



UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

Prediksi Jumlah Kasus Diabetes Tahunan Menggunakan Neural Network Berbasis Regresi

Disusun oleh Kelompok 2



UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

ANGGOTA KELOMPOK

- Muhammad Fakhri Ruslan (2206029935)
- Bryan Jonathan (2206052780)
- Muhammad Faris Naufaldi (2306153843)
- Natalius Desta Riyanto (2306202624)



UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

PENDAHULUAN

LATAR BELAKANG

Diabetes merupakan salah satu penyakit kronis dengan tren peningkatan signifikan setiap tahunnya, terutama di negara berkembang. Prediksi jumlah kasus diabetes sangat penting untuk membantu pemerintah dan instansi kesehatan dalam menyusun kebijakan, perencanaan fasilitas, serta upaya pencegahan yang lebih efektif.

Namun, tren jumlah kasus diabetes tidak selalu bersifat linier. Pola kenaikan bisa dipengaruhi oleh banyak faktor kompleks, seperti rata-rata kadar gula dikonsumsi, usia penduduk, hingga harga gula. Metode konvensional seperti regresi linier sering kali kurang mampu menangkap hubungan tersebut secara akurat.

Oleh karena itu, dalam studi ini digunakan pendekatan berbasis Neural Network (Keras) untuk memprediksi jumlah kasus diabetes dari tahun 2010 hingga 2024. Model ini diharapkan mampu mengenali pola non-linear dan memberikan hasil prediksi yang lebih akurat sebagai dasar pengambilan keputusan.



UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

RUMUSAN MASALAH

- Bagaimana membangun model neural network untuk memprediksi kenaikan kasus diabetes per tahun?
- Seberapa akurat model yang telah dibuat dalam melakukan prediksi kenaikan kasus diabetes per tahun?



UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

TUJUAN

- Membuat model untuk memprediksi jumlah kenaikan kasus diabetes per tahun dengan neural network
- Mengukur akurasi model dan mengetahui tren kenaikan kasus diabetes per tahun yang diprediksi



UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

DATASET DAN PREPROCESS DATA

UNIVERSITAS
INDONESIA

Virtus Proficit Suficit

FMIPA

DATASET

Dataset ini terdiri dari data tahunan periode 2010–2024 yang mencakup variabel-variabel utama yang dianggap relevan terhadap jumlah kasus diabetes, yaitu :

- Populasi (Juta) : total populasi Indonesia per tahun
- Rata-rata kadar gula per minggu (ons) : estimasi konsumsi gula per orang
- Total populasi umur >40 (Juta) : populasi berisiko tinggi terkena diabetes
- Harga gula per 1 kg: sebagai proksi ekonomi dan konsumsi
- Total Kasus Diabetes (Juta): data target (label) yang akan diprediksi

Kelemahan Dataset

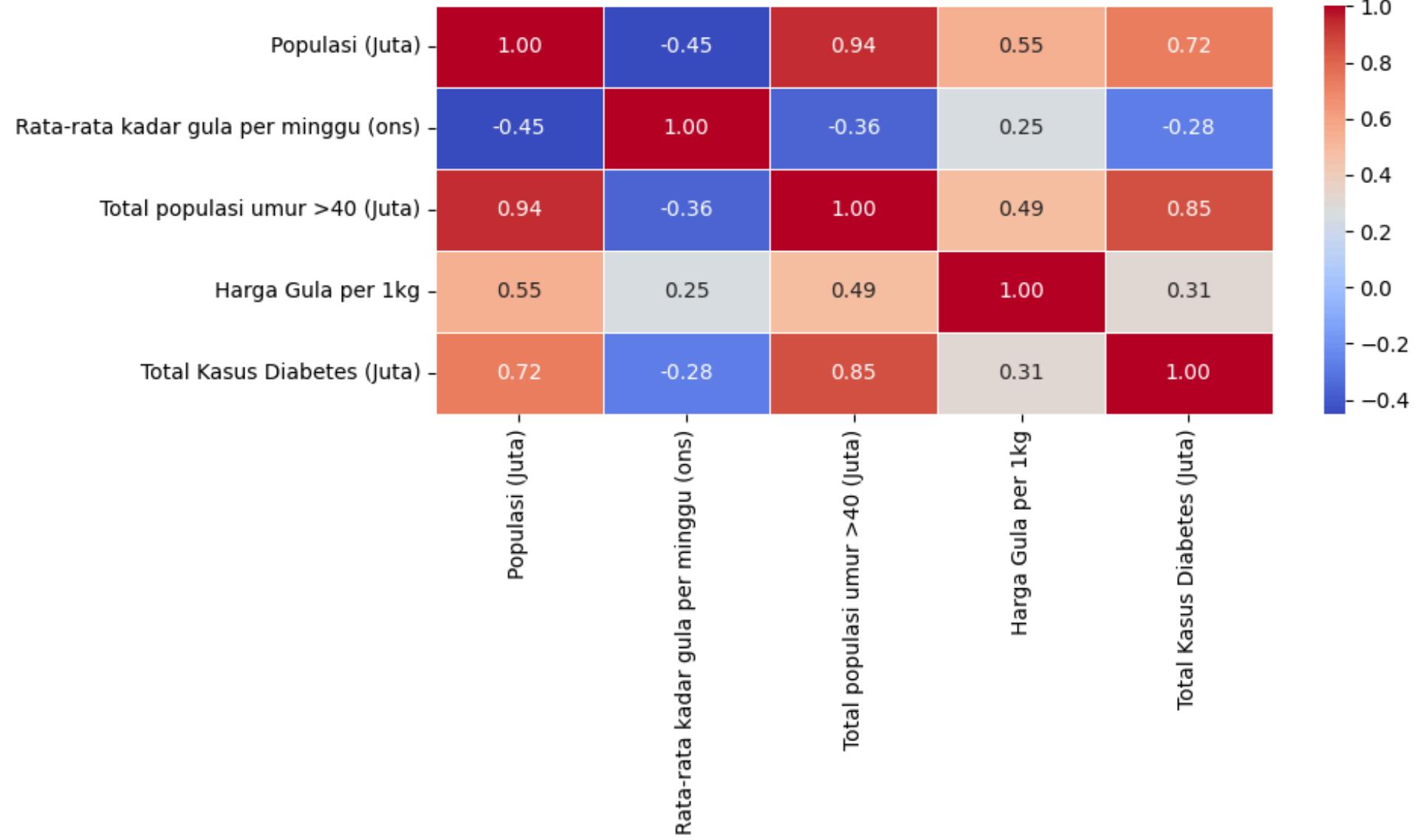
- Jumlah data terbatas: hanya mencakup 15 tahun (2010–2024), sehingga model memiliki keterbatasan dalam belajar pola jangka panjang.
- Potensi overfitting: karena data sedikit, model kompleks seperti neural network berisiko belajar "terlalu dalam" terhadap data latih.

