ANALISIS DETERMINAN PENYEBAB *STUNTING* PROVINSI DI INDONESIA: APLIKASI MODEL *RANDOM FOREST* DAN INTERPRETASI SHAP PADA DATA SSGI TAHUN 2024

TUGAS AKHIR

Diajukan sebagai syarat menyelesaikan jenjang strata Satu (S-1) di
Program Studi Teknik Informatika, Fakultas Teknologi Industri,
Institut Teknologi Sumatera

Oleh:

Cindy Nadila Putri

122140002

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2026

LEMBAR PENGESAHAN

Saya menyatakan bahwa Tugas Akhir berjudul "Analisis Determinan Penyebab *Stunting* Provinsi di Indonesia: Aplikasi Model *Random Forest* dan Interpretasi SHAP pada Data SSGI Tahun 2024" merupakan hasil karya saya sendiri dan belum pernah diajukan, baik sebagian maupun seluruhnya, di Institut Teknologi Sumatera atau institusi pendidikan lain oleh saya maupun pihak lain.

Lampung Selatan, 30 Januari 2026
Penulis,

Foto 2x3

Cindy Nadila Putri
NIM. 122140002

Diperiksa dan disetujui oleh,

Pembimbing

1. Martin Clinton Tosima Manullang, Ph.D.
   NIP. 19930109 2019 03 1 017 .....

2. Martin Clinton Tosima Manullang, Ph.D.
   NIP. 19930109 2019 03 1 017 .....

Penguji

1. Dosen Penguji I
   NIP. 19900000 2000 00 0 000 .....
2. Dosen Penguji II
   NIP. 19900000 2000 00 0 000 .....

Disahkan oleh,
Koordinator Program Studi Teknik Informatika
Fakultas Teknologi Industri
Institut Teknologi Sumatera

Andika Setiawan, S.Kom., M.Cs.
NIP. 19911127 2022 03 1 007

HALAMAN PERNYATAAN ORISINALITAS

Tugas Akhir dengan judul "Analisis Determinan Penyebab *Stunting* Provinsi di Indonesia: Aplikasi Model *Random Forest* dan Interpretasi SHAP pada Data SSGI Tahun 2024" adalah karya saya sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan benar.

Nama : Cindy Nadila Putri

NIM : 122140002

Tanda Tangan : .....

Tanggal : .....

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Institut Teknologi Sumatera, saya yang bertanda tangan di bawah ini:

Nama : Cindy Nadila Putri
NIM : 122140002
Program Studi : Teknik Informatika
Fakultas : Teknologi Industri
Jenis Karya : Tugas Akhir

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Institut Teknologi Sumatera **Hak Bebas Royalti Noneksklusif** ***(Non-exclusive Royalty Free Right)*** atas karya ilmiah saya yang berjudul:

Analisis Determinan Penyebab *Stunting* Provinsi di Indonesia: Aplikasi Model *Random Forest* dan Interpretasi SHAP pada Data SSGI Tahun 2024

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Institut Teknologi Sumatera berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Lampung Selatan
Pada tanggal : 30 Januari 2026

Yang menyatakan

Cindy Nadila Putri

KATA PENGANTAR

Pada halaman ini mahasiswa berkesempatan untuk menyatakan terima kasih secara tertulis kepada pembimbing dan pihak lain yang telah memberi bimbingan, nasihat, saran dan kritik, kepada mereka yang telah membantu melakukan penelitian, kepada perorangan atau lembaga yang telah memberi bantuan keuangan, materi dan/atau sarana. Cara menulis kata pengantar beraneka ragam, tetapi hendaknya menggunakan kalimat yang baku. Ucapan terima kasih agar dibuat tidak berlebihan dan dibatasi pada pihak yang terkait secara ilmiah (berhubungan dengan subjek/materi penelitian).

Puji syukur kehadirat Allah SWT/Tuhan Yang Maha Esa atas limpahan rahmat, karunia, serta petunjuk-Nya sehingga penyusunan tugas akhir ini telah terselesaikan dengan baik. Dalam penyusunan tugas akhir ini penulis telah banyak mendapatkan arahan, bantuan, serta dukungan dari berbagai pihak. Oleh karena itu pada kesempatan ini penulis mengucapan terima kasih kepada:

1. [Rektor ITERA] selaku Rektor Institut Teknologi Sumatera.
2. [Dekan FTI] selaku Dekan Fakultas Teknologi Industri.
3. [Koor Prodi IF] selaku Ketua Program Studi Teknik Informatika.
4. [Dosen Pembimbing] selaku Dosen Pembimbing atas ide, waktu, tenaga, perhatian, dan masukan yang telah disumbangsihkan kepada penulis.
5. [Isi nama lainnya]

Akhir kata penulis berharap semoga tugas akhir ini dapat memberikan manfaat bagi kita semua.

RINGKASAN

Analisis Determinan Penyebab *Stunting* Provinsi di Indonesia: Aplikasi Model *Random Forest* dan Interpretasi SHAP pada Data SSGI Tahun 2024

Cindy Nadila Putri

Halaman Ringkasan berisi uraian singkat tentang latar belakang masalah, rumusan masalah, tujuan, metodologi penelitian, hasil dan analisis data, serta kesimpulan dan saran. Isi ringkasan tidak lebih dari 1000 kata (sekitar maksimal 2 halaman).

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

ABSTRAK

Analisis Determinan Penyebab *Stunting* Provinsi di Indonesia: Aplikasi Model *Random Forest* dan Interpretasi SHAP pada Data SSGI Tahun 2024

Cindy Nadila Putri

Halaman ABSTRAK berisi uraian tentang latar belakang, tujuan, metodologi penelitian, hasil / kesimpulan. Ditulis dalam BAHASA INDONESIA tidak lebih dari 250 kata, dengan jarak antar baris satu spasi. Pada akhir abstrak ditulis kata “Kata Kunci” yang dicetak tebal, diikuti tanda titik dua dan kata kunci yang tidak lebih dari 5 kata. Kata kunci terdiri dari kata-kata yang khusus menunjukkan dan berkaitan dengan bahan yang diteliti, metode/instrumen yang digunakan, topik penelitian. Kata kunci diketik pada jarak dua spasi dari baris akhir isi abstrak.

Kata Kunci: kunci1, kunci2

ABSTRACT

Determinant Analysis of Stunting Causes Across Indonesian Provinces: Application of Random Forest Model and SHAP Interpretation on SSGI 2024 Data

Cindy Nadila Putri

Halaman ABSTRACT berisi uraian tentang latar belakang, tujuan, metodologi penelitian, hasil / kesimpulan. Ditulis dalam BAHASA INGGRIS tidak lebih dari 250 kata, dengan jarak antar baris satu spasi. Secara khusus, kata dan kalimat pada halaman ini tidak perlu ditulis dengan huruf miring meskipun menggunakan Bahasa Inggris, kecuali terdapat huruf asing lain yang ditulis dengan huruf miring (misalnya huruf Latin atau Greek, dll). Pada akhir abstract ditulis kata “Keywords” yang dicetak tebal, diikuti tanda titik dua dan kata kunci yang tidak lebih dari 5 kata. Keywords terdiri dari kata-kata yang khusus menunjukkan dan berkaitan dengan bahan yang diteliti, metode/instrumen yang digunakan, topik penelitian. Keywords diketik pada jarak dua spasi dari baris akhir isi abstrak.

Keywords: keywords1, keywords2

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR RUMUS

DAFTAR KODE

BAB I
PENDAHULUAN

1.1 Latar Belakang

Kualitas sumber daya manusia yang unggul sangat ditentukan oleh status gizi dan kesehatan yang optimal pada masa pertumbuhan. Di Indonesia, pencapaian kualitas tersebut masih terhambat oleh tingginya prevalensi masalah gizi kronis atau yang dikenal sebagai *stunting*. Masalah ini telah menjadi perhatian serius karena dampaknya yang meluas, terutama pada anak-anak sebagai generasi penerus bangsa. Berbagai faktor, seperti pola hidup, akses terhadap makanan bergizi, serta kualitas pelayanan kesehatan, turut berkontribusi terhadap tingginya prevalensi masalah ini.

Penyakit *stunting* merupakan kondisi di mana tubuh mengalami kekurangan gizi secara berlebihan dan terjadi pada rentang waktu yang cukup lama **said2024pencegahan**. Dampak dari masalah ini akan mengakibatkan kendala pertumbuhan pada anak sehingga tinggi badan anak cenderung menjadi lebih pendek. Tak hanya memengaruhi pertumbuhan fisik, kondisi *stunting* juga akan berpengaruh ke dalam aspek pertumbuhan lainnya seperti mental, intelektual, dan kognitif anak **rahagia2023upaya**. Oleh karena itu, penting untuk memahami faktor-faktor yang menyebabkan terjadinya *stunting* agar dapat mengambil langkah-langkah pencegahan yang efektif.

Di Indonesia, kasus *stunting* terjadi pada balita usia 0-5 tahun berada pada persentase sebesar 19,8% menurut Survei Status Gizi Indonesia (SSGI) **kemenkes2025ssgi**. Meskipun telah mengalami penurunan dari tahun sebelumnya yaitu angka 21,5%, hasil ini menunjukan bahwa target pemerintah Indonesia yaitu menurunkan prevalensi *stunting* sampai 14,4% di tahun 2029 masih belum tercapai **bkpk2023ski**, **setneg2024strategi**. *Stunting* dapat berasal dari faktor-faktor yang sangat kompleks dilihat dari aspek sosial,

biologis, maupun lingkungan. Biasanya, penyebab utama pada *stunting* dapat terjadi karena kurangnya asupan gizi, buruknya sanitasi, serta akses yang rendah terhadap pelayanan kesehatan **supriyanto2023implementasi**, **saleh2023edukasi**. Terdapat faktor penting lain yang menyebabkan terjadinya stunting yaitu kurangnya edukasi ibu tentang betapa penting untuk menjaga gizi seimbang pada masa kehamilan, masa menyusui, dan masa pertumbuhan anak **saleh2023edukasi**. Pada beberapa kasus, ibu hamil yang melahirkan bayi dengan kondisi berat badan lahir rendah (BBLR) mengalami fase kekurangan gizi selama masa kehamilannya **saleh2023edukasi**. Hal ini dapat mengakibatkan terjadinya *stunting* di kemudian hari.

Tingkat *stunting* di Indonesia sendiri dinilai tidak merata di seluruh wilayah. Beberapa provinsi seperti Bali dan DI Yogyakarta menunjukkan angka prevalensi yang relatif rendah, berbanding terbalik dengan provinsi Nusa Tenggara Timur dan Sulawesi Barat yang mencatat angka jauh lebih tinggi **kemenkes2025ssgi**. Situasi yang timpang ini menunjukkan adanya variasi faktor determinan yang memengaruhi kejadian *stunting* di tiap daerah, baik dari segi sosial, ekonomi, pendidikan, maupun kondisi lingkungan **zemariam2025prediction**. Faktor lainnya seperti status gizi ibu, akses terhadap layanan kesehatan, dan sanitasi dasar memiliki kontribusi yang berbeda-beda pada setiap wilayah. Adanya variasi faktor risiko ini mengindikasikan bahwa pendekatan penanganan yang bersifat generalis memiliki keterbatasan dalam menjangkau akar masalah di tiap daerah, sehingga dibutuhkan landasan data yang spesifik per wilayah untuk mendukung perumusan strategi intervensi yang lebih presisi dan tepat sasaran. Maka dari itu, diperlukan analisis yang lebih mendalam untuk memahami faktor-faktor yang paling berpengaruh untuk situasi *stunting* tiap provinsi.

Mengingat kompleksitas faktor determinan yang memengaruhi kejadian *stunting*, metode analisis statistik konvensional sering kali memiliki keterbatasan dalam menangkap pola hubungan antarvariabel yang rumit. Oleh

karena itu, pendekatan komputasi modern berbasis *Machine Learning* kini menjadi alternatif solusi yang krusial untuk memetakan faktor risiko dengan akurasi yang lebih tinggi. Pemanfaatan algoritma cerdas ini memungkinkan pengolahan data dalam skala besar sekaligus mengungkap interaksi non-linear yang sulit dideteksi oleh metode tradisional. Efektivitas pendekatan tersebut telah terbukti secara empiris dalam studi yang berhasil mengidentifikasi determinan utama malnutrisi di negara berkembang dengan performa prediksi yang sangat kompetitif **tamanna2025identifying**.

