PROJECT UJIAN AKHIR SEMESTER

PENGANTAR DATA SCIENCE

PERMODELAN KLASIFIKASI PASIEN AUTISME BERBASIS *ARTIFICIAL NEURAL NETWORK*

Anggota Kelompok : 1. Amalia Nur Zahro (22/492729/PA/21138)

2. Gabriella Yoanda Pelawi (22/493047/PA/21159)

3. Muhammad Husni Zahran (22/498283/PA/21501)

4. Mahardi Nalendra Syafa (22/502515/PA/21558)

Dosen Pengampu : Mohamad Fahruli Wahyujati, S.Si., M.Si.

PROGRAM STUDI SARJANA STATISTIKA

DEPARTEMEN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS GADJAH MADA

2024

Abstrak

Autism Spectrum Disorder (ASD) merupakan kondisi *neurodevelopmental* yang ditandai dengan kesulitan dalam interaksi sosial, komunikasi, dan perilaku repetitif. Penelitian ini bertujuan untuk mengembangkan model klasifikasi ASD menggunakan *Artificial Neural Network* (ANN) dan membandingkannya dengan model *K-Nearest Neighbors* (K-NN) dan *Support Vector Classification* (SVC). Dataset yang digunakan diambil dari situs Kaggle dengan judul "Autism Spectrum Disorder (ASD) Data Set", yang merupakan data dari proses screening autisme menggunakan AQ-10. *Hyperparameter tuning* pada model ANN dengan pendekatan bayesian atau *Bayesian Optimization* menaikkan nilai akurasi model ANN dimana menunjukkan performa yang paling baik dalam klasifikasi ASD dengan akurasi sebesar 92%. Model K-NN mencapai akurasi sebesar 90% dalam klasifikasi ASD. Model SVC mencapai akurasi sebesar 79% dalam klasifikasi ASD. Model ANN berpotensi untuk menjadi alternatif metode skrining autisme yang lebih akurat dan efisien.

Kata kunci: *Autism Spectrum Disorder* (ASD), Klasifikasi, *Artificial Neural Network* (ANN), *K-Nearest Neighbors* (K-NN), *Support Vector Classification* (SVC), *Bayesian Optimization, Hyperparameter tuning, Screening*

LATAR BELAKANG

Menurut *World Health Organization* (WHO, 2023), sekitar 1 dari 100 orang di seluruh dunia diperkirakan memiliki gangguan spektrum autisme. *Autism Spectrum Disorder (ASD)* merupakan kondisi *neudevelopmental* dimana seseorang mengalami kesulitan dalam keterampilan sosial, perilaku *repetitive*, komunikasi verbal dan non-verbal (Erkan & Thanh, 2019). Proses screening ASD adalah langkah awal penting dalam mengidentifikasi individu yang mungkin memiliki autisme. Screening melibatkan serangkaian tes atau observasi untuk mendeteksi tanda-tanda awal autisme pada anak-anak maupun orang dewasa, dengan tujuan memastikan penanganan lebih lanjut.

Proses *screening* ASD bisa berbeda-beda tergantung usia dan alat yang digunakan. Alat yang sering digunakan untuk sistem ini adalah ***Autism Spectrum Quotient (AQ)***. AQ dirancang untuk membantu orang dewasa mendeteksi ciri-ciri autistik melalui kuesioner yang diisi sendiri. Versi singkat dari tes ini, disebut **AQ-10-Adult**, terdiri dari 10 pertanyaan. Skor lebih dari 6 pada AQ-10-Adult menunjukkan kemungkinan autisme. Hasil dari AQ dapat memberikan indikasi awal yang kemudian dapat ditindaklanjuti dengan evaluasi lebih lanjut.

Berdasarkan masalah ini, akan dilakukan permodelan klasifikasi pasien autisme menggunakan ANN dan membandingkannya dengan model *machine learning* lainnya. Analisis ini bertujuan untuk memodelkan klasifikasi pasien autisme menggunakan ANN dan membandingkannya dengan model *machine learning* lainnya, dengan tujuan mengidentifikasi fitur penting dari dataset dan meningkatkan akurasi serta efisiensi diagnosis autisme.