Tahun	Populasi (Juta)	Rata-rata kadar gula per minggu (ons)	Total populasi umur >40 (Juta)	Harga Gula per 1kg	Total Kasus Diabetes (Juta)
2010	237,6	1,617	70,73121	12.000	6,9
2011	249,5	1,555	71,441722	12.000	7,3
2012	252,7	1,344	74,370283	12.170	7
2013	255,9	1,38	77,305066	13.041	7,6
2014	258,9	1,328	79,642965	11.782	9
2015	261,8	1,441	81,397514	12.570	9,1
2016	264,6	1,581	83,947193	15.715	9,6
2017	266,9	1,462	86,492771	14.378	10,3
2018	267,3	1,433	88,6273	13.010	10,2
2019	268	1,391	91,1508	12.917	10,7
2020	270,2	1,376	93,6748	13.596	18,69
2021	272,6	1,415	96,3164	13.236	19,47
2022	275,7	1,337	99,1444	13.392	19,5
2023	278,6	1,228	101,6368	14.327	20,4
2024	281,6	1,133	104,3365	18.628	20

KORELASI ANTAR FITUR

- Dari hasil korelasi, fitur yang paling kuat hubungannya dengan total kasus diabetes adalah Total populasi umur >40 tahun, dengan nilai korelasi sebesar 0.85. Ini berarti semakin banyak populasi usia lanjut, semakin tinggi pula jumlah kasus diabetes. Korelasi ini sangat kuat dan positif.
- Fitur Populasi (Juta) juga menunjukkan hubungan positif yang kuat (0.72), artinya secara umum, peningkatan jumlah penduduk juga disertai peningkatan kasus diabetes. Hal ini masuk akal karena jumlah penduduk menjadi dasar proporsional penyebaran kasus.
- Harga Gula per 1kg memiliki korelasi positif lemah (0.31) dengan kasus diabetes. Ini menunjukkan bahwa harga gula mungkin punya sedikit pengaruh terhadap tren kasus diabetes, tetapi tidak dominan. Kemungkinan besar ini hanya mewakili faktor ekonomi atau konsumsi secara tidak langsung.
- Sementara itu, Rata-rata kadar gula per minggu (ons) justru menunjukkan korelasi negatif sebesar -0.28, yang berarti semakin tinggi kadar gula yang dikonsumsi justru sedikit menurunkan jumlah kasus menurut data ini. Namun korelasi ini sangat lemah, dan bisa disebabkan oleh noise atau ketidaksesuaian antar skala variabel.

Heatmap Korelasi antar Fitur dan Target (Train Set)





UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

Pada projek kami, kami membagi dataset menjadi data train dan data testing. Data tahun 2010-2021 kami buat sebagai data train. Data tahun 2022-2024 kami buat sebagai data testing.

SPLIT DATASET

Tahun	Populasi (Juta)	Rata-rata kadar gula per minggu (ons)	Total populasi umur >40 (Juta)	Harga Gula per 1kg	Total Kasus Diabetes (Juta)
2010	237,6	1,617	70,73121	12.000	6,9
2011	249,5	1,555	71,441722	12.000	7,3
2012	252,7	1,344	74,370283	12.170	7
2013	255,9	1,38	77,305066	13.041	7,6
2014	258,9	1,328	79,642965	11.782	9
2015	261,8	1,441	81,397514	12.570	9,1
2016	264,6	1,581	83,947193	15.715	9,6
2017	266,9	1,462	86,492771	14.378	10,3
2018	267,3	1,433	88,6273	13.010	10,2
2019	268	1,391	91,1508	12.917	10,7
2020	270,2	1,376	93,6748	13.596	18,69
2021	272,6	1,415	96,3164	13.236	19,47
2022	275,7	1,337	99,1444	13.392	19,5
2023	278,6	1,228	101,6368	14.327	20,4
2024	281,6	1,133	104,3365	18.628	20

STANDARISASI DATASET

Karena satuan pada beberapa fitur yang berbeda, kami melakukan standarisasi pada setiap fitur.

Apa itu Standarisasi?