Penelitian sebelumnya menerapkan algoritma *machine learning* untuk memprediksi *stunting* pada kalangan remaja di Ethiopia dan menemukan bahwa metode konvensional kurang mampu menangkap interaksi kompleks antarvariabel **zemariam2025prediction**. Namun penelitian tersebut dilakukan secara terbatas pada kelompok usia remaja dan wilayah tertentu, sehingga belum secara langsung dapat digeneralisasi dan masih memerlukan kajian tersendiri untuk ke skala nasional Indonesia yang mencakup berbagai karakteristik. Dengan demikian, penting untuk mengeksplorasi pendekatan serupa pada skala provinsi di Indonesia untuk memahami faktor determinan *stunting* berdasarkan kondisi lokal.

Dalam pemodelan prediktif pada bidang kesehatan dan gizi, data yang digunakan umumnya bersifat kompleks karena melibatkan banyak faktor yang saling berinteraksi, sehingga diperlukan algoritma yang mampu mengidentifikasi pola hubungan antarvariabel secara efektif, termasuk hubungan yang bersifat nonlinier **tamanna2025identifying**. Salah satu algoritma yang populer adalah *Random Forest*, yang dikenal memiliki kemampuan tinggi dalam menangani variabel dalam jumlah besar serta menghasilkan prediksi yang akurat. Keunggulan lain dari Random Forest adalah kemampuannya dalam mengukur tingkat kontribusi masing-masing variabel *(feature importance)*, sehingga dapat membantu memahami faktor-faktor mana yang paling berpengaruh terhadap suatu fenomena. Namun

demikian, model ini sering dianggap sebagai *black box* karena sulit dijelaskan secara langsung oleh pengambil kebijakan atau peneliti non-teknis.

Untuk mengatasi keterbatasan interpretasi tersebut, digunakan metode interpretabilitas seperti SHAP (*SHapley Additive exPlanations*) yang mampu menjelaskan kontribusi masing-masing fitur terhadap hasil prediksi model **tamanna2025identifying**. Melalui pendekatan ini, setiap faktor dapat dinilai apakah ia meningkatkan atau menurunkan kemungkinan terjadinya stunting pada suatu wilayah. Dengan demikian, kombinasi antara *Random Forest* dan SHAP tidak hanya memberikan hasil prediksi yang akurat, tetapi juga penjelasan yang dapat dipahami secara intuitif oleh pengambil keputusan. Pendekatan ini memberikan nilai tambah dalam analisis kebijakan berbasis data, terutama untuk menentukan prioritas intervensi di daerah dengan tingkat *stunting* tinggi.

Menjawab permasalahan tersebut, penelitian ini diarahkan untuk mengidentifikasi faktor-faktor determinan utama yang memengaruhi prevalensi *stunting* antar provinsi di Indonesia menggunakan data SSGI tahun 2024 melalui pendekatan *Random Forest* yang diinterpretasikan dengan SHAP. Penelitian ini berfokus pada dua aspek utama, yaitu menilai performa model dalam memprediksi tingkat *stunting* serta memahami peran masing-masing variabel terhadap hasil prediksi. Hasil penelitian diharapkan dapat memberikan dasar empiris bagi pemerintah dan pemangku kebijakan dalam merancang strategi intervensi gizi yang lebih efektif dan berbasis bukti ilmiah.

1.2 Rumusan Masalah

Berdasarkan pemaparan latar belakang, dihasilkan rumusan masalah pada penelitian yang akan dijabarkan pada poin berikut:

1. Apa saja faktor-faktor yang menjadi determinan utama penyebab *stunting* antar provinsi di Indonesia berdasarkan data SSGI tahun 2024?
2. Bagaimana performa model *Random Forest* dalam memprediksi

prevalensi *stunting* antar provinsi di Indonesia?

3. Bagaimana interpretasi SHAP digunakan untuk menjelaskan kontribusi masing-masing faktor terhadap hasil prediksi model *Random Forest*?

1.3 Tujuan Penelitian

Setelah penjabaran rumusan masalah, pada penelitian ini dibuat tujuan penelitian yaitu sebagai berikut:

1. Mengidentifikasi faktor-faktor determinan utama penyebab *stunting* antar provinsi di Indonesia berdasarkan data SSGI tahun 2024.
2. Menganalisis performa model *Random Forest* dalam memprediksi prevalensi *stunting* antar provinsi di Indonesia.
3. Menerapkan interpretasi SHAP untuk menjelaskan kontribusi masing-masing faktor terhadap hasil prediksi model *Random Forest*.

1.4 Batasan Masalah

Agar penelitian tidak terlalu luas dan tidak keluar dari pokok permasalahan, maka ditentukan batasan oleh beberapa poin berikut:

1. Data yang digunakan untuk penelitian ini terbatas pada Survei Status Gizi Indonesia (SSGI) tahun 2024 yang telah dipublikasikan oleh Kementerian Kesehatan RI. Perlu diketahui bahwa penelitian ini tidak menggunakan data tambahan dari sumber lain. Sehingga penelitian yang dilakukan bukan analisis *time-series* antartahun dan memiliki ruang lingkup spasial terbatas pada 34 provinsi di Indonesia sesuai ketersediaan data.
2. Populasi yang diambil untuk menjadi fokus penelitian ini adalah anak balita usia 0-59 bulan (5 tahun), sesuai dengan sasaran pengukuran prevalensi *stunting* pada SSGI.
3. Variabel yang digunakan untuk penelitian ini mencakup berbagai faktor yang memiliki potensi untuk memengaruhi prevalensi penyebab stunting,

seperti karakteristik balita, karakteristik ibu, serta kondisi rumah tangga yang meliputi status gizi, pendidikan, ekonomi, dan akses terhadap fasilitas kesehatan, air bersih, serta sanitasi. variabel-variabel ini merupakan sebagian dari indikator yang tersedia dalam dataset SSGI tahun 2024 dan dipilih berdasarkan relevansi terhadap topik penelitian.

4. Analisis yang dilakukan pada penelitian ini difokuskan pada tingkat provinsi di Indonesia, sehingga hasilnya mencerminkan kondisi agregat dan tidak mempertimbangkan variasi di tingkat kabupaten/kota atau individu.
5. Model yang digunakan untuk penelitian ini adalah algoritma *Random Forest* untuk memprediksi prevalensi *stunting*, dengan interpretasi hasil menggunakan metode SHAP (*SHapley Additive exPlanations*) untuk memahami kontribusi setiap variabel terhadap prediksi prevalensi *stunting*.
6. Penelitian ini berfokus pada analisis faktor determinan penyebab *stunting*, bukan pada pengembangan sistem prediksi *stunting* secara *real-time* atau aplikasi praktis lainnya.

1.5 Manfaat Penelitian

Penelitian ini dilaksanakan dengan harapan agar nantinya dapat memiliki manfaat sebagai berikut:

1. Bagi masyarakat, penelitian ini diharapkan dapat memberi gambaran empiris mengenai faktor-faktor utama yang memiliki kontribusi terhadap *stunting* di berbagai provinsi di Indonesia terutama di Sumatera. Hasil penelitian diharapkan dapat menjadi dasar bagi tenaga kesehatan dan pemerintah dalam menyusun strategi intervensi gizi yang lebih terarah dan sesuai dengan karakteristik wilayah masing-masing.
2. Bagi peneliti, penelitian ini bermanfaat untuk mengembangkan pemahaman mengenai penerapan algoritma *machine learning*,

khususnya *Random Forest* dengan interpretasi SHAP dalan menganalisis data kesehatan masyarakat.

3. Bagi akademisi, dapat dijadikan referensi mahasiswa lain yang ingin meneliti terkait faktor-faktor penyebab *stunting* maupun penerapan model *machine learning* untuk analisis data survei kesehatan.

1.6 Sistematika Penulisan

Secara keseluruhan, struktur penulisan laporan ini disusun untuk memberikan gambaran umum setiap tahapan penelitian mengenai analisis determinan penyebab *stunting* di provinsi-provinsi di Indonesia. Setiap bab disusun secara sistematis agar pembahasan mengenai penerapan model *Random Forest* dan interpretasi SHAP pada data SSGI tahun 2024 dapat diikuti dengan runtut dan mudah dipahami.

Bab I

Bab I berisi pendahuluan. Pendahuluan menguraikan latar belakang, rumusan masalah, tujuan, manfaat, batasan masalah dan sistematika penulisan dari penelitian.

Bab II

Bab II berisi tinjauan pustaka. Pada bab ini akan dibahas teori-teori mengenai *stunting*, faktor-faktor determinan penyebab *stunting*, Survei Status Gizi Indonesia (SSGI), serta penerapan algoritma *machine learning* khususnya *Random Forest* pada model prediksi risiko *stunting*. Selain itu, bab ini juga membahas hasil penelitian terdahulu yang relevan sebagai landasan konseptual dan pembeda penelitian ini dari studi yang telah dilakukan sebelumnya.

Bab III

Bab ini menjelaskan metode yang akan digunakan dalam penelitian, termask sumber dan jenis data, variabel penelitian, tahapan pengolahan data,

serta rancangan model analisis menggunakan *Random Forest* dan SHAP. Selain itu, bab ini juga menjelaskan mengenai teknik evaluasi model serta prosedur interpretasi hasil untuk mengidentifikasi faktor determinan utama penyebab *stunting*.

Bab IV

Bab ini menguraikan hasil analisis data, performa model *Random Forest*, hasil interpretasi SHAP, serta pembahasan mengenai faktor-faktor determinan yang memiliki pengaruh prevalensi *stunting* antar provinsi di Indonesia.

Bab V

Bab V berisi Kesimpulan dan saran. Pada bab ini akan diberikan kesimpulan dari hasil penelitian yang menjawab rumusan masalah berdasarkan hasil analisis yang telah dilakukan sebelumnya. Saran-saran juga diberikan sebagai masukan bagi pihak terkait dan peneliti selanjutnya untuk pengembangan kebijakan dan penelitian lanjutan di bidang *stunting*.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

2.2 Tinjauan Pustaka

Bagian ini memaparkan tinjauan terhadap delapan jurnal utama yang menjadi acuan dan rujukan dalam pelaksanaan penelitian ini. Jurnal-jurnal tersebut memberikan landasan mengenai pemilihan algoritma, penggunaan variabel determinan, serta metode evaluasi model yang diterapkan dalam tugas akhir ini.

1. Pande *et al.* pada tahun 2023 meneliti faktor-faktor penentu pertumbuhan linear anak di India guna memahami penyebab masalah gizi kronis. Studi ini menggunakan pendekatan *multilevel mixed-effect logistic regression* untuk menguji pengaruh variabel dari tingkat individu hingga lingkungan komunitas dengan memanfaatkan data NFHS-5 yang mencakup lebih dari 146 ribu balita. Hasil analisis menunjukkan bahwa prevalensi *stunting* mencapai angka 36%, di mana faktor individu seperti jenis kelamin laki-laki dan berat lahir rendah terbukti meningkatkan risiko secara signifikan. Lingkungan dengan tingkat kemiskinan tinggi meningkatkan kemungkinan *stunting* hingga 68% pada tingkat komunitas, sedangkan komunitas dengan literasi tinggi justru mampu menurunkan risiko tersebut secara efektif. Penelitian ini menjadi referensi yang sangat relevan karena memberikan landasan empiris bahwa variabel sosial-ekonomi dan lingkungan memiliki kontribusi besar terhadap kejadian *stunting*, selaras dengan variabel yang akan dianalisis dalam penelitian ini **pande_indian_stunting_2023**.
2. Zemariam *et al.* pada tahun 2025 mengkaji kemampuan berbagai algoritma *machine learning* untuk memprediksi status *stunting* serta

determinan sosial-ekonominya pada remaja perempuan di Ethiopia. Penelitian ini membandingkan kinerja delapan algoritma berbeda, termasuk *Random Forest*, dengan menerapkan teknik penyeimbangan data (*SMOTE*) serta seleksi fitur algoritma Boruta untuk menyaring variabel yang paling informatif. Hasil evaluasi menunjukkan bahwa *Random Forest* menjadi model dengan performa terbaik dibandingkan algoritma lainnya, mencatatkan akurasi sebesar 77% dan nilai AUC mencapai 85%. Model ini juga berhasil mengidentifikasi bahwa faktor wilayah, indeks kekayaan rendah, serta minimnya pendidikan formal merupakan prediktor utama kejadian *stunting*. Temuan ini sangat mendukung penelitian ini karena membuktikan efektivitas *Random Forest* dalam menangani data survei kesehatan yang kompleks, serta kemampuannya dalam menyoroti kontribusi variabel sosial-ekonomi terhadap isu gizi secara akurat **zemariam2025prediction**.

