeda dan bervariasi untuk dapat melihat perbedaan dan perbandingan diantara masing – masing.
3. Pembagian data sebelum masuk pada pemodelan dilakukan secara manual dengan membaginya menjadi 80% data latih dan 20% data uji, dan dapat dipertimbangkan untuk menggunakan angka persentase pembagian data lainnya dan atau menggunakan metode pembagian data yang berbeda.
4. Selain metode *Principal Component Analysis* (PCA) dan klasifikasi *random forest*, masih banyak metode seleksi fitur untuk mereduksi dimensi data dan metode klasifikasi lainnya yang dapat dipertimbangkan dan dieksplorasi pada penelitian selanjutnya.

Referensi

[1] Adinugroho. 2018. *Implementasi Data Mining Menggunakan Weka*. (n.p.): Universitas Brawijaya Press.

[2] Nasution dkk. 2019. *Penerapan Principal Component Analysis (PCA) Dalam Penentuan Faktor Dominan Yang Mempengaruhi Pengidap Kanker Serviks (Studi Kasus : Cervical Cancer Dataset).* Jurnal Mantik Penusa Vol. 3, No. 1. Juni. Hal. 204 – 210.

[3] *National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention*. URL: (https://www.cdc.gov/ncbddd/autism/facts.html) Diakses pada 20 November 2021.

[4] Pangestu dan Fibriana. 2017. *Faktor Risiko Kejadian Autisme*. Higeia Journal Of Public Health Research And Development. P Issn 1475-362846, E Issn 1475-222656. Hal. 141 – 150.

[5] Tahyudin dkk. 2021. *Data Mining Dan Data Warehouse Menggunakan Aplikasi KNIME*. (n.p.): Zahira Media Publisher. Hal. 39.

[6] *The Pareto Principal*. 2019. (n.p.): Can Akdeniz.

[7] University of California Irvine (UCI) Machine Learning Data Repository. *Autistic Spectrum Disorder Screening Data for Children Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++ Diakses pada 9 Oktober 2021.