Standarisasi adalah proses mengubah skala fitur agar memiliki:

- Rata-rata (mean) = 0
- Standar deviasi (std) = 1

Rumusnya:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Di mana:

- X = nilai asli
- μ = rata-rata (mean) dari training set
- σ = standar deviasi dari training set

Tahun	Populasi (Juta)	Rata-rata kadar gula per minggu (ons)	Total populasi umur >40 (Juta)	Harga Gula per 1kg	Total Kasus Diabetes (Juta)
2010	237,6	1,617	70,73121	12.000	6,9
2011	249,5	1,555	71,441722	12.000	7,3
2012	252,7	1,344	74,370283	12.170	7
2013	255,9	1,38	77,305066	13.041	7,6
2014	258,9	1,328	79,642965	11.782	9
2015	261,8	1,441	81,397514	12.570	9,1
2016	264,6	1,581	83,947193	15.715	9,6
2017	266,9	1,462	86,492771	14.378	10,3
2018	267,3	1,433	88,6273	13.010	10,2
2019	268	1,391	91,1508	12.917	10,7
2020	270,2	1,376	93,6748	13.596	18,69
2021	272,6	1,415	96,3164	13.236	19,47
2022	275,7	1,337	99,1444	13.392	19,5
2023	278,6	1,228	101,6368	14.327	20,4
2024	281,6	1,133	104,3365	18.628	20

STANDARISASI DATASET



Sebelum Standarisasi

1. Populasi (Juta) – Garis Biru
- Nilai absolut jauh lebih besar dibanding fitur lain.
 - Paling dominan, algoritma model akan menganggap fitur ini sangat penting hanya karena skalanya besar.
 - Bisa menyebabkan model terlalu “memperhatikan” populasi dibanding fitur lainnya.

2. Total Populasi Umur >40 (Juta) – Garis Hijau

- Nilainya masih cukup besar, tapi tidak sedominan populasi total.
- Cukup berpengaruh, tapi tetap di bawah populasi total.

3. Harga Gula per 1kg – Garis Merah

- Nilainya tampak hampir mendatar karena jauh lebih kecil skalanya.
- Terlihat kurang berpengaruh secara numerik padahal bisa jadi relevan.

4. Rata-rata Kadar Gula per Minggu (ons) – Garis Orange

- Nilai sangat kecil, nyaris flat di plot atas.
- Paling tidak berpengaruh secara numerik, padahal mungkin signifikan secara biologis atau medis.

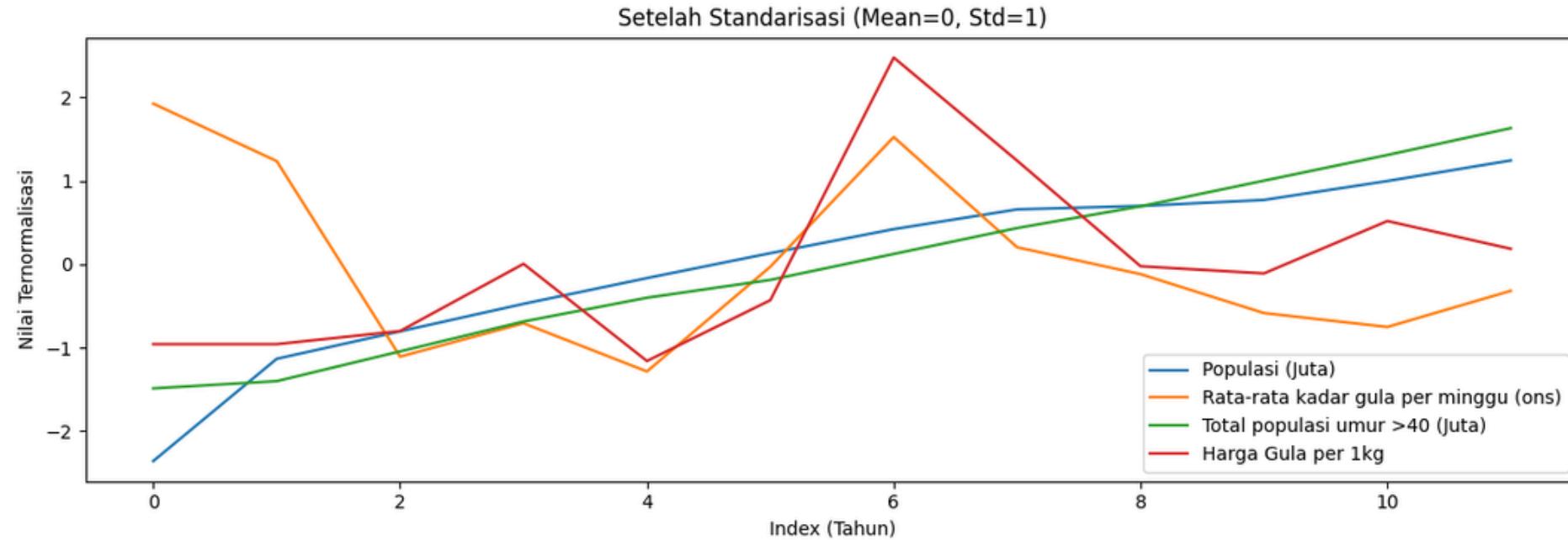
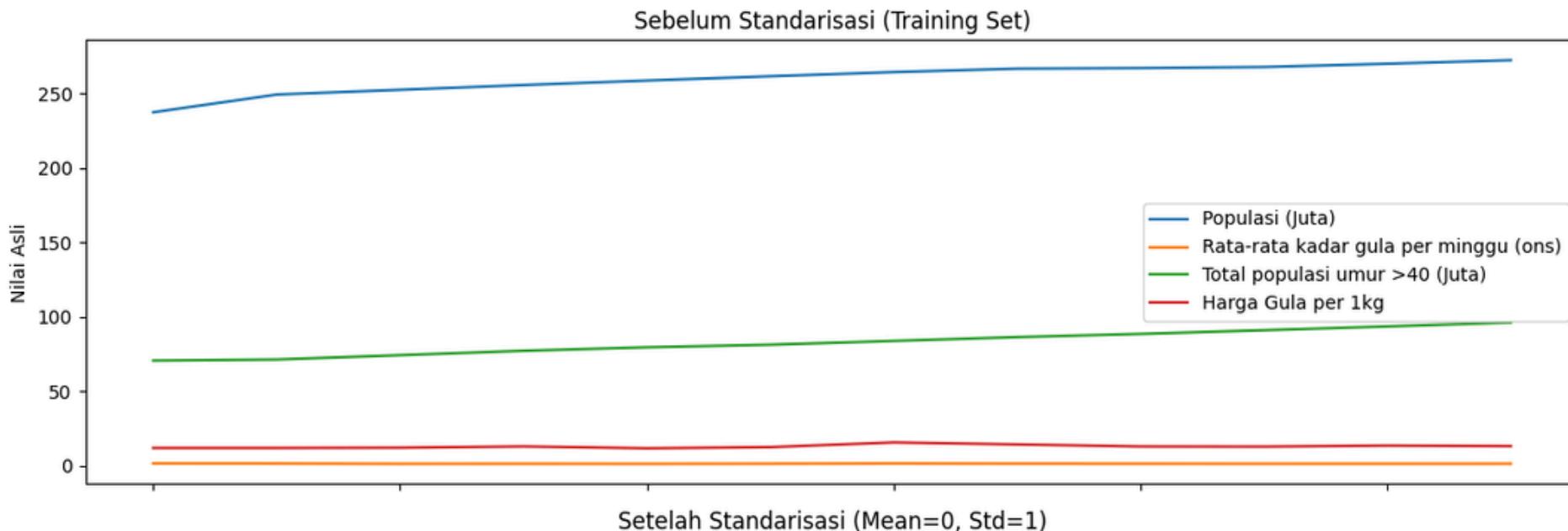
Setelah Standarisasi