3. Jemil *et al.* pada tahun 2026 bertujuan memprediksi kejadian *severe stunting* serta determinan utamanya pada balita di 12 negara Afrika Timur. Studi ini membandingkan kinerja delapan algoritma klasifikasi—termasuk *Random Forest*—dengan menerapkan teknik *SMOTE* untuk penyeimbangan data serta *Stratified 10-Fold Cross-Validation* sebagai metode evaluasi menggunakan data DHS dengan sampel lebih dari 76 ribu anak. Hasil eksperimen menempatkan *Random Forest* sebagai model terbaik dengan akurasi mencapai 87% dan nilai AUC sebesar 0,83. Studi ini berhasil mengungkap bahwa faktor seperti kurangnya ASI eksklusif, status ekonomi rendah, serta sanitasi buruk merupakan pemicu utama risiko *stunting* parah melalui analisis interpretabilitas (SHAP). Referensi ini sangat krusial bagi penelitian ini karena memberikan validasi empiris mengenai keandalan kombinasi *Random Forest* dan metode SHAP dalam membedah faktor risiko malnutrisi secara mendetail **jemil2026predicting**.

4. Tamanna *et al.* pada tahun 2025 memanfaatkan data *Bangladesh Demographic and Health Survey* (BDHS) untuk mengevaluasi determinan berbagai bentuk malnutrisi, termasuk *stunting*, *wasting*, dan *underweight*. Studi ini menerapkan algoritma Boruta untuk seleksi fitur sebelum menguji performa berbagai model *machine learning*, seperti *Random Forest*, XGBoost, dan SVM, dalam memprediksi status gizi balita. Hasil evaluasi menunjukkan bahwa *Random Forest* mampu memodelkan kondisi malnutrisi secara kompetitif, dengan akurasi prediksi *stunting* mencapai 64,19% dan *wasting* sebesar 76,68%. Penelitian ini juga mengidentifikasi bahwa pendidikan ibu, indeks kekayaan, serta fasilitas sanitasi merupakan prediktor utama status gizi anak melalui analisis nilai SHAP. Temuan ini memperkuat landasan penelitian ini karena membuktikan bahwa *Random Forest* efektif digunakan untuk memodelkan masalah gizi yang kompleks, sekaligus menegaskan peran krusial variabel sosial-ekonomi yang akan menjadi fokus analisis pada tingkat provinsi di Indonesia **tamanna2025identifying**.
5. Pratama *et al.* pada tahun 2024 membandingkan efektivitas berbagai algoritma *machine learning* untuk memprediksi prevalensi *stunting* tingkat provinsi di Indonesia dalam konteks nasional. Studi ini menguji performa model seperti SVR, GAM, dan *Random Forest* menggunakan metrik evaluasi regresi dengan memanfaatkan dataset gabungan dari Survei Status Gizi Indonesia (SSGI) tahun 2021–2022 dan data publikasi resmi lainnya. Hasilnya menunjukkan bahwa *Random Forest Regression* memberikan kinerja paling unggul dengan nilai R^2 sebesar 0,703 dan tingkat kesalahan (MAPE) di bawah 10%. Korelasi kuat antara akses sanitasi layak dan kemiskinan terhadap angka prevalensi juga terkonfirmasi dalam penelitian tersebut. Temuan ini menjadi landasan empiris yang sangat krusial bagi penelitian ini,

karena membuktikan bahwa *Random Forest* merupakan metode yang tangguh untuk menangani karakteristik data agregat SSGI, sehingga sangat relevan untuk diterapkan kembali pada data terbaru tahun 2024 **pratama2024comparison**.

6. Kutlu, Donmez, dan Freeman pada tahun 2024 berfokus pada penerapan *machine learning* untuk menilai risiko diabetes sekaligus mengatasi keterbatasan model *black-box* dalam konteks klinis. Studi ini mengombinasikan seleksi fitur *Recursive Feature Elimination* (RFE) dan algoritma XGBoost yang menghasilkan akurasi prediksi sebesar 86,6% dengan memanfaatkan dataset masif berisi lebih dari 250 ribu pasien. Analisis menggunakan metode SHAP berhasil mengungkap bahwa status kesehatan umum, tekanan darah, serta BMI merupakan variabel paling berpengaruh dalam pengambilan keputusan model. Integrasi SHAP mampu menjembatani celah antara akurasi algoritma dan kebutuhan interpretasi medis, sehingga studi ini menjadi rujukan metodologis yang kuat dan selaras dengan tujuan penelitian ini dalam menjelaskan determinan *stunting* **kutlu2024machine**.
7. Orji dan Ukwandu pada tahun 2024 menawarkan perspektif penting mengenai standar evaluasi model regresi dalam prediksi biaya kesehatan yang transparan. Studi ini membandingkan algoritma seperti *Random Forest* dan XGBoost, dengan fokus khusus pada penggunaan metrik evaluasi yang komprehensif, mulai dari R^2 hingga *Mean Absolute Percentage Error* (MAPE). Hasil pengujian menunjukkan bahwa *Random Forest* memiliki performa yang sangat kompetitif dengan nilai R^2 mencapai 0,82 dan MAPE sebesar 12,7%, sembari tetap mempertahankan kemampuan interpretasi fitur yang krusial bagi pembuat kebijakan. Kerangka validasi model regresinya yang ketat menjadi relevansi utama jurnal ini bagi penelitian. Hal tersebut menjadi acuan metodologis dalam memilih metrik pengukuran *error* yang tepat

untuk memodelkan prevalensi *stunting* **orji2024machine**.

8. Irisson *et al.* pada tahun 2022 menyoroti tantangan validasi pada data spasial yang sering kali memiliki autokorelasi, di mana metode pengujian konvensional cenderung memberikan estimasi *error* yang terlalu optimis. Studi ini memperkenalkan pendekatan *Iterative Spatial Leave-One-Out Cross-Validation* (SLOOCV) yang menjamin independensi antara data latih dan data uji untuk mengatasi hal tersebut. Hasil eksperimen menunjukkan bahwa metode ini mampu menghasilkan estimasi kinerja yang jauh lebih realistis serta mengurangi bias prediksi secara signifikan dibandingkan teknik pembagian acak biasa. Penggunaan skema *Leave-One-Out Cross-Validation* (LOOCV) dalam penelitian ini didasari oleh prinsip kuat tersebut. Data agregat provinsi memiliki karakteristik spasial terbatas yang membutuhkan strategi evaluasi model yang ketat **irisson2022iterative**.

Ringkasan sistematis dari seluruh jurnal rujukan ditampilkan pada Tabel **??** untuk mempermudah pemetaan posisi penelitian terhadap studi-studi terdahulu. Tabel ini merangkum aspek-aspek kunci yang meliputi judul, metode, serta temuan utama guna memperjelas perbedaan pendekatan dan kontribusi kebaruan dalam penelitian ini.

Tabel 2.1 Ringkasan Penelitian Terdahulu

No.	Judul & Penulis	Masalah	Metode	Hasil Utama
1.	*Analyzing determinants from both compositional and contextual level impeding desired linear growth of children in Indian context* **pande_indian_stunting_2023**	Identifikasi determinan *stunting* anak (individu & kontekstual) di India.	*Multilevel Mixed-Effect Logistic Regression*.	Prevalensi *stunting* 36%; faktor risiko utama meliputi jenis kelamin laki-laki, BBLR, dan kemiskinan komunitas; literasi tinggi menjadi faktor protektif.
2.	*Prediction of stunting and its socioeconomic determinants among adolescent girls in Ethiopia using machine learning algorithms* **zemariam2025prediction**	Prediksi *stunting* pada remaja perempuan di Ethiopia beserta faktor sosial-ekonominya.	*Random Forest* (RF), SMOTE, Boruta.	RF menjadi model terbaik (Akurasi 77%, AUC 85%). Faktor dominan: wilayah, indeks kekayaan rendah, dan kurangnya pendidikan.

No.	Judul & Penulis	Masalah	Metode	Hasil Utama
3.	*Predicting severe stunting and its determinants among under-five in Eastern African Countries: A machine learning algorithms* **jemil2026predicting**	Prediksi kejadian *severe stunting* balita di 12 negara Afrika Timur.	RF, SMOTE, *Stratified 10-Fold CV*, SHAP.	RF unggul dengan Akurasi 87% dan AUC 0.83. SHAP mengungkap determinan kunci: kurang ASI eksklusif, ekonomi lemah, dan sanitasi buruk.
4.	*Identifying determinants of malnutrition in under-five children in Bangladesh: insights from the BDHS-2022 cross-sectional study* **tamanna2025identifying**	Evaluasi determinan malnutrisi (*stunting, wasting*) pada balita di Bangladesh.	Boruta *Feature Selection*, RF, XGBoost.	RF kompetitif dengan akurasi *stunting* 64,19% dan *wasting* 76,68%. Determinan utama: pendidikan ibu, kekayaan, dan sanitasi.

No.	Judul & Penulis	Masalah	Metode	Hasil Utama
5.	*Comparison of Machine Learning Algorithms for Predicting Stunting Prevalence in Indonesia* **pratama2024comparison**	Prediksi prevalensi *stunting* tingkat provinsi di Indonesia (Data SSGI).	*Random Forest Regression*, SVR, GAM.	*RF Regression* terbaik ($R^2 = 0,703$, MAPE <10%). Menunjukkan korelasi kuat antara sanitasi layak dan kemiskinan terhadap prevalensi.
6.	*Machine learning interpretability in diabetes risk assessment: a SHAP analysis* **kutlu2024machine**	Mengatasi sifat *black-box* model prediksi risiko kesehatan (Diabetes).	XGBoost, RFE, SHAP.	Akurasi 86,6%. SHAP berhasil memberikan interpretasi transparan terhadap fitur dominan (kesehatan umum, TD, BMI) dalam model medis.

No.	Judul & Penulis	Masalah	Metode	Hasil Utama
7.	*Machine learning for an explainable cost prediction of medical insurance* **orji2024machine**	Prediksi biaya asuransi kesehatan yang transparan (*explainable*).	RF, XGBoost, SHAP, Evaluasi Regresi (R^2, MAPE).	RF mencapai R^2 0,82 dan MAPE 12,7%. Menetapkan standar evaluasi regresi yang komprehensif serta pentingnya interpretabilitas fitur.
8.	*Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing* **irisson2022iterative**	Bias validasi pada data spasial yang memiliki autokorelasi antar wilayah.	*Iterative Spatial Leave-One-Out Cross-Validation* (SLOOCV).	Metode validasi LOOCV spasial menghasilkan estimasi *error* yang lebih realistis dan tidak bias dibandingkan pembagian acak biasa pada data berbasis wilayah.

2.3 Dasar Teori

Berisi teori/konsep yang berkaitan/digunakan dalam tugas akhir yang dikerjakan. Gunakanlah data melalui buku/jurnal referensi, publikasi tugas

akhir, penelitian, buku, dan informasi web yang dapat dipertanggungjawabkan, hindari penggunaan dasar teori melalui tautan Wikipedia, surat kabar, atau portal berita, yang dapat memiliki isi yang tidak bersifat fakta.

2.3.1 Teori 1

Berikut adalah contoh penyisipan tabel menggunakan `\begin{longtable}{}`:

Tabel 2.2 Contoh Tabel

Col1	Col2	Col2	Col3
1	6	87837	787
2	7	78	5415
3	545	778	7507
4	545	18744	7560
5	88	788	6344

2.3.2 Dasar Teori

2.3.2.1 Stunting

Stunting merupakan kondisi kegagalan pertumbuhan pada anak akibat kekurangan gizi kronis serta infeksi berulang dalam jangka waktu lama (SITASI). Secara teknis, kondisi ini diukur menggunakan indeks Tinggi Badan menurut Umur (TB/U) dengan ambang batas nilai *Z-score* kurang dari -2 Standar Deviasi (SD) (SITASI). Penentuan status *stunting* melalui perhitungan *Z-score* secara matematis dinyatakan sebagai berikut:

$$Z = \frac{x - \text{median}}{\text{SD}} \quad \text{(Rumus 2.1)}$$

Kondisi ini memiliki dampak jangka pendek berupa penurunan daya tahan tubuh, serta dampak jangka panjang yang menghambat perkembangan kognitif dan kapasitas intelektual anak (SITASI). Hal tersebut menjadikannya sebagai masalah multidimensi yang dipengaruhi oleh berbagai faktor risiko, mulai dari

pola asuh dan nutrisi hingga faktor lingkungan. Karakteristik data stunting yang dipengaruhi oleh interaksi variabel yang kompleks membuat pendekatan komputasi diperlukan untuk mengidentifikasi pola determinan secara lebih akurat. Melalui pemodelan yang tepat, hubungan antar-variabel ini dapat dipetakan guna mendukung strategi pencegahan yang lebih efektif di tingkat wilayah.

