BAB I

PENDAHULUAN

1.1 Latar Belakang

Stunting merupakan sebuah penyakit yang saat ini dianggap sebagai permasalahan serius dan menjadi fokus perhatian pemerintah sebagai prioritas riset nasional tahun 2020 – 2024. Berdasarkan data dari kementrian kesehatan pada tahun 2022, prevalensi stunting pada anak indonesia dengan usia dibawah lima tahun masih tergolong tinggi yakni sebesar 21,6% (Cindy, 2023). Dengan artian, penyakit stunting setidaknya mempengaruhi 22 anak dari 100 anak di bawah usia lima tahun. Prevalensi stunting di indonesia juga tergolong tinggi jika dibandingkan dengan negara negara yang ada di kawasan ASEAN. Sebagai salah satu kota penyangga IKN, Samarinda memiliki prevalensi stunting tertinggi kedua setelah kabupaten Kutai Kartanegara di provinsi Kalimantan Timur tahun 2022 dengan besar persentase 25,3% (Cindy, 2023). Stunting pada anak balita dapat menyebabkan berbagai masalah kesehatan jangka panjang dan menghambat perkembangan kognitif pada anak sehingga menyebabkan prestasi belajar yang lebih rendah jika dibandingkan dengan anak yang tidak stunting (Pratiwi et al., 2021). Oleh karena itu, penting untuk mengembangkan strategi penggunaan data mining dalam mengklasifikasi status stunting pada anak di bawah lima tahun. Metode ini diharapkan dapat menghasilkan informasi dan pengetahuan yang berguna sebagai dasar dalam pengambilan keputusan dan pertimbangan untuk mendeteksi stunting pada anak.

Berdasarkan penelitian yang sudah pernah dilakukan dalam bidang data mining dalam klasifikasi status stunting telah dilakukan dengan beberapa pendekatan, seperti algoritma *Random Forest, Support vector machine, KNN, Neural Network, Naive Bayes, classification tree* dan lain lain. Penelitian yang dilakukan oleh Apriyani & Kurniati (2020) dilakukan pengujian terhadap metode algoritma *Naive Bayes dan Support vector machine* (SVM*)* hasil penelitiaan menunjukan akurasi yang terbilang cukup baik dengan persentase akurasi sebesar 90%. Namun pada penelitian terkait masih menggunakan data berdimensi rendah (low – dimensional) atau data yang memiliki atribut sedikit untuk digunakan sebagai analisis. Data berdimensi rendah juga memiliki kelemahan seperti overfitting, dan sulit diinterpretasikan (Hakimah et al., 2022). Berdasarkan penelitian yang dilakukan Gebeye *et al.,*(2023) Model *Random Forest* (RF), *Support Vector Machine* (SVM), *Logistic Regression* (LR), *Neural Network* (NN*), dan Naive Bayes* (NB) mengalami penurunan performa saat berhadapan dengan data berdimensi tinggi, yang mempengaruhi akurasi metode klasifikasi tersebut. Hasil akurasi yang diperoleh menunjukkan bahwa Support Vector Machine (SVM) mencapai akurasi sebesar 71.03%, Random Forest (RF) mencapai 72.41%, Neural Network (NN) mencapai 71.03%, dan Naive Bayes (NB) mencapai 67.93%. Temuan ini menyoroti tantangan yang dihadapi oleh model klasifikasi saat beroperasi pada data dengan dimensi tinggi, yang dapat mengurangi efektivitas dan akurasi dari metode yang digunakan.

Data berdimensi tinggi merupakan sebuah data yang memiliki banyak atribut atau fitur yang digunakan dalam menganalisis. Sebagai contoh, bila suatu data yang memiliki atribut berjumlah puluhan bahkan ratusan atribut, maka data tersebut dapat dikategorikan sebagai data berdimensi tinggi. Data berdimensi tinggi memiliki beberapa tantangan, termasuk peningkatan kompleksitas model, risiko overfitting, dan kesulitan dalam visualisasi data (Hakimah et al., 2022). Model-model seperti *Random Forest, Support Vector Machine* , *Logistic Regression*, *Neural Network*, dan *Naive Bayes* dapat mengalami penurunan performa ketika berhadapan dengan data berdimensi tinggi (Gebeye et al., 2023), sehingga mempengaruhi akurasi klasifikasi yang didapatkan. Oleh karena itu, strategi seperti reduksi dimensi dan pemilihan fitur diperlukan untuk mengatasi masalah ini. Pemilihan fitur memungkinkan model untuk fokus pada fitur-fitur yang paling relevan dengan masalah yang dihadapi.