Setelah dilakukan StandardScaler, seluruh fitur:

- Memiliki rata-rata = 0
- Standar deviasi = 1

Ini membuat semua fitur berada dalam skala yang setara, sehingga:

- Model bisa belajar dengan adil dari semua fitur.
- Proses pelatihan menjadi lebih stabil dan efisien.



METODOLOGI

EARLY STOPPING & LEARNING RATE SCHEDULE

• Early Stopping

Artinya Model akan berhenti jika val_loss tidak membaik selama 10 epoch berturut-turut. Digunakan untuk menghentikan pelatihan secara otomatis saat model tidak lagi menunjukkan peningkatan pada data validasi.

- Bobot terbaik selama training akan dipulihkan (restore_best_weights=True).
- Tujuannya untuk menghindari overfitting dan menghemat waktu pelatihan.

• Learning Rate Scheduler

Learning rate adalah parameter penting yang menentukan seberapa besar langkah pembelajaran model. Digunakan ReduceLROnPlateau untuk menurunkan learning rate secara otomatis saat model stagnan. Artinya:

- Jika val_loss tidak membaik selama 5 epoch, learning rate akan dikalikan 0.5.
- Tidak akan turun di bawah $1e-10$.

MODEL NEURAL NETWORK

Input Layer (\mathbb{R}^4)

- Terdiri dari 4 neuron, masing-masing mewakili satu fitur:
 - Populasi (Juta)
 - Rata-rata kadar gula per minggu
 - Total populasi usia >40 tahun
 - Harga gula per 1kg

Output Layer (\mathbb{R}^1)

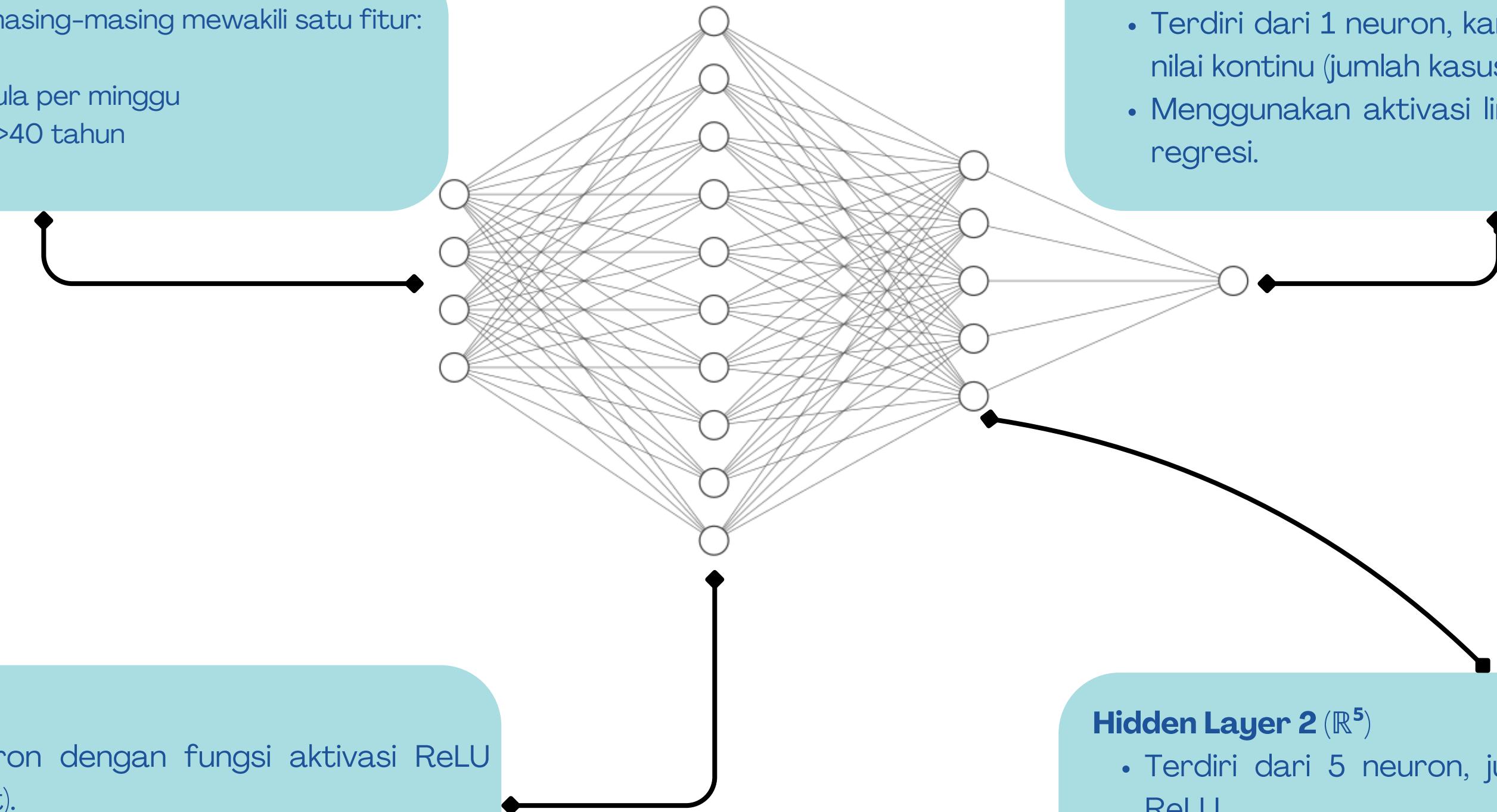
- Terdiri dari 1 neuron, karena target prediksi adalah nilai kontinu (jumlah kasus diabetes).
- Menggunakan aktivasi linear, sesuai untuk masalah regresi.