2.3.2.2 Determinan Stunting

Determinan *stunting* dalam skala populasi diukur melalui berbagai indikator agregat yang mencerminkan kualitas kesehatan, sosial ekonomi, dan lingkungan di suatu wilayah. Penentuan variabel determinan dalam penelitian ini merujuk pada kerangka konseptual yang ditetapkan oleh *World Health Organization* (WHO), yang membagi faktor risiko menjadi tingkat konteks sosial, politik, serta faktor rumah tangga dan lingkungan (SITASI). Visualisasi kerangka faktor risiko tersebut disajikan pada Gambar **??** untuk memberikan gambaran menyeluruh mengenai interaksi antarvariabel.

Gambar 2.1 Kerangka Konseptual Faktor Penyebab *Stunting* **SITASI_SUMBER_GAMBAR**

Indikator kesehatan ibu dan anak menjadi variabel utama yang diukur melalui persentase cakupan layanan di suatu provinsi. Hal ini mencakup variabel seperti persentase ibu hamil risiko tinggi, persentase balita dengan ASI eksklusif, serta cakupan imunisasi dasar lengkap (SITASI). Merujuk pada Gambar **??**, indikator-indikator ini merepresentasikan kualitas pengasuhan dan asupan yang merupakan faktor penyebab langsung di tingkat wilayah. Tinggi rendahnya persentase capaian ini mencerminkan sejauh mana perlindungan kesehatan anak telah terimplementasi secara merata.

Variabel sosial ekonomi dan lingkungan yang direpresentasikan melalui indikator makro juga memberikan kontribusi besar terhadap fluktuasi prevalensi. Indikator lingkungan diukur melalui persentase rumah tangga

dengan akses air minum aman dan sanitasi layak, sementara indikator sosial ekonomi mencakup tingkat kemiskinan provinsi (SITASI). Faktor-faktor tersebut bertindak sebagai determinan dasar yang memengaruhi kemampuan masyarakat dalam mengakses sumber daya gizi secara kolektif. Peningkatan kualitas infrastruktur dasar di tingkat provinsi dipercaya memiliki korelasi linear terhadap penurunan angka *stunting* secara signifikan.

Pemetaan seluruh indikator determinan berbasis persentase ini berfungsi sebagai landasan dalam tahap pemilihan fitur (*feature selection*) pada pemodelan regresi. Melalui pendekatan ini, model *Random Forest* dapat mempelajari pola hubungan antara proporsi variabel determinan terhadap nilai kontinu prevalensi *stunting* di tingkat provinsi. Analisis berbasis indikator wilayah ini sangat relevan untuk mengidentifikasi area yang membutuhkan intervensi kebijakan lebih mendalam.

2.3.2.3 Survei Status Gizi Indonesia (SSGI)

Survei Status Gizi Indonesia (SSGI) merupakan instrumen nasional utama yang digunakan oleh pemerintah untuk memantau status gizi balita secara berkala di seluruh wilayah Indonesia (SITASI). Survei ini bertujuan untuk menyediakan data prevalensi *stunting*, *wasting*, *underweight*, dan *overweight* yang akurat dan representatif dari tingkat nasional hingga kabupaten/kota (SITASI). Dalam penelitian ini, laporan resmi SSGI digunakan sebagai sumber data primer untuk mengekstraksi variabel target dan variabel prediktor dalam skala provinsi. Penggunaan data sekunder publikasi resmi ini menjamin bahwa basis analisis yang digunakan telah melalui proses validasi teknis oleh lembaga berwenang sebelum dipublikasikan secara nasional.

Indikator yang dikumpulkan dalam SSGI mencakup intervensi gizi spesifik dan intervensi gizi sensitif yang disajikan dalam bentuk proporsi atau persentase populasi (SITASI). Indikator spesifik meliputi cakupan layanan kesehatan langsung seperti pemberian ASI eksklusif, sedangkan indikator

sensitif mencakup faktor pendukung seperti persentase rumah tangga dengan akses sanitasi layak (SITASI). Metodologi perolehan data prevalensi dilakukan melalui pengukuran antropometri yang terstandarisasi guna menghasilkan estimasi angka yang konsisten antarwilayah. Proses standardisasi ini sangat penting untuk meminimalisir bias pengukuran yang mungkin terjadi saat pengambilan data di lapangan.

Data agregat yang dihasilkan SSGI memiliki karakteristik nilai kontinu dalam rentang 0–100 persen (SITASI). Struktur ini merepresentasikan kondisi kesehatan dan sosial ekonomi pada 38 provinsi di Indonesia secara makro. Format persentase yang konsisten pada seluruh variabel memudahkan perbandingan antarwilayah serta identifikasi variasi determinan terhadap prevalensi *stunting*. Pendekatan berbasis data agregat ini memberikan gambaran populasi yang lebih luas dibandingkan data individu dalam mendukung analisis kebijakan di tingkat wilayah.

2.3.2.4 *Machine Learning*

Machine Learning merupakan cabang dari kecerdasan buatan yang memungkinkan sistem untuk mempelajari pola secara mandiri dari kumpulan data tanpa melalui pemrograman instruksi secara eksplisit (SITASI). Teknologi ini bekerja dengan membangun model statistik yang mampu mengenali tren tersembunyi guna melakukan generalisasi pada data baru. Di bidang kesehatan, pendekatan ini sangat berguna untuk menganalisis hubungan antar variabel yang kompleks dan sering kali bersifat non-linear. Penggunaan algoritma ini berpotensi menghasilkan estimasi yang lebih adaptif terhadap kompleksitas data dibandingkan dengan pendekatan statistik konvensional (SITASI).

Prinsip kerja yang paling relevan dalam pemetaan faktor kesehatan adalah *supervised learning*, di mana model dilatih menggunakan pasangan data masukan dan target yang sudah diketahui nilainya (SITASI). Melalui proses pelatihan ini, algoritma akan mencoba menemukan fungsi pemetaan paling

optimal untuk meminimalkan kesalahan prediksi. Pendekatan ini secara khusus dapat diterapkan dalam bentuk regresi jika variabel yang diprediksi berupa nilai kontinu, seperti angka prevalensi *stunting* di tingkat provinsi. Pemodelan ini menjadi dasar teknis sebelum melakukan implementasi algoritma yang lebih spesifik guna mendapatkan performa prediksi yang stabil dan akurat.

2.3.2.5 Supervised Learning

Supervised learning adalah kategori dalam pembelajaran mesin yang beroperasi menggunakan pasangan data masukan dan target yang telah memiliki label atau nilai kebenaran sebelumnya (SITASI). Dalam mekanismenya, algoritma akan mempelajari fungsi pemetaan antara fitur masukan (*input*) dan variabel target (*output*) untuk meminimalkan selisih kesalahan prediksi. Peran fitur dalam proses ini sangat krusial, karena kualitas informasi yang dikandung oleh fitur akan menentukan kemampuan model dalam mengenali karakteristik unik dari data (SITASI). Tujuan akhir dari model ini adalah untuk memprediksi label atau nilai pada data baru dengan tingkat akurasi yang optimal berdasarkan pengalaman pelatihan. Berdasarkan tipe variabel targetnya, *supervised learning* dapat diklasifikasikan lebih spesifik menjadi tugas klasifikasi atau tugas regresi.

2.3.2.6 Regresi dalam Machine Learning

Pendekatan regresi dalam *machine learning* digunakan secara khusus untuk menangani permasalahan di mana variabel target yang diprediksi merupakan nilai kontinu atau numerik (SITASI). Berbeda dengan klasifikasi yang mengelompokkan data ke dalam kategori diskrit, regresi berfokus pada estimasi hubungan kuantitatif antar variabel untuk menghasilkan keluaran dalam rentang angka tertentu. Ilustrasi dasar dari hubungan ini dapat digambarkan melalui persamaan regresi linear sederhana sebagai berikut:

$$y = \beta_0 + \beta_1 x + \epsilon$$

(Rumus 2.2)

Di mana y adalah variabel target, x adalah fitur masukan, β adalah koefisien model, dan ϵ merupakan *error* (SITASI). Penggunaan metode regresi ini memungkinkan peneliti untuk memodelkan fenomena numerik seperti angka prevalensi dengan tingkat presisi yang terukur. Sebelum model melakukan estimasi terhadap target tersebut, diperlukan tahapan pemilihan variabel yang paling relevan melalui proses *feature selection*.

2.3.2.7 Feature Selection

Feature selection adalah proses identifikasi dan pemilihan subset fitur yang paling relevan untuk digunakan dalam pembangunan model prediktif (SITASI). Tujuan utama dari tahapan ini adalah untuk mengurangi dimensi data, meningkatkan efisiensi komputasi, serta menghindari masalah *overfitting* yang dapat menurunkan performa generalisasi model. Secara umum, metode ini terbagi menjadi tiga kategori utama, yaitu *filter methods*, *wrapper methods*, dan *embedded methods* (SITASI). Dengan mengeliminasi fitur yang redundan atau tidak informatif, model dapat lebih fokus pada variabel yang memiliki korelasi kuat terhadap variabel target. Setelah fitur yang relevan terpilih, langkah selanjutnya adalah memastikan data tersebut berada dalam format yang seragam melalui transformasi data.

2.3.2.8 Transformasi Data

Transformasi data merupakan prosedur modifikasi nilai data asli ke dalam format atau skala tertentu guna meningkatkan performa algoritma pembelajaran mesin (SITASI). Proses ini mencakup normalisasi skala agar seluruh fitur memiliki rentang nilai yang seragam, penggabungan kategori yang serupa, hingga konversi indikator tertentu agar lebih representatif. Dalam analisis data kesehatan, transformasi sering kali diperlukan untuk memastikan bahwa perbedaan satuan antar variabel tidak memengaruhi bobot keputusan model secara bias (SITASI). Keberhasilan transformasi akan menentukan kualitas

representasi informasi yang kemudian dapat diproses lebih lanjut ke tahap agregasi. Setelah data ditransformasi, informasi tersebut sering kali perlu dikelompokkan kembali melalui mekanisme agregasi data.

2.3.2.9 Agregasi Data

Agregasi data adalah proses pengumpulan dan perangkuman data dari tingkat individu menjadi format kelompok atau unit yang lebih besar, seperti wilayah administratif (SITASI). Dalam konteks spasial, agregasi memungkinkan peneliti untuk melihat representasi makro dari suatu populasi yang sering kali lebih stabil dan bermakna untuk kebijakan publik dibandingkan data individu yang fluktuatif. Meskipun memberikan keunggulan dalam penyederhanaan informasi, agregasi memiliki keterbatasan berupa hilangnya variasi detail di tingkat mikro (SITASI). Namun, penggunaan data agregat sangat efektif untuk memetakan determinan kesehatan masyarakat pada lingkup wilayah seperti provinsi. Data agregat yang telah tersusun rapi ini kemudian siap digunakan sebagai masukan bagi algoritma *Random Forest Regression*.

2.3.2.10 Random Forest Regression

Random Forest Regression merupakan algoritma *ensemble learning* yang bekerja dengan membangun sekumpulan pohon keputusan (*decision trees*) secara independen selama tahap pelatihan (SITASI). Algoritma ini menerapkan mekanisme *bootstrap aggregating* (*bagging*) dan seleksi fitur acak untuk menciptakan variasi antar pohon guna meningkatkan stabilitas model. Perbedaan utama *Random Forest* untuk regresi dibandingkan klasifikasi terletak pada *output* akhirnya yang merupakan nilai rata-rata dari seluruh prediksi pohon (SITASI). Struktur model ini dapat diilustrasikan sebagai berikut:

Gambar 2.2 Struktur Umum *Random Forest Regression* **SITASI_GAMBAR**

Secara matematis, hasil prediksi akhir ($\hat{y}$) dari total K pohon keputusan

dapat dirumuskan sebagai:

$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} f_k(x)$$

(Rumus 2.3)

Model ini dikenal sangat tangguh dalam menangani data non-linear dan memiliki ketahanan terhadap *outlier* di dalam dataset (SITASI). Setelah model berhasil dibangun dan melakukan prediksi, kinerjanya harus diukur secara objektif melalui evaluasi model regresi.