METODE PENELITIAN

Dataset

Dataset yang digunakan diambil dari situs Kaggle dengan judul Autism Screening on Adults, yang merupakan data dari proses screening autisme menggunakan **Autism Spectrum Quotient (AQ)**, sebuah alat yang dirancang untuk mendeteksi ciri-ciri autistik pada orang dewasa melalui kuesioner yang diisi

sendiri. Dataset ini mencakup berbagai fitur yang relevan untuk analisis dan klasifikasi autisme, yaitu :

Fitur	Deskripsi
index	ID pasien
AX_Score	Skor ASD berdasarkan alat skrining AQ-10
age	Usia pasien ASD
gender	Jenis kelamin pasien (‘m’ untuk laki-laki dan ’f’ untuk perempuan)
ethnicity	Kategori etnis pasien ('White-European', 'Latino', 'Black', 'Asian', dll)
jaundice	Riwayat penyakit kuning pasien (‘no’ dan ’yes’)
autism	Apakah keluarga pasien pernah didiagnosis autisme (’no’ dan ’yes’)
country_of_res	Kategori negara asal pasien
used_app_before	Apakah pasien pernah menggunakan aplikasi skrining sebelumnya (’no’ dan ’yes’’)
result	Skor dari alat skrining AQ-10
age_desc	Usia pasien dalam bentuk kategori ’18 and more’
relation	Kategori hubungan pengisi tes dengan pasien (’self’, ’parent’, dll)
Class/ASD	Klasifikasi hasil diagnosis autisme pasien (’no’ dan ’yes’)

Tabel 1. Deskripsi Fitur

Artificial Neural Network (ANN)

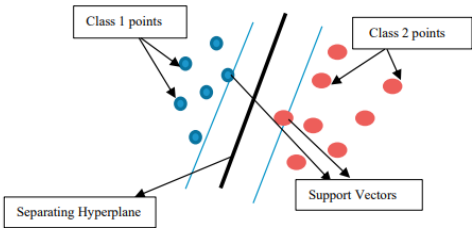
Artificial Neural Network (ANN) adalah model matematis yang meniru cara kerja otak manusia, terdiri dari neuron buatan yang terhubung dan mampu memproses serta belajar dari data. Setiap neuron menerima input, memprosesnya dengan fungsi aktivasi, dan menghasilkan output yang dikirim ke neuron lain.

Analisis ANN untuk klasifikasi ASD melibatkan *pre-processing* data (imputasi nilai hilang, encoding data kategorik, dan normalisasi fitur), eksplorasi data, dan pembentukan model menggunakan Sequential dari Keras. Model ini memiliki lapisan input, beberapa lapisan tersembunyi dengan aktivasi ReLU, dan lapisan output dengan aktivasi sigmoid. Model dikompilasi dengan *optimizer* Adam dan *loss function binary crossentropy*, lalu dilatih dengan data pelatihan dan pengujian (80:20). Setiap iterasi menyesuaikan bobot koneksi neuron untuk meminimalkan kesalahan prediksi. Evaluasi dilakukan dengan prediksi terhadap data pengujian dan menghitung metrik seperti akurasi, precision, recall, dan *confusion matrix*. Validasi menggunakan *K-Fold Cross-Validation* memastikan kemampuan generalisasi model dengan menghitung akurasi rata-rata dari setiap fold.

Machine Learning

Penelitian ini menggunakan beberapa metode *machine learning* untuk mengklasifikasikan autisme dan membandingkannya dengan model ANN. Setelah *preprocessing* dan normalisasi data, data dibagi menjadi set pelatihan dan pengujian (80:20).

Metode *machine learning* yang pertama digunakan adalah *Support Vector Machine* (SVM). SVM adalah algoritma *machine learning* yang mencari hyperplane optimal untuk memisahkan kelas-kelas dalam n-dimensional data (Sujatha R, Chatterjee, Alaboudi, & Jhanjhi, 2021). Algoritma ini efektif untuk klasifikasi linier dan dapat diperluas untuk data non-linier menggunakan kernel trick.



Gambar 1. Visualisasi SVM

Metode kedua adalah *K-Nearest Neighbors* (KNN). KNN adalah algoritma berbasis instance yang mengklasifikasikan data baru berdasarkan mayoritas dari k tetangga terdekat dalam data pelatihan.