Hidden Layer 1 (\mathbb{R}^{10})

- Terdiri dari 10 neuron dengan fungsi aktivasi ReLU (Rectified Linear Unit).
- Ini adalah lapisan yang pertama kali memproses input dan mulai belajar pola non-linear.

Hidden Layer 2 (\mathbb{R}^5)

- Terdiri dari 5 neuron, juga menggunakan aktivasi ReLU.
- Bertugas menyempurnakan representasi fitur sebelum dipetakan ke output.



FUNGSI AKTIVASI YANG DIGUNAKAN

1. ReLU (Rectified Linear Unit)

- Digunakan di hidden layer.
- Fungsi:

$$f(x) = \max(0, x)$$

- Kenapa dipakai?
 - Sederhana dan cepat dihitung.
 - Menghindari masalah vanishing gradient.
 - Mampu memperkenalkan non-linearitas dalam jaringan.

2. Linear

- Digunakan di output layer.
- Fungsi:

$$f(x) = x$$

- Kenapa dipakai?
 - Cocok untuk regresi, karena kita ingin prediksi berupa angka kontinu (jumlah kasus diabetes).
 - Tidak mengubah bentuk output terakhir → memudahkan evaluasi dengan MSE atau MAE.

METRIK EVALUASI

1. Mean Squared Error (MSE)

Metrik utama yang digunakan saat training dan testing.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Mengukur rata-rata kuadrat selisih antara nilai aktual dan prediksi.
- Sensitif terhadap error besar.
- Semakin kecil MSE, semakin baik model.

2. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- Mengukur rata-rata dari selisih absolut antara nilai aktual dan prediksi.
- Lebih robust terhadap outlier dibanding MSE.

METRIK EVALUASI

3. R² Score (Koefisien Determinasi)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Mengukur seberapa besar variasi data yang bisa dijelaskan oleh model.
- Nilai mendekati 1 → model sangat baik.
- Nilai mendekati 0 → model buruk (tidak lebih baik dari rata-rata).
- Jika nilai R² negatif, artinya prediksi model lebih buruk daripada tebakan rata-rata.

Ketiga metrik ini digunakan untuk menilai akurasi model dalam memprediksi data testing (2022–2024).

OUTPUT MODEL

TRAINING LOSS VS VAL LOSS

1. Pola Awal (Epoch 0–5)

- Terlihat training loss turun cepat, sedangkan validation loss fluktuatif namun masih tinggi.
- Ini menunjukkan model masih adaptasi awal pada data latih, namun belum mampu generalisasi ke data validasi.