2.3.2.11 Evaluasi Model Regresi

Evaluasi model regresi bertujuan untuk mengukur sejauh mana prediksi yang dihasilkan oleh model mendekati nilai aktual dari data observasi (SITASI). Terdapat beberapa metrik statistik utama yang umum digunakan, antara lain *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), serta Koefisien Determinasi (R^2). Masing-masing metrik memberikan perspektif berbeda mengenai besaran kesalahan dan kualitas kecocokan model terhadap data (SITASI). Adapun rumus untuk metrik-metrik tersebut adalah sebagai berikut:

$$MAE = \frac{1}{n} \sum |y - \hat{y}|, \quad RMSE = \sqrt{\frac{1}{n} \sum (y - \hat{y})^2}$$

(Rumus 2.4)

Pemilihan metrik yang tepat sangat menentukan interpretasi peneliti terhadap akurasi model yang dikembangkan. Untuk menjamin bahwa nilai evaluasi tersebut tidak bersifat bias, diperlukan prosedur validasi yang ketat melalui *cross validation*.

2.3.2.12 Cross Validation

Cross validation merupakan teknik validasi model yang bertujuan untuk menilai kemampuan generalisasi model pada dataset yang independen (SITASI). Prosedur ini dilakukan dengan membagi data menjadi bagian latih (*training set*) dan bagian uji (*testing set*) secara berulang dengan skema tertentu.

Tujuan utamanya adalah untuk mengurangi bias evaluasi yang mungkin muncul jika model hanya diuji pada satu bagian data acak saja (SITASI). Melalui proses ini, peneliti dapat memperoleh estimasi performa model yang lebih stabil dan objektif. Salah satu varian dari teknik ini yang sangat intensif dan sistematis adalah *Leave-One-Out Cross Validation* (LOOCV).

2.3.2.13 Leave-One-Out Cross Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) merupakan bentuk ekstrem dari *K-Fold Cross Validation* di mana jumlah lipatan (*fold*) sama dengan jumlah total sampel data (N) (SITASI). Dalam setiap iterasinya, satu sampel digunakan sebagai data uji sementara $N - 1$ sampel lainnya digunakan sebagai data latih, dan proses ini diulang sebanyak N kali. Skema ini sangat cocok digunakan untuk dataset dengan jumlah sampel yang kecil karena mampu memaksimalkan penggunaan data untuk pelatihan model (SITASI). Kelebihan utamanya adalah menghasilkan estimasi yang hampir tidak bias, meskipun membutuhkan biaya komputasi yang lebih tinggi jika jumlah data meningkat.

Gambar 2.3 Skema Partisi Data pada LOOCV **SITASI_GAMBAR**

Meskipun model telah tervalidasi dengan baik, kompleksitas algoritma *ensemble* sering kali menciptakan hambatan dalam memahami alasan di balik sebuah prediksi, yang dikenal sebagai masalah *black-box*.

2.3.2.14 Interpretabilitas Model

Interpretabilitas model merujuk pada sejauh mana manusia dapat memahami alasan dan proses pengambilan keputusan di balik keluaran sebuah algoritma (SITASI). Masalah *black-box* sering kali ditemukan pada algoritma kompleks seperti *Random Forest*, di mana hubungan antar variabel sulit dijelaskan secara intuitif. Di bidang kesehatan, interpretasi menjadi aspek yang sangat krusial karena pembuat kebijakan perlu mengetahui variabel apa yang

paling memengaruhi suatu kondisi (SITASI). Penjelasan model dapat dibagi menjadi *global explanation* untuk tren populasi secara menyeluruh dan *local explanation* untuk kasus spesifik. Untuk menjembatani kebutuhan interpretasi pada model kompleks tersebut, metode SHAP menjadi salah satu solusi yang paling banyak digunakan.

2.3.2.15 SHAP (SHapley Additive exPlanations)

SHAP (*SHapley Additive exPlanations*) merupakan metode interpretasi model yang didasarkan pada teori permainan (*game theory*) untuk menjelaskan kontribusi setiap fitur terhadap prediksi tertentu (SITASI). Metode ini menghitung nilai Shapley yang merepresentasikan besaran kontribusi rata-rata suatu fitur terhadap seluruh kemungkinan kombinasi fitur lainnya. Nilai ini memberikan keadilan dalam distribusi atribusi fitur sehingga peneliti dapat memahami arah dan kekuatan pengaruh setiap variabel (SITASI). Besarnya kontribusi fitur i dalam model f dengan himpunan fitur S dirumuskan sebagai berikut:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$

(Rumus 2.5)

Keunggulan utama SHAP terletak pada konsistensi teoritisnya dan kemampuannya untuk memberikan visualisasi yang jelas baik di tingkat global maupun lokal. Melalui penerapan SHAP, hasil prediksi prevalensi *stunting* tidak hanya menjadi sekadar angka, tetapi juga memberikan wawasan mengenai faktor determinan utama di tiap wilayah.

2.3.2.16 Subsubbab

Berikut adalah contoh subsubbab. Ini adalah level subbab maksimal dalam laporan Tugas Akhir, dan tidak boleh lebih dalam.

Gambar **??** adalah contoh Gambar yang diambil dari internet yang harus dicantumkan sumbernya dan memiliki lisensi Creative Common. Jika

gambar adalah milik peneliti lain atau tidak dibuat atau diambil sendiri maka peneliti wajib meminta izin kepada peneliti lain tersebut untuk mencantumkan gambar. Gunakan `\begin{figure}` untuk memasukkan gambar. Gunakan `\caption{[nama caption]}` untuk memberikan caption gambar. Nomor caption akan diurutkan secara otomatis. Jangan lupa untuk melabeli setiap gambar dengan `\label{[nama label]}`, agar bisa direferensi menggunakan `\ref{[nama label]}`

Gambar 2.4 Contoh gambar dan caption
Sumber: Contoh

2.3.3 Teori 2

Untuk membuat sebuah rumus persamaan, gunakan kode `\begin{equationcaptioned}` seperti dibawah:

$$x + 1 = 2$$

(Rumus 2.6)

Teks caption rumus tidak akan muncul di teks, tetapi akan muncul di Daftar Rumus.

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

(Rumus 2.7)

Berikut adalah contoh penulisan persamaan yang lebih kompleks, yaitu persamaan distribusi normal.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

(Rumus 2.8)

Jika menuliskan banyak persamaan secara berurutan, gunakan `\begin{split}`:

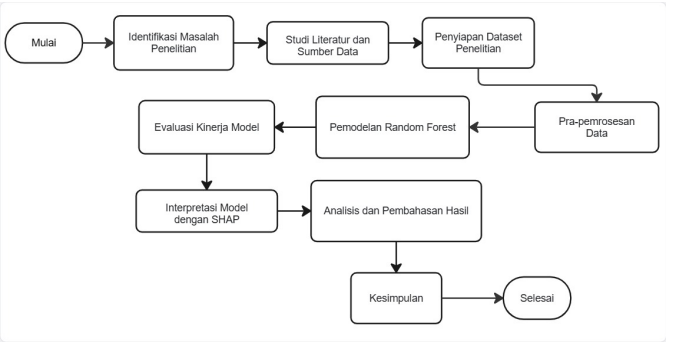
$$\begin{aligned} 2x - 5y &= 8 \\ 3x + 9y &= -12 \end{aligned}$$

(Rumus 2.9)

BAB III
METODE PENELITIAN

3.1 Alur Penelitian

Penelitian ini dirancang untuk menganalisis faktor-faktor determinan penyebab *stunting* pada tingkat provinsi di Indonesia dengan memanfaatkan data sekunder *Annual Report* SSGI tahun 2024. Pendekatan yang digunakan mengombinasikan pemodelan *machine Learning* regresi menggunakan *Random Forest* dan metode interpretasi model SHAP. Model ini dibangun menggunakan dataset agregat tingkat provinsi yang memuat berbagai indikator status gizi balita serta faktor determinan yang mencakup aspek kesehatan ibu dan anak, pola asuh, layanan kesehatan, penyakit, perlindungan sosial, dan kondisi sanitasi. Penelitian ini bertujuan untuk mengidentifikasi faktor determinan utama yang berkontribusi terhadap tingginya prevalensi *stunting* antar provinsi, serta memberikan pemahaman yang bersifat interpretatif sebagai dasar perumusan kebijakan dan intervensi gizi yang lebih tepat sasaran. Secara garis besar, tahapan penelitian digambarkan dalam diagram alir (*flowchart*) pada Gambar **??**.



Gambar 3.1 Diagram Alir Penelitian

Berdasarkan Gambar **??**, alur penelitian disusun secara sistematis yang diawali dengan tahap identifikasi masalah mengenai urgensi penanganan *stunting* di Indonesia dan kompleksitas faktor determinan yang mempengaruhinya di tingkat provinsi. Tahap ini didukung oleh studi literatur mendalam untuk memahami variabel-variabel determinan multidimensi yang mencakup berbagai aspek, serta metode pengolahan data dan pemodelan *machine learning* yang relevan. Berlandaskan pemahaman tersebut, selanjutnya dilakukan pengumpulan dataset yang bersumber dari Laporan Tahunan SSGI tahun 2024. Laporan Tahunan SSGI 2024 kemudian digunakan untuk menyiapkan dataset agregat tingkat provinsi yang memuat variabel target (prevalensi *stunting*) dan variabel fitur (faktor determinan). Data mentah yang diperoleh selanjutnya diproses melalui tahapan pra-pemrosesan data yang meliputi pembersihan data, penyaringan data, dan pemisahan variabel target dan fitur.

Setelah data siap, tahap berikutnya adalah pembangunan model regresi menggunakan algoritma *Random Forest*. Proses ini mencakup penentuan parameter terbaik melalui *Grid Search* dan validasi silang (*Cross Validation*) untuk memastikan performa model yang optimal. Model yang telah dilatih kemudian dievaluasi kinerjanya menggunakan metrik statistik seperti MAE, MSE, RMSE, MAPE, serta koefisien determinasi (R^2 Score). Langkah terakhir adalah menginterpretasikan hasil prediksi model *Random Forest* menggunakan SHAP untuk mengetahui kontribusi setiap fitur determinan terhadap angka *stunting*, baik secara global skala nasional maupun lokal per provinsi.

3.2 Langkah Penelitian

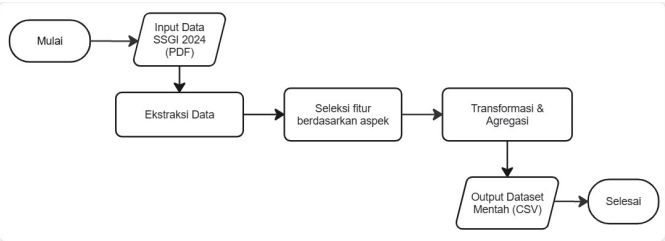
Tahapan penelitian disusun secara sistematis untuk menjawab rumusan masalah yang telah ditetapkan. Secara garis besar, alur penelitian terdiri dari identifikasi masalah, studi literatur, penyiapan data, pra-pemrosesan, pembangunan model, hingga evaluasi dan interpretasi hasil. Penjelasan rinci

mengenai masing-masing tahapan diuraikan sebagai berikut.

3.2.1 Penyiapan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari Laporan Publikasi SSGI tahun 2024 yang diterbitkan oleh Kementerian Kesehatan Republik Indonesia **kemenkes2025ssgi**. Laporan ini disajikan dalam bentuk dokumen PDF dan memuat informasi statistik yang komprehensif, mencakup ±5.189 tabel statistik yang menggambarkan indikator kesehatan ibu dan anak, pola asuh, akses pelayanan kesehatan, kondisi sanitasi dan lingkungan, serta karakteristik demografi rumah tangga. Secara spesifik, populasi target dalam survei ini difokuskan pada rumah tangga kelompok balita usia 0-59 bulan, dengan data yang disajikan menggunakan pendekatan *cross-sectional* untuk menggambarkan kondisi kesehatan pada satu titik waktu tertentu. Seluruh informasi statistik tersebut disajikan dalam bentuk numerik atau persentase yang merepresentasikan prevalensi kejadian di wilayah terkait.

Informasi dalam laporan SSGI 2024 disajikan dalam berbagai tingkat agregasi wilayah, mulai dari data nasional, provinsi, hingga rincian kabupaten/kota. Dari berbagai tingkatan tersebut, penelitian ini membatasi unit analisis pada skala provinsi guna menganalisis variabilitas determinan *stunting* antarwilayah secara makro. Namun, kompleksitas struktur data ini menyebabkan data tidak dapat langsung digunakan sebagai *dataset* penelitian kuantitatif tanpa melalui proses penyiapan data terlebih dahulu. Oleh karena itu, diperlukan suatu tahapan penyiapan data untuk mengekstraksi, menyeleksi, dan menyusun kembali informasi yang relevan agar diperoleh *dataset* terstruktur yang sesuai dengan tujuan penelitian. Alur penyiapan data yang dilakukan dalam penelitian ini ditunjukkan pada Gambar **??**.