Hyperprameter Optimization : Bayesian Approcah

Penelitian ini bertujuan untuk mengembangkan model klasifikasi optimal guna mengidentifikasi apakah pasien dewasa mengidap autisme. Pendekatan utamanya adalah mengoptimalkan *hyperparameter* dari model dasar. Dalam *machine learning*, terdapat dua jenis parameter: parameter yang dipelajari dari data, seperti bobot jaringan saraf, dan hiperparameter yang diatur sebelum pelatihan, seperti jumlah *node* atau nilai *dropout* dalam model.

Optimasi *hyperparameter* bertujuan untuk memaksimalkan kinerja algoritma *machine learning* yang diberikan dengan memilih hyperparameter yang paling sesuai. Berdasarkan Persamaan (1) di mana f menyatakan performa, x

dikatakan sebagai beberapa pengaturan hyperparameter, dan pilihan optimal adalah x_{opt}

$$x_{opt} = arg\max_{x \in X} f(x) \quad (1)$$

Dalam penelitian ini, teknik untuk mendapatkan hyperparameter optimal digunakan yaitu *Bayesian Optimization*. Bayesian Optimization mampu secara efisien menemukan hyperparameter yang optimal dengan iterasi yang lebih sedikit (Jia Wu a, 2019). Semua Optimasi Bayesian adalah langkah untuk membangun model probabilitas menggunakan fungsi objektif. Probabilitas Bayes ditulis dalam Persamaan

$$p(m|n) = \frac{p(m|n) * p(n)}{p(m)} \quad (2)$$

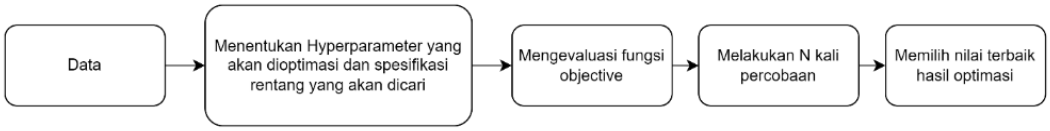
Dengan $p(m|n)$ = probabilitas hyperparameter yangmenilai fungsi objektif. Nilai keluaran dapat dapat dituliskan dalam Persamaan (3)

$$p(a|b) = \begin{cases} l(a), & jika\ b < b^* \\ g(a), & jika\ b \geq b^* \end{cases} \quad (3)$$

Nilai $b < b^*$ menunjukkan biaya yang lebih rendah dari fungsi objektif dari ambang batas, diberi label $l(a)$, dan jika lebih besar, maka diberi label $g(a)$

Bayesian Optimization menggunakan model pengganti yang sesuai dengan pengamatan model nyata. Dalam kasus ini, sebuah observasi adalah proses training dari model dasar ANN dengan *hyperparameter* yang dipilih secara khusus untuk observasi tersebut. Satu set *hyperparameter* dipilih untuk setiap iterasi, dan sebuah observasi kemudian dibuat. Akurasi validasi digunakan untuk evaluasi pengamatan. *Hyperparameter* dipilih menggunakan fungsi akuisisi yang menyeimbangkan pilihan antara mengeksplorasi seluruh ruang pencarian dan menentukan area yang berkinerja baik dari ruang pencarian.

Eksperimen optimasi Bayesian. Gambar (2) menunjukkan alur dari Bayesian Optimization.