2. Tanda Overfitting (Epoch 6–10)

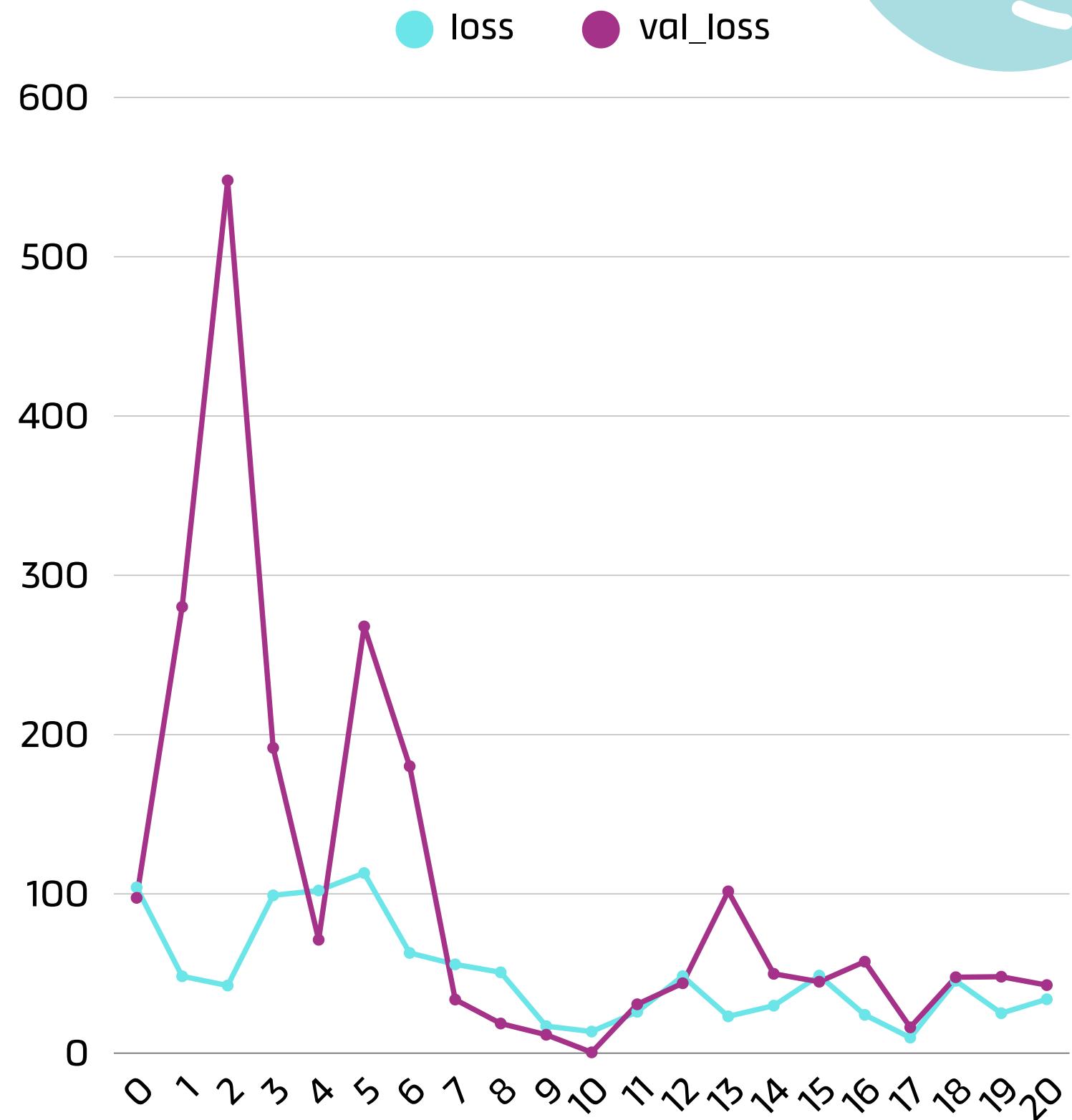
- Training loss terus menurun, namun validation loss meningkat tajam di beberapa titik (terlihat spike besar).
- Hal ini menunjukkan model mulai menghafal data training (overfitting) dan gagal generalisasi ke data baru.
- Model menjadi sensitif terhadap noise pada data latih.

3. Penyesuaian Learning Rate (Epoch 11–14)

- Setelah ReduceLROnPlateau aktif, validation loss mulai menurun kembali.
- Ini mengindikasikan bahwa learning rate yang lebih kecil membantu model keluar dari kondisi stagnan dan memperbaiki performa validasi.

4. Stabilisasi Akhir (Epoch 15–20)

- Di akhir, baik training loss maupun validation loss menjadi lebih stabil.
- Tidak ada lagi penurunan besar, menandakan model mulai mencapai titik optimal dan early stopping akan segera aktif jika tidak ada perbaikan lagi.



PLOT SELURUH TAHUN

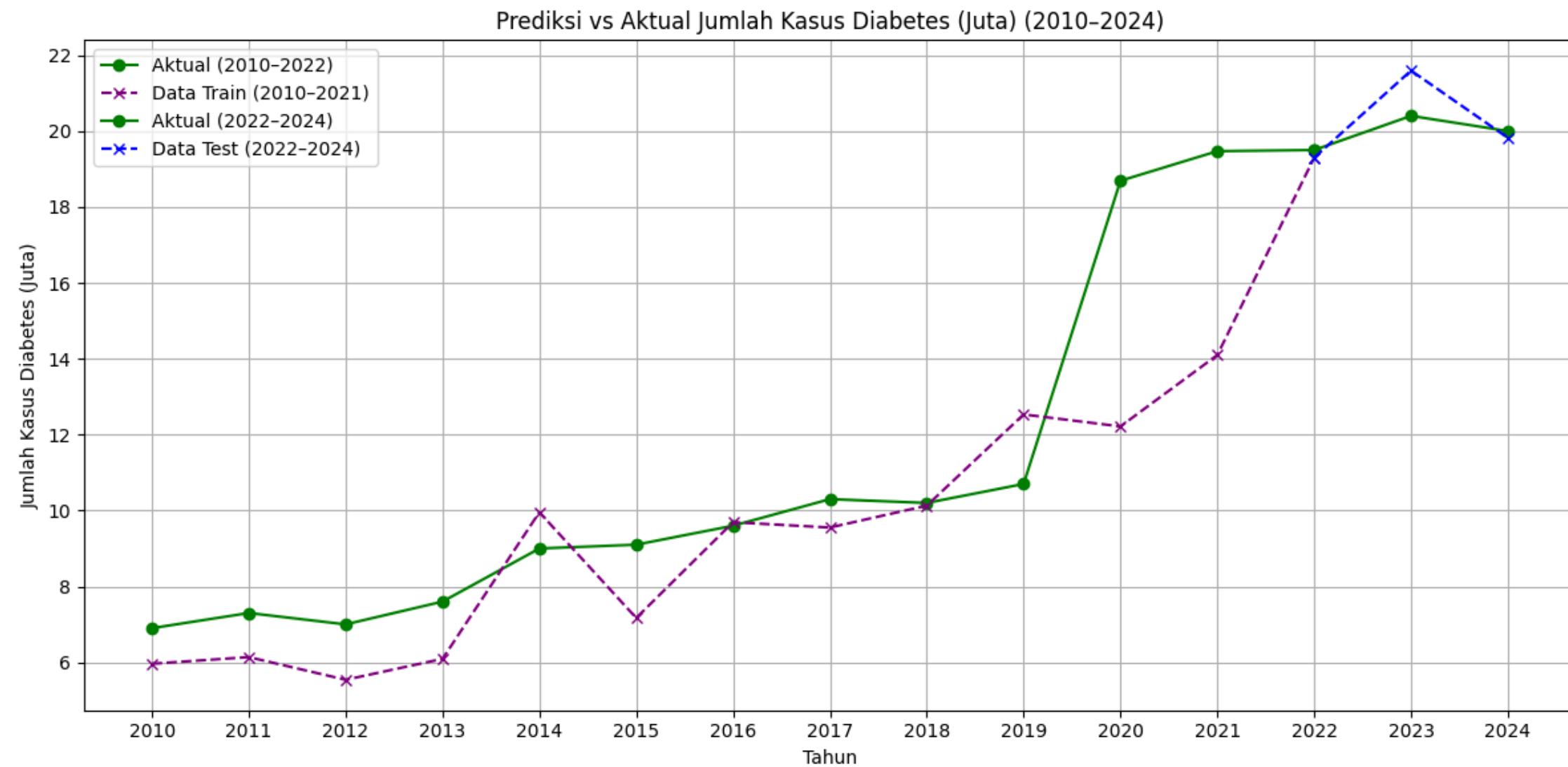
Plot ini bukan untuk menilai performa keseluruhan, melainkan untuk menggambarkan bagaimana model mempelajari tren data historis (2010–2021) dan bagaimana hasil prediksinya diuji pada data baru (2022–2024).