Gambar 3.2 Alur Penyiapan Data

Alur penyiapan data pada Gambar **??** menggambarkan tahapan yang dilakukan untuk mengonversi data mentah SSGI 2024 dalam bentuk PDF menjadi *dataset* terstruktur yang siap digunakan pada tahap pra-pemrosesan dan pemodelan. Penjelasan lebih lanjut mengenai masing-masing tahapan penyiapan data diuraikan sebagai berikut:

3.2.1.1 Ekstraksi Data

Tahap pertama difokuskan pada pengumpulan data mentah dari laporan SSGI 2024 yang tersedia dalam bentuk dokumen PDF dengan ribuan tabel statistik yang kompleks. Mengingat data tersebar dalam berbagai sub-bab laporan, dilakukan identifikasi manual terlebih dahulu untuk memetakan nomor halaman tabel yang memuat indikator determinan *stunting* dan prevalensi gizi balita. Informasi dari tabel-tabel tersebut kemudian diekstraksi dan dikonversi dari bentuk dokumen statis menjadi lembar kerja digital (*spreadsheet*) yang lebih fleksibel. Karena data asli menggunakan pendekatan *cross-sectional* dengan banyak kategori, proses ekstraksi dilakukan dengan memastikan angka persentase yang diambil sesuai dengan label baris dan kolom yang benar.

Pada tahap ini juga ditetapkan skala analisis yang digunakan dalam penelitian, yaitu skala provinsi sebagai representasi kondisi nasional. Pemilihan skala provinsi dilakukan dengan mempertimbangkan tujuan penelitian yang berfokus pada pemetaan faktor determinan *stunting* secara makro. Berdasarkan keputusan tersebut, seluruh tabel yang dianalisis dibatasi pada tabel tingkat

provinsi yang tersedia dalam laporan SSGI 2024.

3.2.1.2 Seleksi Fitur Determinan

Tahap seleksi fitur dilakukan untuk menentukan variabel determinan *stunting* yang relevan dari 100 tabel tingkat provinsi hasil ekstraksi data SSGI 2024. Proses seleksi ini mempertimbangkan kesesuaian variabel dengan tujuan penelitian, keterwakilan indikator determinan, serta potensi redundansi informasi antar tabel. Variabel yang bersifat terlalu rinci, deskriptif kategori, atau tidak lagi informatif pada tingkat agregasi provinsi dieliminasi dan, apabila memungkinkan, direpresentasikan melalui variabel status atau skor risiko komposit.

Pendekatan seleksi ini diterapkan untuk menyesuaikan karakteristik data dengan tingkat analisis penelitian, yaitu pada level provinsi, sehingga kompleksitas yang tidak diperlukan dapat dikurangi tanpa menghilangkan makna substantif indikator. Berdasarkan proses tersebut, sebanyak 71 tabel dipertahankan dan digunakan sebagai dasar pembentukan fitur penelitian. Hasil seleksi ini menghasilkan *dataset* yang lebih ringkas, informatif, dan relevan dalam mengidentifikasi faktor-faktor determinan *stunting* antarprovinsi.

3.2.1.3 Transformasi dan Agregasi Data

Setelah melalui proses seleksi tabel, langkah selanjutnya adalah transformasi data untuk memastikan setiap variabel yang terbentuk memiliki representasi informasi yang kuat terhadap risiko *stunting*. Pada tahap ini, tidak semua informasi dari tabel mentah digunakan secara langsung; sebaliknya, dilakukan peninjauan ulang untuk menetapkan satu indikator yang paling representatif dari setiap tabel. Untuk tabel yang hanya memuat indikator tunggal, nilai persentase diadopsi langsung sebagai fitur numerik. Namun, pada tabel yang memiliki kategori majemuk, dilakukan strategi transformasi khusus dengan memilih indikator yang mencerminkan risiko tertinggi atau melakukan

penggabungan kategori agar menghasilkan variabel baru yang lebih ringkas dan substantif.

Selanjutnya, proses agregasi dilakukan untuk menyederhanakan struktur data sehingga seluruh variabel memiliki skala yang seragam dan mudah dibandingkan antar provinsi. Pada tabel yang memiliki lebih dari satu kategori indikator, proses agregasi dilakukan melalui pendekatan penggabungan indikator berbasis risiko. Sebagai contoh, pada tabel usia kehamilan ibu yang terdiri dari kategori usia <21 tahun, 21–39 tahun, dan >40 tahun, hanya kategori yang merepresentasikan kondisi berisiko yang dipertahankan. Nilai persentase ibu dengan usia kehamilan <20 tahun dan >35 tahun dijumlahkan untuk membentuk satu variabel baru yang merepresentasikan proporsi kehamilan berisiko pada usia ekstrem. Secara matematis, proses ini dapat dinyatakan sebagai berikut:

$$X_{\text{risiko usia kehamilan}} = X_{<20} + X_{>35}$$

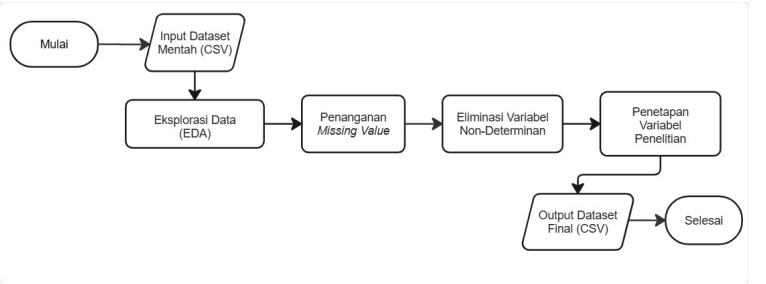
Di mana $X_{<20}$ adalah persentase kehamilan usia di bawah 20 tahun dan $X_{>35}$ adalah persentase kehamilan usia di atas 35 tahun. Pendekatan serupa diterapkan pada tabel lain yang memiliki kategori majemuk, dengan prinsip utama mempertahankan indikator yang memiliki relevansi risiko terhadap kejadian *stunting* serta menghindari redundansi informasi antarvariabel. Seluruh variabel direpresentasikan dalam bentuk persentase guna menjaga konsistensi satuan data dan memudahkan proses pemodelan. Hasil dari tahap ini adalah *dataset* terstruktur berbasis provinsi yang siap digunakan sebagai input pada tahapan analisis selanjutnya.

3.2.2 Pra-pemrosesan Data

Setelah proses penyiapan dan pembentukan *dataset* terstruktur selesai dilakukan, tahap selanjutnya adalah pra-pemrosesan data. Tahapan ini

bertujuan untuk memastikan bahwa data yang digunakan sebagai input model memiliki kualitas yang baik, konsisten, dan layak secara statistik sebelum masuk ke proses pemodelan *machine learning*. Mengingat *dataset* disusun dari hasil ekstraksi dan transformasi data sekunder, terdapat potensi munculnya permasalahan seperti nilai kosong atau ketidakkonsistenan format.

Alur pra-pemrosesan data ditunjukkan pada Gambar **??**. Proses ini diawali dengan pemuatan *dataset* ke dalam lingkungan komputasi serta pemeriksaan struktur dan konsistensi data. Selanjutnya dilakukan *Exploratory Data Analysis* (EDA) untuk memperoleh gambaran umum karakteristik data. Tahap akhir pra-pemrosesan mencakup eliminasi variabel non-determinan serta penetapan prevalensi *stunting* sebagai variabel target (y) dan indikator determinan sebagai variabel prediktor (X).

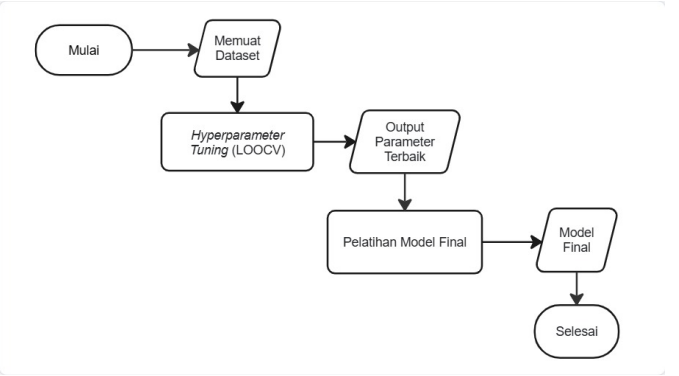


Gambar 3.3 Alur Pra-pemrosesan Data

3.2.3 Pembangunan Model *Random Forest*

Setelah melalui tahapan penyiapan dan pra-pemrosesan data, *dataset* yang telah bersih dan terstruktur selanjutnya digunakan pada tahap pemodelan. Pada penelitian ini, pendekatan yang digunakan adalah *Random Forest Regression* untuk memodelkan hubungan antara variabel determinan *stunting* dan nilai prevalensi *stunting* pada tingkat provinsi yang memiliki nilai numerik kontinu. Pemilihan metode ini didasarkan pada kemampuannya dalam menangani data dengan jumlah fitur yang relatif banyak serta hubungan non-linier antarvariabel

tanpa asumsi distribusi data yang ketat **shen2023machine**. Alur proses pemodelan disajikan secara ringkas pada Gambar **??**.



Gambar 3.4 Alur Pembangunan Model *Random Forest Regression*

3.2.3.1 *Hyperparameter Tuning* Berbasis *Cross Validation*

Tahap awal pembangunan model difokuskan pada pencarian konfigurasi model yang paling optimal melalui strategi validasi yang ketat. Mengingat keterbatasan jumlah sampel data yang hanya tersedia pada tingkat provinsi, penelitian ini menerapkan metode *Leave-One-Out Cross Validation* (LOOCV) sebagai pengganti pembagian data latih-uji konvensional **cha2021comparison**. Dalam skema ini, proses pelatihan dilakukan secara iteratif sebanyak jumlah sampel, di mana pada setiap iterasi, satu provinsi digunakan sebagai data uji sementara provinsi sisanya menjadi data latih. Pendekatan ini dipilih untuk memaksimalkan pemanfaatan informasi dari data yang terbatas serta menghasilkan estimasi performa model yang tidak bias terhadap pemilihan acak data uji tertentu.

Bersamaan dengan proses validasi LOOCV, dilakukan optimasi kinerja model melalui mekanisme *Hyperparameter Tuning* menggunakan metode *Grid Search*. Tahap ini bertujuan untuk mencari kombinasi parameter terbaik meliputi jumlah pohon (`n_estimators`), kedalaman maksimal (`max_depth`),

dan jumlah sampel minimum percabangan yang menghasilkan tingkat kesalahan prediksi terendah. Eksplorasi ruang parameter ini dinilai krusial untuk mencegah terjadinya *overfitting*[1] yang sering menjadi kendala pada dataset berskala kecil **fatmawati2024random**. Kombinasi parameter yang menghasilkan nilai rata-rata *Root Mean Squared Error* (RMSE) terendah selama proses validasi silang kemudian ditetapkan sebagai konfigurasi model optimal.

3.2.3.2 Pelatihan Model Final

Setelah parameter optimal diperoleh dari tahap validasi, langkah selanjutnya adalah pembangunan model final (*final model fitting*). Pada tahap ini, model *Random Forest* dilatih ulang menggunakan keseluruhan *dataset* provinsi yang tersedia dengan menerapkan parameter terbaik hasil optimasi sebelumnya. Tujuannya adalah untuk menangkap pola data secara utuh tanpa menyisakan data untuk pengujian, karena evaluasi performa telah diselesaikan pada tahap validasi silang. Model terbaik ini kemudian digunakan untuk memetakan hubungan non-linier yang kompleks antara seluruh variabel determinan dengan prevalensi *stunting*.

Selain menghasilkan prediksi angka prevalensi, model final ini memiliki fungsi krusial untuk mengekstraksi nilai kepentingan fitur (*feature importance*). Nilai ini mengkuantifikasi kontribusi relatif setiap indikator determinan terhadap hasil prediksi model secara global. Informasi mengenai bobot kontribusi fitur inilah yang nantinya menjadi landasan utama dalam analisis prioritas intervensi penanganan *stunting* pada tahap interpretasi hasil menggunakan metode SHAP.