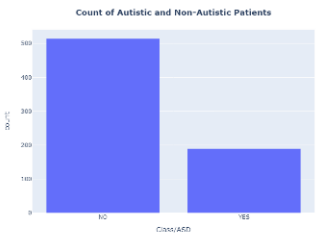
Gambar 2. Alur Optimasi Bayesian

HASIL DAN ANALISIS

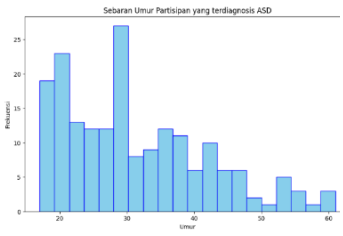
Prepocessing Data

Outliers dideteksi menggunakan *boxplot*, kemudian *interquartile range* (IQR) ditentukan menggunakan batas atas dan batas bawah dari kolom tersebut. Nilai-nilai yang berada di luar batas dihapus. Setelah dideteksi, terdapat *outliers* pada variabel 'age'. *Outliers* tersebut dihapus menggunakan teknik ini. Karena *missing value* hanya terdapat pada satu variabel yaitu variabel 'age' dengan dua nilai yang *missing*, maka data dengan *missing value* ditangani menggunakan imputasi dengan nilai rata-rata. Variabel 'age_desc' dianggap tidak relevan atau tidak berguna untuk model sehingga dihapus untuk meningkatkan kinerja model dan mengurangi kompleksitas. Fitur kategorikal perlu diubah menjadi format numerik agar dapat digunakan pada model. Dataset pasien autisme dibagi menjadi dua bagian untuk pelatihan dan pengujian. Bagian data latih berisi 80% data sedangkan 20% data sisanya digunakan untuk tujuan pengujian. Dari total 703 baris, 562 baris pelatihan digunakan untuk membangun model klasifikasi dan 141 baris sisanya digunakan untuk bagian pengujian untuk mengevaluasi model yang telah dibangun.

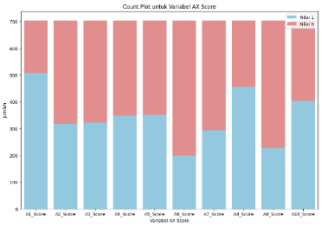
Exploratory Data Analysis (EDA)



Gambar 3. Count of Autistic and Non-Autistics Patients



Gambar 4. Sebaran Umur Partisipan yang Terdiagnosis ASD



Gambar 5. Count Plot untuk Variabel AX Score

Penelitian ini berfokus pada distribusi pasien yang didiagnosis dengan Autism Spectrum Disorder (ASD) dan distribusi umur partisipan yang terdeteksi

mengalami ASD. Distribusi jumlah pasien yang didiagnosis dengan ASD dan yang tidak, menunjukkan bahwa pasien non-ASD lebih banyak (Gambar 3). Distribusi umur partisipan yang mengalami ASD memperlihatkan mayoritas berada di kelompok umur remaja dan dewasa (Gambar 4). Selain itu, distribusi jumlah partisipan berdasarkan skor AX mereka, yang digunakan untuk mengukur berbagai aspek terkait ASD (Gambar 5)

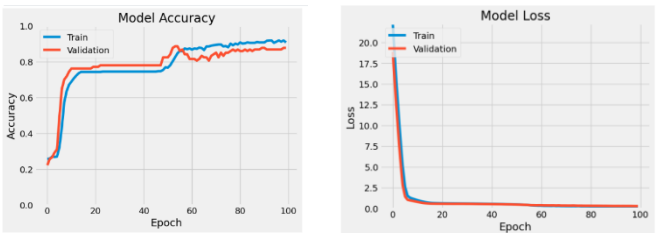
Model *Artificial Neural network*

Analisis klasifikasi untuk autisme pada orang dewasa dilakukan pemodelan dengan membuat *layer* ANN. Arsitektur yang digunakan pada pemodelan ini adalah sebagai berikut.

Nama Layer	Bentuk Input	Bentuk Output	Parameter
Input Layer	(None, 19)	(None, 19)	-
dense	(None, 19)	(None, 10)	200
dense_1	(None, 10)	(None, 8)	88
dense_2	(None, 8)	(None, 1)	9

Tabel 2. Layer awal ANN

Proses *training* dari model dilakukan dengan menggunakan *optimizer* Adam *(Adaptive Moment)* dengan *learning rate* 0,001 dan *epoch* sebesar 100. Berikut merupakan *plot error (loss)* dan akurasi model setiap *epoch* selama proses *training.*



Gambar 6. Plot loss dan accurcy Model ANN

Didapatkan nilai akurasi untuk data validasi sebesar 87,6%. Selanjutnya, pemodelan tersebut diuji pada data *test* dan memperoleh nilai metrik evaluasi akurasi sebesar 87,2%. Sehingga, diperoleh model terbukti dapat mengklasifikasikan autisme dengan cukup baik.

Bayesian Optimization

Pada pembentukan model klasifikasi dasar dengan ANN, dapat ditinjau nilai akurasi yang didapat adalah 87% di mana nilai ini dapat dikatakan model sudah cukup baik dalam mengklasifikasikan. Selanjutnya akan dilakukan upaya menaikkan nilai akurasi dengan hyperparameter tuning pada model ANN dengan pendekatan bayesian atau Bayesian Optimization.

Hyperparameter yang dipertimbangkan dalam percobaan *Optimasi Bayesian* diberikan dalam Tabel 2 bersama dengan dimensi pencarian untuk setiap hyperparameter.