1. Periode Data Latih (2010–2021):

- Garis hijau dan ungu cukup dekat, menunjukkan model mampu mengikuti tren umum kenaikan jumlah kasus diabetes dari tahun ke tahun.
- Fluktuasi kecil pada prediksi bisa terjadi karena jumlah data yang terbatas, namun tidak menunjukkan kesalahan prediksi ekstrem.

2. Periode Data Uji (2022–2024):

- Garis biru (prediksi) dan hijau (aktual) masih relatif dekat, menunjukkan bahwa model berhasil menangkap arah tren, meskipun ada sedikit overestimasi di tahun 2023.
- Hal ini wajar, karena data uji tidak pernah dilihat oleh model saat training.



PLOT PREDIKSI DATA TEST

1. Tahun 2022:

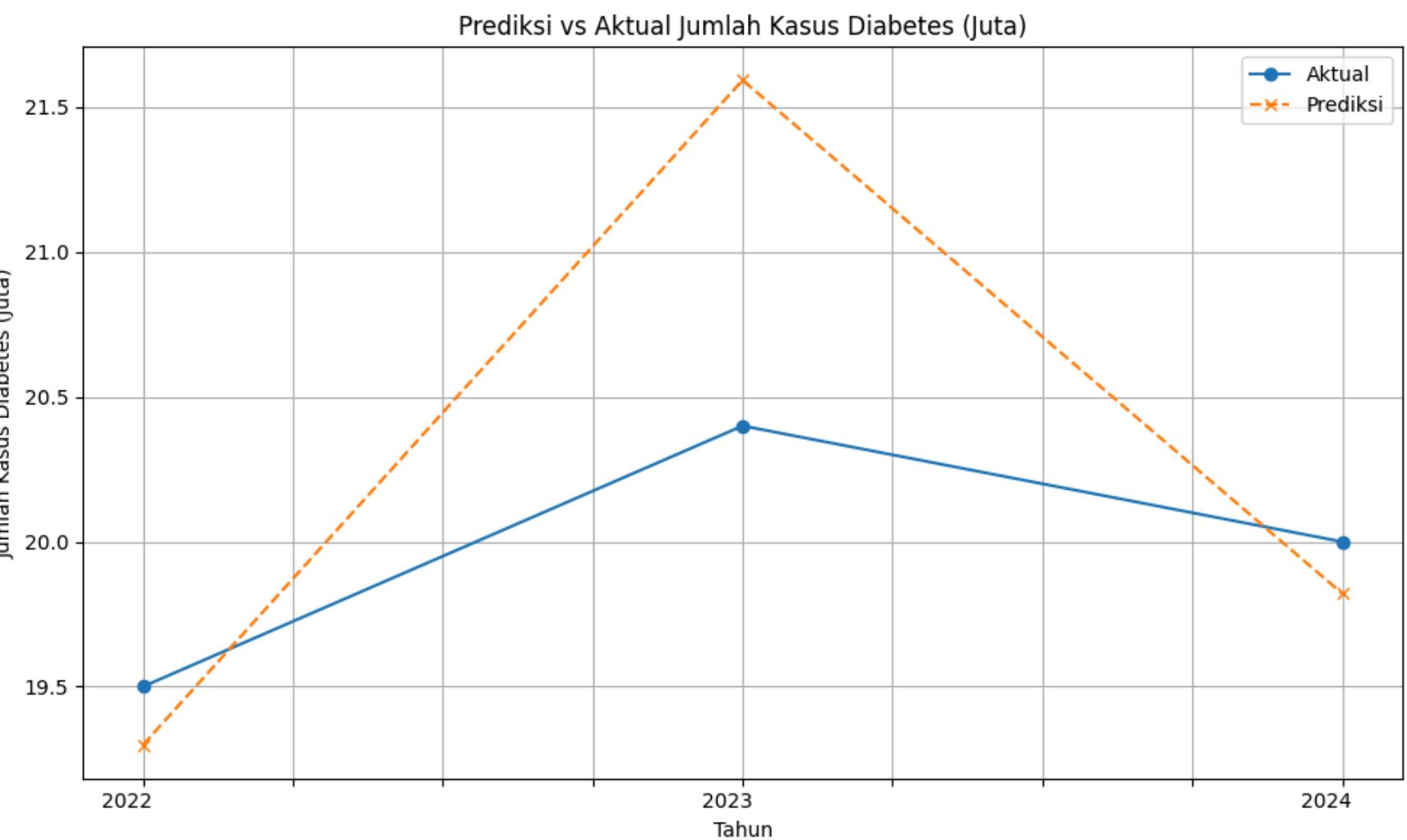
- Aktual: 19.5 juta kasus
- Prediksi: 19.30 juta kasus
- Model sedikit underestimate → prediksi sedikit lebih rendah dari aktual (~0.2 juta), tapi sangat dekat → akurasi tinggi.