[1] *Overfitting* merupakan kondisi saat model terlalu mempelajari *noise* data latih sehingga kehilangan kemampuan generalisasi pada data baru.

3.2.4 Evaluasi Performa Model

Evaluasi performa model dilakukan untuk mengukur sejauh mana model *Random Forest Regression* mampu memprediksi prevalensi *stunting* secara akurat pada tingkat provinsi. Mengingat penelitian ini bersifat regresi, digunakan beberapa metrik evaluasi berbasis kesalahan prediksi dan kekuatan penjelasan model. Penggunaan lebih dari satu metrik bertujuan untuk memberikan gambaran performa model secara komprehensif, baik dari sisi rata-rata kesalahan absolut, sensitivitas terhadap kesalahan besar, maupun kemampuan model dalam menjelaskan variasi data target. Seluruh metrik dievaluasi pada skema validasi silang yang telah ditetapkan pada tahap pemodelan.

1. ***Mean Absolute Error*** **(MAE)**

   MAE digunakan untuk mengukur rata-rata selisih absolut antara nilai prevalensi *stunting* hasil prediksi model dengan nilai aktual pada masing-masing provinsi. Metrik ini memberikan gambaran kesalahan prediksi secara langsung dalam satuan persentase, sehingga memudahkan interpretasi seberapa jauh hasil prediksi model menyimpang dari data SSGI 2024 yang sebenarnya.

2. ***Mean Squared Error*** **(MSE)**

   MSE digunakan untuk mengevaluasi kesalahan prediksi dengan memberikan penalti lebih besar pada selisih prediksi yang tinggi antarprovinsi. Penggunaan metrik ini penting untuk memastikan bahwa model tidak menghasilkan kesalahan ekstrem pada provinsi tertentu yang dapat mengaburkan pola determinan *stunting* secara nasional.

3. ***Root Mean Squared Error*** **(RMSE)**

   RMSE digunakan sebagai indikator utama akurasi model karena mengembalikan nilai kesalahan ke dalam skala yang sama dengan prevalensi *stunting*. Dalam konteks penelitian ini, RMSE membantu

menilai seberapa besar deviasi prediksi model terhadap nilai aktual *stunting* antarprovinsi secara keseluruhan.

4. ***R-squared*** **(R^2)**

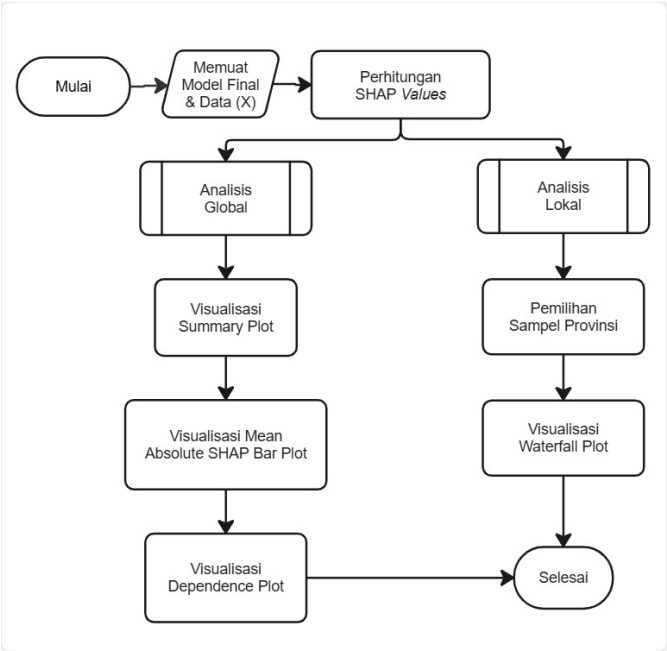
   Koefisien determinasi R^2 digunakan untuk mengukur sejauh mana variasi prevalensi *stunting* antarprovinsi dapat dijelaskan oleh variabel determinan yang dimasukkan ke dalam model. Metrik ini memberikan gambaran kemampuan model dalam menangkap hubungan kompleks antara faktor kesehatan ibu, balita, dan lingkungan dengan tingkat *stunting*.

5. ***Mean Absolute Percentage Error*** **(MAPE)**

   MAPE digunakan untuk mengukur kesalahan prediksi dalam bentuk persentase relatif terhadap nilai aktual prevalensi *stunting*. Metrik ini membantu melihat konsistensi performa model antarprovinsi, terutama ketika terdapat perbedaan tingkat *stunting* yang cukup besar antara satu wilayah dan wilayah lainnya.

3.3 Interpretasi Hasil dengan SHAP

Interpretasi keputusan model *Random Forest* yang bersifat *black-box* dilakukan menggunakan metode SHAP guna mengidentifikasi besaran pengaruh faktor determinan secara transparan. Analisis mencakup tinjauan global skala nasional dan tinjauan lokal spesifik per provinsi. Alur kerja analisis interpretabilitas model ini diilustrasikan pada Gambar **??**.



Gambar 3.5 Alur Interpretasi Model dengan SHAP

3.3.1 Interpretasi Global

Interpretasi global bertujuan untuk memberikan gambaran menyeluruh mengenai faktor-faktor yang paling dominan mempengaruhi prevalensi *stunting* di tingkat nasional. Analisis ini menggabungkan nilai SHAP mutlak dari seluruh sampel data untuk mengidentifikasi hirarki kepentingan fitur serta pola hubungan umum antara determinan dan kejadian *stunting*. Dalam penelitian ini, interpretasi global diuraikan melalui dua komponen visualisasi utama:

1. ***Feature Importance Plot***

   Visualisasi berbentuk diagram batang ini mengurutkan variabel determinan berdasarkan besaran rata-rata nilai SHAP absolut, yang dinyatakan secara matematis sebagai $\frac{1}{n}\sum_{i=1}^{n}|\phi_i^{(j)}|$. Grafik ini berfungsi untuk menentukan fitur mana yang memiliki dampak paling signifikan terhadap model secara keseluruhan tanpa memandang arah pengaruhnya,

baik positif maupun negatif.

2. ***Summary Plot (Beeswarm)***

   Visualisasi distribusi nilai SHAP yang menyajikan arah pengaruh setiap variabel terhadap prediksi. Pada grafik ini, titik berwarna merah merepresentasikan nilai fitur yang tinggi, sedangkan biru merepresentasikan nilai rendah. Sebaran titik tersebut menunjukkan bagaimana nilai fitur tertentu dapat menaikkan (nilai SHAP positif) atau menurunkan (nilai SHAP negatif) risiko *stunting*.

3.3.2 Interpretasi Lokal

Interpretasi lokal dilakukan untuk memahami mekanisme prediksi model *Random Forest* pada tingkat wilayah spesifik, yaitu provinsi. Berbeda dengan interpretasi global yang menyoroti pola umum secara nasional, pendekatan ini berfokus pada kontribusi variabel determinan terhadap prediksi prevalensi *stunting* pada masing-masing provinsi. Analisis ini penting karena faktor dominan penyebab *stunting* dapat bervariasi antarwilayah meskipun memiliki tingkat prevalensi yang relatif serupa. Alur interpretasi lokal menggunakan metode SHAP dalam penelitian ini ditunjukkan pada Gambar **??**.

3.3.2.1 Pemilihan Sampel Provinsi

Mengingat interpretasi lokal bersifat mendalam dan kontekstual, tidak seluruh provinsi dianalisis secara individual. Oleh karena itu, dilakukan pemilihan sampel provinsi yang bertujuan untuk merepresentasikan variasi kondisi *stunting* di Indonesia. Pemilihan sampel didasarkan pada distribusi nilai prevalensi *stunting* hasil prediksi model, sehingga mencakup provinsi dengan tingkat prevalensi relatif tinggi, rendah, serta kondisi menengah. Pendekatan ini memungkinkan analisis yang lebih terfokus tanpa mengurangi representativitas karakteristik wilayah secara nasional.

3.3.2.2 Visualisasi dan Analisis *Waterfall Plot*

Setelah sampel provinsi ditetapkan, dilakukan interpretasi lokal menggunakan visualisasi *Waterfall Plot*. Visualisasi ini digunakan untuk menguraikan proses pembentukan nilai prediksi prevalensi *stunting* oleh model pada satu provinsi tertentu. Proses dimulai dari nilai rata-rata prediksi model (*base value*), kemudian setiap variabel determinan ditampilkan sebagai kontribusi aditif yang dapat meningkatkan atau menurunkan nilai prediksi hingga mencapai nilai akhir.

Melalui *Waterfall Plot*, kontribusi masing-masing variabel dapat diamati secara eksplisit baik dari segi arah maupun besar pengaruhnya. Variabel dengan kontribusi positif merepresentasikan faktor yang mendorong peningkatan risiko *stunting*, sedangkan kontribusi negatif menunjukkan faktor yang berperan dalam menurunkan risiko tersebut. Pendekatan ini memberikan transparansi terhadap keputusan model serta menjadi dasar analisis perbedaan determinan *stunting* antarprovinsi pada tahap pembahasan hasil.

3.4 Konfigurasi Model dan Data

3.4.1 Lingkungan Pengembangan Model

Proses pembangunan model dan pengolahan data dilakukan dalam lingkungan bahasa pemrograman `Python` versi 3.10. Untuk menjaga reproduksibilitas penelitian, berikut adalah daftar pustaka (*library*) utama beserta versinya dan peran spesifiknya dalam sistem:

1. `scikit-learn` (v1.7.2), digunakan untuk implementasi algoritma *Random Forest*, pembagian data *Leave-One-Out Cross-Validation*, serta optimasi *hyperparameter*.
2. `shap` (v0.49.1), digunakan sebagai instrumen interpretasi model untuk menghitung nilai kontribusi tiap fitur melalui pendekatan *Shapley values*.
3. `pandas` (v2.3.2), digunakan untuk manipulasi struktur data, khususnya dalam proses agregasi dari data individu ke tingkat provinsi.

4. `numpy` (v2.3.3), mendukung operasi komputasi matriks dan pemrosesan *array* numerik yang efisien.
5. `matplotlib` (v3.10.6) & `seaborn` (v0.13.2), digunakan untuk memvisualisasikan distribusi variabel dan hasil interpretasi fitur secara grafis.
6. `scipy` (v1.16.2), digunakan untuk pengolahan statistik tingkat lanjut yang diperlukan dalam pembersihan data.

3.4.2 Karakteristik Data

Dataset yang digunakan dalam penelitian ini bersumber dari Laporan Publikasi SSGI tahun 2024 yang diterbitkan oleh Kementerian Kesehatan Republik Indonesia **kemenkes2025ssgi**. Data tersebut merupakan data sekunder berskala nasional yang menyajikan berbagai indikator status gizi, kesehatan ibu dan anak, serta kondisi lingkungan dan sosial ekonomi pada tingkat provinsi. Setelah melalui tahap penyiapan dan seleksi data, diperoleh satu dataset terstruktur yang terdiri dari 56 variabel independen dan satu variabel dependen berupa prevalensi *stunting*. Untuk memudahkan pemahaman struktur data dan menjaga konsistensi analisis determinan, seluruh variabel independen kemudian dikelompokkan ke dalam beberapa kategori tematik berdasarkan karakteristik indikator yang diwakilinya.

Tabel 3.1 Distribusi Variabel Determinan Stunting berdasarkan Kategori

No	Kategori Variabel	Jumlah Fitur
1	Riwayat Kehamilan dan Kelahiran	15
2	Riwayat Pemberian ASI dan MP-ASI	17
3	Akses dan Pemanfaatan Pelayanan Kesehatan	9
4	Imunisasi dan Morbiditas Balita	8
5	Kepemilikan dan Pemanfaatan Jaminan Kesehatan	2
6	Pendampingan Keluarga dan Pengetahuan Stunting	4
7	Kontrasepsi Pasca Bersalin	1
8	Bantuan Sosial	1
9	Akses Air Minum dan Sanitasi	4
Total		**56**

Tabel **??** menunjukkan bahwa variabel determinan yang digunakan dalam penelitian ini didominasi oleh indikator terkait riwayat kehamilan, pola pemberian makan bayi dan anak, serta akses pelayanan kesehatan. Komposisi ini mencerminkan pendekatan multidimensi dalam menganalisis *stunting*, di mana faktor biologis, perilaku, layanan kesehatan, serta kondisi lingkungan dianalisis secara simultan. Pengelompokan ini juga menjadi dasar dalam interpretasi model, khususnya pada tahap analisis kontribusi fitur menggunakan SHAP.

