BAB II
TINJAUAN PUSTAKA

Pada tahap ini, peneliti akan menjelaskan dasar teori yang mendukung penulisan dan pengerjaan Tugas Akhir. Dasar teori ini diperoleh dari berbagai sumber, termasuk jurnal ilmiah dan skripsi temuan penelitian yang berkaitan dengan topik yang dibahas.

2.1 Peneliti Terdahulu

Beberapa penelitian yang menjadi acuan pada tugas akhir.

No	Peneliti	Judul penelitian	Tahun	Metode	Akurasi
1	Josua Fernandes Nababan	Klasifikasi penderita stunting dengan metode support vector machine	2019	SVM	80,8%
2	Mahmin, Dinda & Oloan	Klasifikasi penyakit stunting dengan menggunakan algoritma SVM dan Random Forest	2023	SVM, Random Forest	81%
3	Otong Saeful	Implementasi Algoritma K-Nearest Neighbor untuk prediksi stunting pada anak	2020	KNN	84,37%
4	Widya, Fida & Agung	Prediksi stunting pada balita di rumah sakit kota	2023	Naïve Bayes	85,33%

		semarang menggunakan naïve bayes			

Tabel 1 Literatur Review

Tabel ini menggambarkan berbagai metode dan pendekatan yang telah digunakan dalam penelitian terkait stunting. Pada pada penelitian pertama yang dilakukan pada tahun 2021 memiliki kelebihan dalam penggunaan metode svm ini akurasi yang tinggi dan memiliki kekurangan pemilihan parameter yang kompleks seperti jenis kernel, parameter c, dan gamma untuk kernel rbf [3]. Penelitian kedua terbit pada tahun 2023 memiliki kelebihan dalam menggabungkan 2 metode yaitu svm dan random forest dan memiliki kekurangan pada pemilihan parameter seperti pada penelitian pertama demikian pula, random forest membutuhkan penentuan jumlah pohon dan kedalaman maksimum pohon yang optional [7]. Penelitian ketiga terbit pada tahun 2020 memiliki kelebihan pelatihan model yang tidak rumit dengan kesederhanaan dan implementasi namun kekurangannya ketidakefektifan pada data berukuran besar [4]. Penelitian ke-empat terbit pada tahun 2023 memiliki kelebihan kesederhanaan dan kecepatan yang mudah diimplementasikan untuk melakukan prediksi dan kekurangannya asumsi independen yang kuat mengacu pada anggapan bahwa setiap fitur dalam dataset tidak memiliki hubungan langsung satu sama lain saat melakukan prediksi Ini berarti Naive Bayes menganggap bahwa nilai-nilai fitur (seperti faktor kesehatan dalam prediksi stunting) tidak saling mempengaruhi secara langsung. Namun, dalam kenyataannya, fitur-fitur dalam data sering kali saling terkait atau berkorelasi. Oleh karena itu, jika fitur-fitur tersebut berkorelasi secara signifikan, asumsi independensi Naive Bayes dapat mengurangi akurasi model dalam memprediksi stunting atau kondisi lainnya [8]. Stunting adalah kondisi kekurangan gizi kronis yang mengakibatkan pertumbuhan fisik dan perkembangan anak terganggu, sehingga tinggi badannya lebih rendah dibandingkan anak-anak seusianya [9].

2.1.1 Faktor Penyebab Stunting

Beberapa factor penyebab stunting antara lain ialah kekurangan nutrisi asupan makanan yang tidak mencukupi kebutuhan gizi anak, infeksi berulang seperti diare yang sering menyerang anak, sanitasi dan kebersihan lingkungan yang tidak sehat dan sanitasi

yang buruk, akses layanan kesehatan yang terbatas terhadap kesehatan ibu dan anak terakhir pendidikan dan pengetahuan orang tua tentang gizi dan kesehatan [10].