2. Tahun 2023:

- Aktual: 20.4 juta kasus
- Prediksi: 21.59 juta kasus
- Model overestimate sekitar 1.2 juta kasus → prediksi agak meleset, tapi tetap mengikuti tren naik dari tahun sebelumnya.

3. Tahun 2024:

- Aktual: 20.0 juta kasus
- Prediksi: 19.82 juta kasus
- Model kembali underestimate sedikit, selisih hanya ~0.18 juta → masih sangat mendekati aktual.



Kesimpulan:

- Akurasi model cukup baik untuk ketiga tahun terakhir (data test).
- Prediksi mengikuti pola tren aktual meskipun ada deviasi kecil di tiap titik.
- Ini membuktikan bahwa model berhasil melakukan generalisasi, terutama untuk memprediksi tren diabetes.

Tahun	Aktual	Prediksi
2022	19.5	19.298351
2023	20.4	21.592810
2024	20.0	19.821127

EVALUASI

1. MSE (Mean Squared Error): 0.50

- Mengukur rata-rata dari kuadrat selisih antara nilai aktual dan nilai prediksi.
- MSE yang rendah menunjukkan kesalahan prediksi yang kecil, namun karena bersifat kuadrat, outlier sangat memengaruhi nilai ini.
- Dalam konteks ini, MSE sebesar 0.50 juta kasus² tergolong cukup rendah, menandakan prediksi relatif dekat dengan nilai aktual.

2. MAE (Mean Absolute Error): 0.52

- Mengukur rata-rata selisih absolut antara nilai aktual dan prediksi.
- Lebih mudah diinterpretasi dibanding MSE karena berada dalam satuan yang sama dengan target (juta kasus).
- MAE sebesar 0.52 juta kasus berarti rata-rata prediksi meleset ±520 ribu kasus, menunjukkan tingkat akurasi yang cukup.

3. R² Score (Koefisien Determinasi): -2.6773

- Mengukur seberapa besar variasi pada data target yang bisa dijelaskan oleh model.
- Nilai idealnya mendekati 1, dan nilai < 0 menunjukkan bahwa model lebih buruk daripada hanya menebak rata-rata.
- R² negatif ini mengindikasikan bahwa meskipun error absolut rendah, model belum menangkap pola hubungan antar fitur dan target dengan baik dalam konteks regresi.

Tahun	Aktual	Prediksi
2022	19.5	19.298351
2023	20.4	21.592810
2024	20.0	19.821127
Test MSE: 0.50, MAE: 0.52		
1/1 0s 89ms/step		
R ² Score: -2.6773		

KESIMPULAN

1. Penelitian ini berhasil membangun model prediksi jumlah kasus diabetes menggunakan metode Neural Network dengan data tahun 2010–2024.
2. Model mampu mengikuti arah tren peningkatan maupun penurunan kasus, khususnya pada tahun 2022 dan 2024 yang menunjukkan prediksi mendekati nilai aktual.
3. Namun, pada tahun 2023 terjadi deviasi terbesar (overestimate ~1,2 juta kasus), menunjukkan adanya kelemahan dalam menangani anomali atau fluktuasi tajam.
4. Hasil evaluasi:
 - MAE = 0.52 juta kasus → kesalahan rata-rata masih tergolong rendah.
 - MSE = 0.50 juta → menunjukkan stabilitas prediksi.
 - $R^2 = -2.68$ → model belum dapat menjelaskan variasi target secara baik dalam konteks regresi.
5. Kekuatan model: mampu menangkap pola umum dan cukup akurat dalam nilai prediksi absolut.
6. Keterbatasan model: belum optimal dalam memahami kompleksitas hubungan antar fitur, serta sangat sensitif terhadap fitur yang tersedia.

SARAN DAN PENGEMBANGAN MODEL KE DEPAN

1. Evaluasi lebih lanjut perlu dilakukan dengan data uji yang lebih panjang agar hasil metrik (terutama R^2) lebih stabil dan representatif.
2. Perlu ditambahkan fitur relevan lain, misalnya konsumsi makanan manis, aktivitas fisik, atau prevalensi obesitas, untuk meningkatkan kekuatan prediksi.
3. Fine tuning lebih lanjut, terutama untuk mengurangi deviasi besar seperti yang terjadi di tahun 2023.

DAFTAR PUSTAKA

1. Badan Pusat Statistik. (2024). Jumlah Penduduk Menurut Kelompok Umur dan Jenis Kelamin, 2024. <https://www.bps.go.id/id/statistics-table/3/WVc0MGEyMXBkVFUxY25KeE9HdDZkbTQzWkVkb1p6MDkjMw==/jumlah-penduduk-menurut-kelompok-umur-dan-jenis-kelamin--2023.html?year=2024>
2. Chollet, F. (2017). Deep Learning with Python. Manning Publications.
3. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and Tensor-Flow. O'Reilly Media.
4. International Diabetes Federation. (2025). Indonesia Diabetes Trends & Prevalence. <https://diabetesatlas.org/data-by-location/country/indonesia/>
5. Kementerian Kesehatan Republik Indonesia. (2023). Profil Kesehatan Indonesia 2022. <https://www.kemkes.go.id>
6. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.



UNIVERSITAS
INDONESIA

Veritas, Probitas, Justitia

FMIPA

**THANK YOU
ANY QUESTIONS?**