Hyperparameter	**Jangkauan Pencarian**
n_layers	*(1,10)*
n_units	*(16,512)*
Learning_rate	*(1e-6, 1e-2)*

Tabel 3. Jangkauan Pencarian Hyperparameter

Hyperparameter yang dipilih mencakup jumlah lapisan (1-10), jumlah unit per lapisan (16-512), dan learning rate (1e-6 hingga 1e-2). Jumlah lapisan dan unit yang lebih tinggi memungkinkan penangkapan fitur lebih kompleks namun berisiko *overfitting*. Rentang *learning rate* dipilih untuk menghindari konvergensi lambat atau kegagalan mencapai konvergensi.

Dalam proses Bayesian Optimization, dilakukan 15 iterasi. Awalnya, *optimizer* mencoba 5 kombinasi *hyperparameter* secara acak untuk memahami ruang *hyperparameter*. Kemudian, *optimizer* melakukan 10 iterasi tambahan, memprediksi dan memilih kombinasi *hyperparameter* yang diharapkan meningkatkan performa model. Teknik *callback* ‘EarlyStopping’ digunakan untuk menghentikan pelatihan lebih awal jika performa model tidak membaik secara signifikan selama 3 *epoch* berturut-turut pada data validasi.

Summary dari model optimal tiap iterasi yang didapat adalah berikut

Iter	***Target***	***Learning_rate***	***n_layers***	***N_units***
1	0.8794	0.004171	7.483	16.06
2	0.9007	0.003024	2.321	61.8
3	0.8936	0.001863	4.11	212.8

4	0.8865	0.005389	4.773	355.9
5	0.8865	0.002045	8.903	29.58
6	0.8936	0.002928	2.371	60.78
7	0.9149	0.003375	1.624	62.39
8	0.8865	0.00382	1.099	63.12
9	0.8794	0.002276	2.4	485.9
10	0.6525	0.007263	7.558	112.6
11	0.8156	0.001907	6.944	208.7
12	0.8936	0.002746	1.654	216.4
13	0.766	0.004839	6.232	217.1
14	0.8652	0.003723	1.23	210.8
15	0.8652	0.006581	8.46	25.2

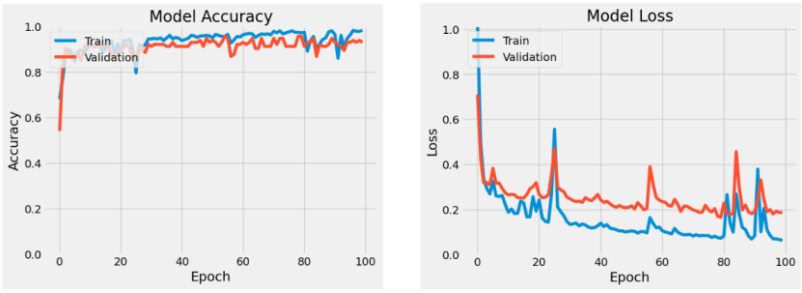
Tabel 4. Iterasi Optimisasi

Dari tabel tersebut, kombinasi *hyperparameter* terbaik yang ditemukan adalah pada iterasi ke-7 dengan akurasi sebesar 0,9149 pada data *validation* dengan kombinasi *hyperparameter learning rate* sebesar 0.0033746930961514395, Jumlah lapisan sekitar 1.62 (dibulatkan menjadi 1 lapisan, karena jumlah lapisan harus bilangan bulat). Jumlah unit per lapisan sekitar 62.39 (dibulatkan menjadi 62 unit). Gambaran model dapat dilihat pada tabel ()

Nama Layer	**Bentuk Input**	**Bentuk Output**	**Parameter**
Dense_499_input	(None, 19)	(None, 19)	-
dense_499	(None, 19)	(None, 62)	1240
dense_500	(None, 62)	(None, 1)	63

Tabel 5. Hyperparameter Optimisasi

Menggunakan model terbaik tersebut, dilakukan pemodelan ulang dan didapatkan akurasi sebesar 92% sehingga dapat dikatakan bahwa model sudah sangat baik dalam melakukan klasifikasi autisme. Berikut merupakan *plot error (loss)* dan akurasi model setiap *epoch* selama proses *training* .