3.5 Ilustrasi Perhitungan Metode

Bagian ini menyajikan simulasi pemrosesan data SSGI 2024 oleh model untuk menghasilkan prediksi prevalensi *stunting*. Angka yang digunakan di sini adalah data hipotetis untuk menggambarkan alur kalkulasi dari *input* fitur hingga evaluasi akhir.

3.5.1 Prediksi *Random Forest Regression*

Dalam penelitian ini, prediksi prevalensi *stunting* untuk satu provinsi tidak ditentukan oleh satu pohon keputusan tunggal, melainkan melalui konsensus dari banyak pohon (*ensemble*). Setiap pohon dalam model akan memberikan estimasi angka prevalensi berdasarkan fitur *input* yang dipelajarinya.

Berikut adalah ilustrasi jika model terdiri dari 3 pohon keputusan, di mana masing-masing pohon memberikan estimasi angka *stunting* sebagai berikut:

Tabel 3.2 Data Hipotetis Output Pohon Keputusan

Pohon Keputusan	***Output* Prediksi (h_i)**
Tree 1	18.5
Tree 2	17.9
Tree 3	18.2

Nilai prediksi akhir ($\hat{y}$) untuk provinsi tersebut diperoleh dengan merata-

ratakan seluruh *output* pohon menggunakan persamaan berikut:

Selanjutnya, proses agregasi dilakukan untuk menyederhanakan struktur data sehingga seluruh variabel memiliki skala yang seragam. Pada tabel yang memiliki lebih dari satu kategori indikator, proses agregasi dilakukan melalui pendekatan penggabungan indikator berbasis risiko. Sebagai contoh, pada tabel usia kehamilan ibu, nilai persentase kehamilan usia ekstrem dijumlahkan untuk membentuk satu variabel baru.

$$X_{\text{risiko usia kehamilan}} = X_{<20} + X_{>35}$$

Di mana $X_{<20}$ adalah persentase kehamilan usia di bawah 20 tahun dan $X_{>35}$ adalah persentase kehamilan usia di atas 35 tahun. Pendekatan serupa diterapkan pada tabel lain yang memiliki kategori majemuk dengan prinsip mempertahankan indikator risiko utama terhadap *stunting*. Seluruh variabel direpresentasikan dalam bentuk persentase guna menjaga konsistensi satuan data. Hasil dari tahap ini adalah *dataset* terstruktur berbasis provinsi yang siap digunakan sebagai input pada tahapan analisis selanjutnya. Angka **18.2%** inilah yang menjadi prediksi final prevalensi *stunting* untuk provinsi tersebut.

3.5.2 Perhitungan *Error* Prediksi

Validasi akurasi dilakukan dengan membandingkan angka prediksi model terhadap data aktual SSGI. Selisih absolut antara kedua nilai ini menunjukkan seberapa jauh deviasi estimasi model untuk satu provinsi.

Misalkan data aktual dan hasil prediksi untuk sampel provinsi tersebut adalah:

Tabel 3.3 Data Hipotetis Aktual vs Prediksi

Data	**Nilai (‰)**
Aktual (y)	20.0
Prediksi ($\hat{y}$)	18.2

Perhitungan deviasi atau *error* individu (e) adalah:

$$e = |y - \hat{y}|$$

Penyelesaian:

$$e = |20.0 - 18.2|$$

$$e = \mathbf{1.8}$$

Hasil ini menunjukkan bahwa prediksi model meleset sebesar **1.8 poin persentase** dari angka *stunting* yang sebenarnya.

3.5.3 Perhitungan Metrik Evaluasi

Untuk mengukur performa model secara menyeluruh terhadap seluruh provinsi di Indonesia, digunakan perhitungan agregat. Berikut adalah ilustrasi perhitungan menggunakan sampel data dari 3 provinsi (A, B, dan C).

Tabel 3.4 Data Hipotetis untuk Evaluasi Agregat

Sampel	**Aktual** (y)	**Prediksi** ($\hat{y}$)	**Selisih** ($\lvert e \rvert$)	**Kuadrat** (e^2)
A	20	18	2	4
B	15	16	1	1
C	25	23	2	4

1. MAE

Mean Absolute Error (MAE) dihitung untuk mengetahui rata-rata penyimpangan absolut prediksi *stunting* dalam satuan persen.

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

Penyelesaian:

$$\text{MAE} = \frac{2+1+2}{3} = \mathbf{1.67}$$

Model memiliki rata-rata kesalahan prediksi sebesar **1.67 poin**.

2. MSE

Mean Squared Error (MSE) dihitung untuk memberikan bobot lebih pada kesalahan prediksi yang ekstrem, guna memastikan model sensitif terhadap data provinsi dengan selisih *stunting* yang besar.

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

Penyelesaian:

$$\text{MSE} = \frac{4+1+4}{3} = \mathbf{3.00}$$

3. RMSE

Root Mean Squared Error (RMSE) mengembalikan nilai kesalahan ke satuan asli (persentase) untuk menggambarkan standar deviasi dari residu prediksi model terhadap data SSGI.

$$\text{RMSE} = \sqrt{\text{MSE}}$$

Penyelesaian:

$$\text{RMSE} = \sqrt{3.00} \approx \mathbf{1.73}$$

Deviasi standar kesalahan model berada pada angka **1.73 poin**.

4. MAPE

Mean Absolute Percentage Error (MAPE) digunakan untuk melihat rasio kesalahan model relatif terhadap besaran angka *stunting* aktual di masing-masing provinsi.

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Penyelesaian:

$$\text{MAPE} = \frac{100\%}{3}(0.10 + 0.067 + 0.08)$$

$$= \mathbf{8.22}\%$$

Rata-rata kesalahan model adalah sebesar **8.22%** dari nilai aktual.

3.6 Ilustrasi Rancangan Pengujian

Bagian ini menjabarkan skenario eksperimen yang dilakukan untuk melatih dan memvalidasi model. Mengingat jumlah observasi yang terbatas yaitu provinsi di Indonesia, strategi pembagian data dan pemilihan parameter menjadi krusial untuk mencegah *overfitting* dan memastikan hasil evaluasi yang objektif.

3.6.1 Skema *Leave-One-Out Cross Validation* (LOOCV)

Penelitian ini menerapkan metode *Leave-One-Out Cross Validation* (LOOCV) untuk memaksimalkan penggunaan data. Dalam skema ini, evaluasi dilakukan melalui *N* kali iterasi, di mana pada setiap iterasi, satu provinsi disisihkan sebagai data uji (*testing set*) sementara sisa provinsi lainnya digunakan sebagai data latih (*training set*).

Berikut adalah ilustrasi pembagian data pada setiap iterasi pengujian:

Tabel 3.5 Ilustrasi Skema Pembagian Data LOOCV

Iterasi	**Data Latih (N-1 Provinsi)**	**Data Uji (1 Provinsi)**
1	Prov 2, 3, ..., N	**Provinsi 1**
2	Prov 1, 3, ..., N	**Provinsi 2**
...	...	...
N	Prov 1, 2, ..., N	**Provinsi N**

Dengan mekanisme ini, model diuji pada setiap provinsi tepat satu kali tanpa bias pemilihan data, dan performa akhir dihitung dari rata-rata *error* seluruh iterasi.

3.6.2 *Hyperparameter Tuning*

Untuk mendapatkan konfigurasi model yang paling optimal dalam memprediksi *stunting*, dilakukan pencarian parameter terbaik menggunakan teknik *Grid Search*. Berbagai kombinasi parameter seperti jumlah pohon dan kedalaman pohon diuji performanya menggunakan skema LOOCV di atas.

Tabel berikut mengilustrasikan simulasi hasil pencarian parameter, di mana keputusan pemilihan didasarkan pada nilai rata-rata RMSE terendah.

Tabel 3.6 Data Hipotetis Hasil Tuning Parameter

No	***n_estimators***	***max_depth***	**RMSE**	**Keputusan**
1	100	5	2.10	-
2	**200**	**5**	**1.85**	**Dipilih (Terbaik)**
3	200	10	1.92	-

Berdasarkan ilustrasi Tabel **??**, kombinasi nomor 2 dipilih karena menghasilkan tingkat kesalahan prediksi yang paling minim.

3.6.3 Pemilihan Model Final

Tahap terakhir adalah pembentukan model final untuk kebutuhan interpretasi fitur. Setelah parameter terbaik ditemukan (misal: *n_estimators*=200 dan *max_depth*=5), model tersebut tidak lagi dipisahkan antara data latih dan uji.

Sebagai langkah akhir, model dilatih ulang (*retraining*) menggunakan seluruh *dataset* penuh. Model yang telah mempelajari pola dari seluruh data provinsi inilah yang nantinya digunakan untuk analisis determinan penyebab *stunting* menggunakan metode SHAP pada Bab IV.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

Berisi hasil penelitian berdasarkan rancangan yang sudah dijelaskan pada Bab **??**, terutama dari Subbab **??**. Bagi yang membuat alat, jelaskan alat yang jadi dalam bentuk apa. Bagi yang membuat aplikasi, jelaskan aplikasi yang jadi dalam bentuk seperti apa. Jabarkan dalam bentuk pseudocode dan dijelaskan per bagian kodenya. Gunakan gambar dan tabel sebagai alat bantu menjelaskan hasil.

Contoh implementasi kode dapat ditulis menggunakan `\begin{lstlisting}`. Contoh kode dapat dilihat pada Kode **??**.

Kode 4.1 Akuisisi Gambar

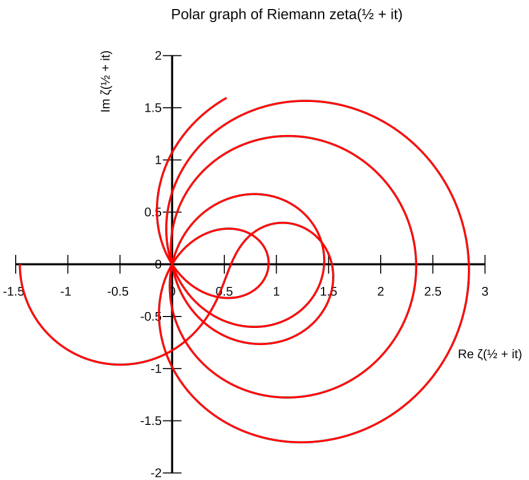
```
def process_dataset(dataset_path):
        image_files = glob(os.path.join(dataset_path, '*.png'))
        image_files.sort()
        for image_file in image_files:
                frame = cv2.imread(image_file)
                if frame is None:
                        continue
                frame_rgb = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)
                cv2.imshow('Frame', frame)
                if cv2.waitKey(1) & 0xFF == ord('q'):
                        break
        cv2.destroyAllWindows()
def main():
        datasets = get_all_dataset_folders(DATASET_ROOT)
        for dataset in datasets:
                process_dataset(dataset)
                print("print string")
```

4.2 Hasil Pengujian

Berikan hasil pengujian berdasarkan rancangan & skenario yang sudah direncanakan sebelumnya pada Subbab **??**.

Tabel 4.1 Data *dummy* Pengujian

Subjek	Hasil Prediksi (BPM)							GT
	F	NA	NO	RC	LC	M	C	
1	68	69	68	70	68	71	69	68
2	69	69	68	70	68	71	69	69
3	70	70	69	71	68	73	69	70
4	71	70	70	72	69	73	70	71
5	72	72	70	72	70	74	70	72



Gambar 4.1 Contoh Graf Pengujian

4.3 Analisis Hasil Penelitian

Berikan analisis hasil penelitian & pengujian, berupa data yang didapatkan dari penelitian & pengujian Tugas Akhir yang sudah anda kerjakan. Gunakan gambar dan tabel sebagai alat bantu menjelaskan analisis hasil. Data luaran penelitian yang dapat dianalisis berupa:

1. Hasil pengujian

2. Hasil kuesioner
3. Aplikasi yang dikembangkan

Analisis dapat membandingkan dengan hasil penelitian sebelumnya yang memiliki kemiripan topik.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berisi kesimpulan dari hasil dan pembahasan terkait penelitian yang dilakukan, dapat juga berupa temuan yang Anda dapatkan setelah melakukan penelitian atau analisis terhadap tugas akhir Anda. Memberikan jawaban dari poin pada subbab **??** dan **??**.

5.2 Saran

Berisi saran mengenai aspek tugas akhir atau temuan yang dapat dikembangkan dan diperkaya di tugas akhir selanjutnya. Saran dapat berkaitan erat pada subbab **??**.

LAMPIRAN

A Dataset

B Hasil Wawancara

C Rincian Kasus Uji