2.2 Machine learning

Pembelajaran mesin adalah sebuah alat kecerdasan buatan (AI) yang memungkinkan komputer belajar dari data untuk membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Terdapat beberapa jenis pembelajaran, yaitu pembelajaran terawasi (supervised learning), tidak terawasi (unsupervised learning), semi-terawasi (semi-supervised learning), dan penguatan (reinforcement learning). Komponen utama machine learning meliputi data, algoritma, model, serta proses pelatihan dan pengujian [11]. Dalam konteks penyakit stunting, teknik seperti XGBoost dan Random Forest digunakan untuk menganalisis faktor risiko dan memprediksi kemungkinan terjadinya stunting berdasarkan data kesehatan [1]. Aplikasi machine learning dalam bidang ini juga dapat membantu dalam pengembangan model intervensi yang lebih efektif. Tantangan utama seperti overfitting, underfitting, dan interpretasi hasil tetap relevan dalam mengimplementasikan teknik ini untuk kesehatan masyarakat.

2.3 XGBoost

XGBoost (Extreme Gradient Boosting) adalah salah satu algoritma boosting yang sangat efisien dan fleksibel. Algoritma ini menggabungkan sejumlah model prediksi yang lemah untuk membentuk model yang kuat, sehingga menghasilkan performa prediktif yang tinggi [12]. Dalam konteks prediksi penyakit stunting, XGBoost dapat digunakan untuk mengidentifikasi faktor-faktor risiko yang signifikan dan memberikan prediksi yang akurat mengenai kemungkinan terjadinya stunting [9].

2.3.1 Mekanisme Kerja XGBoost

XGBoost bekerja dengan membangun model secara iterative, dimana setiap model bertujuan untuk mengurangi kesalahan dari model sebelumnya. Proses ini melibatkan prediksi awal dengan sederhana lalu menghitung gradient kemudian melatih model baru untuk memperbaiki kesalahan dari prediksi sebelumnya terakhir menggabungkan model baru dengan model sebelumnya untuk membuat prediksi yang lebih akurat [13]. Berikut beberapa tahapan cara kerja xgboost secara rinci :

1. **Inisialisasi Model**

   Xgboost mulai dengan prediksi awal yang biasanya adalah nilai rata-rata dari target atau probabilitas kelas mayoritas.

2. **Iterasi Gradient Boosting**

   Gradient Boosting ini dilakukan dalam beberapa ilterasi jadi setiap iterasi menambahkan model baru yang bertujuan untuk memperbaiki kesalahan model sebelumnya.

3. **Pembuatan Pohon Keputusan**

   Setiap iterasi, sebuah pohon keputusan (tree) baru dibangun dan model yang ada digunakan untuk memprediksi nilai target dan juga residu perbedaan antara prediksi dan nilai actual dihitung.

4. **Menghitung Gradient**

   Gradient dari fungsi loss dihitung berdasarkan residu yang dihasilkan dan gradient ini menunjukkan arah dan besarnya perbaikan yang diperlukan untuk mengurangi kesalahan.

5. **Regularisasi**

   Xgboost menambahkan penalty untuk kompleksitas model (jumlah leaf nodes dan besarnya nilai leaf) untuk mencegah overftting. Penalty ini dikendalikan oleh hyperparameter seperti :

   - **n_estimators** : jumlah pohon yang akan dibangun
   - **max_depth** : kedalaman maksimum setiap pohon
   - **learning_rate** : kecepatan pembelajaran
   - **subsample** : proporsi sampel yang digunakan untuk membangun setiap pohon
   - **colsample_bytree** : proporsi fitur yang dipilih secara acak untuk setiap pohon
   - **gamma** : minimum loss reduction untuk membuat split baru di pohon
   - **lambda dan alpha** : parameter regularisasi L2 dan L1

- **min_child_weight** : hyperparameter penting dalam xgboost yang membantu mengendalikan overfitting dengan menentukan jumlah minimum sum dari instansi weight yang diperlukan di setiap leaf node.
- **Random_state** : parameter yang digunakan untuk mengontrol pengacakan randomization dalam proses pelatihan model.
- **Eval_metric** : parameter yang digunakan untuk menentukan metric evaluasi yang akan digunakan selama pelatiha model memantau kinerja.
- **Use_label_encoder** : parameter yang menetukan apakah xgboost akan menggunakan 'labelEncoder' dari scikit-learn untuk mengonversi label (target) menjadi numeric dan disini terdapat nilai false and true yang berarti untuk false (tidak menggunakan labelEncoder dan mengasumsikan label sudah dalam bentuk numeric) dan untuk true (menggunakan labelEncoder utnuk mengonversi label)

2.3.2 Kelebihan dan Kekurangan XGBoost

XGBoost dikenal dengan kecepatan dan efisiensinya dalam pelatihan model serta kemampuan untuk menangani data yang tidak seimbang dan missing values. Namun xgboost juga rentan terhadap overfitting jika tidak diatur dengan benar.