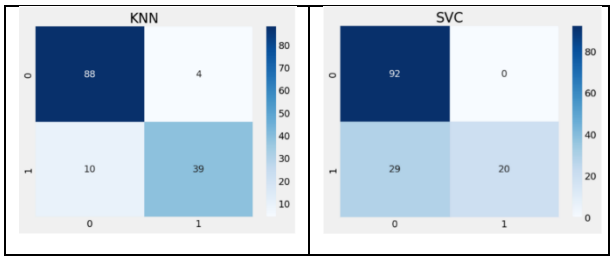


Gambar 7. Akurasi dan Loss Model Optimal

Nilai akurasi pada model hasil *hyperparameter tuning* menghasilkan nilai sebesar 92% pada data *test*, lebih baik dari model awal dengan akurasi sebesar 87%. Dapat disimpulkan bahwa model hasil *hyperparameter tuning* lebih baik daripada model sebelumnya .

Model K-Nearest Neighbors dan Support Vector Classification

Dataset dibagi secara acak menjadi dua set, satu set 562 kasus (80% dari keseluruhan dataset) untuk pelatihan dan 141 kasus untuk menguji model. Model dibangun menggunakan set pelatihan dengan K-Nearest Neightgors dan Support Vector Classification. Analisis K-NN dan SVC dilakukan menggunakan Python dengan librarty “scikit learn” dan parameter default (tanpa pengaturan paerameter).



Gambar 8. Classification Matrix Model Machine Learning

Diperoleh tingkat akurasi dari model K-NN yang terbentuk adalah 90% dengan tingkat akurasi rata-rata klasifikasi untuk setiap kelasnya adalah 89%. Dengan demikian, diperoleh bahwa model K-NN yang terbentuk pada kasus ini dapat dengan baik melakukan klasifikasi autisme pada orang dewasa, sedangkan pada model SVC diperoleh tingkat akurasi sebesar 79% dengan tingkat akurasi rata-rata klasifikasi untuk setiap kelasnya adalah 72%. Pada model SVC yang terbentuk, model masih kurang baik dalam melakukan prediksi pada kelas “orang dewasa dengan autisme" dengan tingkat akurasi yang hanya 58%. Dengan demikian, dapat

dikatakan bahwa model SVC masih kurang baik dalam melakukan klasifikasi autisme pada orang dewasa.

KESIMPULAN DAN SARAN

Diperoleh bahwa tingkat akurasi dari klasifikasi sangat beragam di antra kelas yang di uji. Namun secara keseluruhan algoritma yang digunakan ANN dapat melakukan klasifikasi dengan lebih akurat jika dibandingkan dengan dua algoritma *machine learning* lainnya. ANN mendapat milai akurasi klasifikasi sebesar 92%, sedangkan K-NN dan SVC secara berturut-turut hanya mendapat nilai akurasi sebesar 90% dan 79%. Secara umum, pada kasus ini, algoritma *machine learning* konvensional masih kurang baik dalam melakukan prediksi pada kelas “orang dewasa dengan autisme”.

Hasil ini, membuktikan bahwa algoritma *machine learning* konvensional masih kurang mampu dalam melakukan klasifikasi pada data yang kecil dan tidak seimbang antar kelas, mengingat bahwa pada kasus ini hanya digunakan data dengan 700 observasi dan kelas “orang dewasa dengan autisme” hanya sejumlah 189 obsevasi.

Saran yang diberikan penulis pada pembaca, yaitu dalam kasus pengklasifikasian orang dewasa dengan autisme lebih baik jika menggunakan algoritma ANN karena memiliki tingkat akurasi yang tinggi baik secara umum, maupun pada setiap kelasnya.

Daftar Pustaka

Erkan, U., & Thanh, D. N. (2019). Autism Spectrum Disorder Detection with Machine Learning Methods. *Current Psychiatry Research and Reviews*, 1.

Jia Wu a, X.-Y. C.-D.-H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, 26-40.

Sujatha R, A. S., Chatterjee, J. M., Alaboudi, A., & Jhanjhi, N. Z. (2021). A Machine Learning Way to Classify Autism. *International Journal of Emerging Technologies in Learning (iJET)*, 188.

WHO, W. H. (den 15 November 2023). Autism. *WHO Newsroom Fact Sheets*.

Lampiran

Drive Dataset dan Syntax