2.4 Random Forest

Random Forest adalah algoritma ensemble learning yang menggabungkan sejumlah pohon keputusan untuk meningkatkan akurasi prediksi. Setiap pohon dalam hutan acak dibangun dari subset data yang berbeda dan menggunakan subset fitur yang berbeda, sehingga mengurangi overfitting dan meningkatkan generalisasi model [14]. Dalam analisis penyakit stunting, Random Forest dapat membantu dalam menentukan variabel-variabel penting dan memberikan model prediksi yang handal [9].

2.4.1 Mekanisme Kerja Random Forest

Random Forest bekerja dengan (Bootstrap Sampling). membuat beberapa subset data dari dataset dengan teknik bootstrap lalu melatih pohon keputusan pada setiap subset data dan terakhir menggabungkan hasil prediksi dari semua pohon keputusan untuk mendapatkan prediksi akhir melalui voting atau averaging [15].

2.4.2 Kelebihan dan Kekurangan Random Forest

Random Forest ini sangat baik dalam pencegahan overfitting dan dapat menangani data yang hilang dan tidak seimbang. Namun juga memiliki kekurangan terutama dalam hal kompleksitas model.

2.5 Implementasi kedua algoritma

Implementasi kedua algoritma XGBoost dan Random Forest dalam studi penyakit stunting bertujuan untuk mengidentifikasi faktor risiko utama dan membangun model prediksi yang dapat memberikan informasi yang berharga untuk upaya pencegahan dan intervensi [16]. XGBoost dan Random Forest digunakan untuk menganalisis data kesehatan dan gizi yang kompleks, menghasilkan prediksi dengan akurasi tinggi serta mengidentifikasi variabel penting yang berkontribusi terhadap terjadinya stunting . Dengan pendekatan ini, diharapkan dapat meningkatkan pemahaman terhadap penyakit stunting dan mendukung pengembangan strategi intervensi yang lebih efektif [8].

2.6 Perbandingan Algoritma

Kedua algoritma Xgboost dan Random Forest, memiliki kelebihan dan kekurangan masing-masing dalam analisis penyakit stunting. XGBoost sering kali lebih unggul dalam hal akurasi prediksi dan kemampuan menangani fitur yang tidak seimbang, sementara Random Forest lebih mudah diinterpretasikan dan lebih sedikit memerlukan tuning parameter. Kombinasi penggunaan kedua algoritma ini dapat memberikan pemahaman yang lebih komprehensif tentang faktor-faktornya risiko stunting dan membantu dalam pengembangan strategi intervensi yang efektif [17].

2.7 Evaluasi Model

Evaluasi model dapat mencakup berbagai aspek seperti metrix evaluasi, confusion matrix, cross-validation, feature importance, model comparison, kemudahan implementasi, keterbatasan dan tantangan. Berikut adalah bagian evaluasi yang bisa ditambahkan dalam tinjauan pustaka

2.7.1 Metrix Evaluasi

1. Akurasi merupakan metrix yang mengukur seberapa baik model prediksi dalam mengklasifikasikan data dengan benar. Xgboost sering kali menunjukkan akurasi yang lebih tinggi dibandingkan dengan Random Forest, terutama dalam data yang kompleks dan besar. Ini disebabkan oleh kemampuan XGBoost dalam menangani interaksi non-linear dan melakukan optimasi yang lebih baik melalui boosting. Begitu juga dengan Random Forest memiliki akurasi yang baik dan sering kali stabil. Namun, dalam beberapa kasus, akurasi Random Forest mungkin sedikit lebih rendah dibandingkan XGBoost karena tidak menggunakan boosting [15]. Berikut ini rumus untuk accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
2. Presisi merupakan proporsi prediksi positif yang benar-benar positif
   Berikut rumus untuk Precision = $\frac{TP}{TP+FP}$
3. Recall merupakan proporsi dari total positif yang benar-benar terdeteksi dari model dan ini rumusnya recall = $\frac{TP}{TP+FN}$
4. F1-Score merupakan perbandingan antara data recall dan presisi dari percobaan. Berikut rumusnya F1-score = $2 * \frac{precision * recall}{precision+recall}$

2.8 Confusion Matrix

Sebuah tabel yang digunakan untuk evaluasi kerja algoritma klasifikasi yang berisi jumlah data uji true positive, true negative, false positive dan false negative.

	Prediksi positive	Prediksi negative
Actual positive (1)	True positive (TP)	False negative (FN)
Actual negative (0)	False positive (FP)	True negative (TN)

Tabel 2 Confusion Matrix

Keterangan :

- TP : jumlah kasus yang diklasifikasikan dengan benar sebagai 1
- TN : jumlah non kasus yang diklasifikasikan dengan benar sebagai 0
- FP : jumlah non kasus yang diklasifikasikan yang salah sebagai 1
- FN : jumlah kasus yang telah diklasifikasikan dengan salah sebagai 0

2.9 Cross-Validation

Merupakan teknik yang digunakan untuk mengevaluasi kinerja model machine learning. Validasi silang ini mengurangi kumpulan data menjadi subkumpulan atau fold dan model ini dilatih pada beberapa fold dan juga diuji pada fold yang tersisa. Proses ini diulang beberapa kali yang nantinya hasilnya dirata-rata untuk mendapatkan estimasi kinerja model yang lebih andal. Dilakukannya proses tersebut agar mengurangi overfitting membantu dalam memastikan bahwa model tidak terlalu cocok dengan dataset tertentu, memberikan informasi yang lebih akurat tentang lingkungan kerja model dan menawarkan bantuan juga dalam pemilihan hyperprameter yang optimal untuk model.

Berikut rumusnya accuracy = $\frac{\sum klasifikasi\ benar}{data\ uji\ x\ 100\%}$

2.10 Feature Importance

Feature Importance merupakan teknik yang digunakan untuk mengidentifikasi fitur-fitur mana yang memiliki pengaruh terbesar terhadap prediksi model.

1. **Interpretabilitas Model**

   Random Forest lebih unggul dalam hal Interpretabilitas model dibandingkan dengan XGBoost. Setiap partikel di Random Forest dapat diinterpretasikan secara individual, dan pntingnya setiap fitur dapat diukur dengan mudah. Ini memudahkan

peneliti dan prediksi untuk memahami faktor-faktor utama yang berkontribusi terhadap stunting. Dalam penelitian oleh, (BioMed Central) Random forest berhasil mengidentifikasi variable-variabel penting seperti asupan gizi, kondisi sanitasi, dan ekonomi keluarga dengan jelas [18].

2. **Identifikasi Fitur yang tidak relevan**

   Menentukan fitur-fitur yang kurang penting atau tidak relevan dapat membantu dalam menyederhanakan model, mengurangi overfitting, dan meningkatkan efisiensi model.

2.11 Model Comparison

Model Comparison merupakan proses membandingkan beberapa model yang berbeda berdasarkan metrik evaluasi yang sama untuk menentukan model mana yang paling sesuai dengan data dan tujuan prediksi. Hal ini melibatkan metric evaluasi seperti akurasi, presisi, recall, dan f1-score untuk membandingkan kinerja berbagai model.

2.12 Kemudahan Implementasi

Dari segi ini, Random Forest cenderung lebih mudah diterapkan dibandingkan XGBoost. Random Forest memerlukan lebih sedikit tuning parameter dan lebih robust terhadap overfitting, sehingga cocok digunakan dalam kondisi data yang beragam dan terbatas. Meskipun XGBoost sangat kuat, ia membutuhkan tuning parameter yang lebih kompleks dan pengetahuan mendalam tentang hyperparameter untuk mendapatkan performa optimal [19].

2.13 Keterbatasan dan Tantangan

XGboost dapat menjadi sangat kompleks dan memerlukan sumber daya komputasi yang tinggi, terutama untuk dataset yang sangat besar. Di sisi lain Random Forest, meskipun lebih mudah diimplementasikan bisa mengalami penurunan performa ketika dihadapkan pada data yang sangat bervariasi atau dengan jumlah fitur yang sangat banyak. Perlu penelitian lebih lama untuk memahami fenomena ini dan mengekplorasi metode optimis yang dapat meningkatkan performa kedua algoritma [19].