Algoritma klasifikasi pada penelitian ini akan menggunakan *Random Forest* (RF), hal ini berdasarkan penelitian yang dillakukan oleh (Chilyabanyama et al., 2022; Gebeye et al., 2023; Hemo, and Rayhan, 2021; Talukder and Ahammed, 2020) dimana *RF* memiliki performa terbaik jika dibandingkan dengan algoritma *Support Vector Machine, Logistic Regression, Neural Network, Naïve Bayes, linear discriminant analysis* (LDA), *k-nearest neighbors* (k-NN), *eXtreme Gradient Boosting* (Xg boost) dan *Classification and Regression Tree*s (CART) terkait klaisifikasi data stunting. Penggunaan seleksi fitur *Analysis of variances* (ANOVA) juga akan digunakan untuk menemukan atribut-atribut yang paling relevan dalam klasifikasi. Penelitian Nugroho *et al.,*(2022) metode *ANOVA* terbukti bisa meningkatkan akurasi dari metode klasfikasi KNN dan *decision tree* (DT) dimana sebelum penerapan akurasi yang didapat sebesar 67% untuk KNN dan 57% untuk DT kemudian setelah dilakukan penerapan metode *ANOVA* di dapatkan peningkatan akurasi yang tinggi dimana akurasi KNN menjadi 87% dan DT menjadi 96% pada data stunting. Kemudian dalam penelitian yang dilakukan oleh Hasan (2020) metode *ANOVA* dikombinasikan dengan metode klasifikasi *RF* maka didapatkan akurasi yang cukup tinggi sebesar 91%. Selain itu pada penelitian lain yang melakukan komparasi penggunaa seleksi fitur *ANOVA* terhadap algoritma klasifikasi mendapatkan kenaikan akurasi 0.8% menjadi 94.1% untuk Random Forest dengan tingkat akurasi paling tinggi dibandingkan dengan algoritma klasifikasi lainnya (Yoga Siswa, 2023). Dalam beberapa penelitian terkait stunting juga masih ditemukan ketidakseimbangan kelas seperti penelitian yang dilakukan oleh Sutarmi *et al.* (2023).

Ketidakseimbangan kelas terjadi ketika jumlah sampel dari satu kelas secara signifikan lebih banyak atau lebih sedikit dibandingkan dengan kelas lainnya (Luo et al., 2019). Dalam konteks stunting, hal ini bisa berarti bahwa jumlah anak yang tidak mengalami stunting jauh lebih banyak daripada yang mengalami stunting, atau sebaliknya. Masalah ini dapat menyebabkan bias pada model yang cenderung mengklasifikasi kelas mayoritas dengan lebih baik dibandingkan kelas minoritas. Untuk mengatasi masalah ketidakseimbangan kelas dan memastikan bahwa model dapat memprediksi kedua kelas (stunting dan tidak stunting) dengan akurat, bisa menggunakan teknik *oversampling* dan *undersampling* untuk menyeimbangkan kelas, oleh karena itu penelitian ini juga akan menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE).

Berdasarkan pada penelitian yang sudah pernah dilakukan Santoso et al., (2019) metode *SMOTE* terbukti mampu meningkatkan akurasi pada algoritma klasifikasi dimana Naive Bayes mencapai akurasi sebesar 62.4% dengan data asli dan meningkat menjadi 72.6% setelah menggunakan *SMOTE*. *Support Vector Machine* menunjukkan peningkatan dari 50.3% menjadi 77.0%, dan *Random Forest* meningkat dari 48.6% menjadi 80.8%. Hasil ini membuktikan bahwa penggunaan metode *SMOTE* untuk menangani ketidakseimbangan kelas mampu memberikan kenaikan akurasi yang cukup baik terutama jika dikombinasikan dengan algoritma *Random Forest*.

Penelitian ini akan menggunakan algoritma Random Forest yang sering kali menunjukkan performa lebih baik dibandingkan dengan metode lain dalam mengklasifikasi stunting, terutama ketika dikombinasikan dengan teknik pemilihan fitur dan penanganan ketidakseimbangan kelas. Namun, sebagian besar penelitian sebelumnya menggunakan data dengan dimensi yang berbeda dan teknik pemilihan fitur yang beragam. Kombinasi Random Forest dengan seleksi fitur ANOVA dan teknik SMOTE secara bersamaan bertujuan untuk meningkatkan akurasi klasifikasi stunting dengan menangani tantangan data berdimensi tinggi dan ketidakseimbangan kelas secara lebih efektif. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi baru dalam pemanfaatan teknik data mining untuk klasifikasi stunting pada anak balita, serta memberikan dasar yang lebih kuat untuk pengambilan keputusan dan intervensi yang lebih tepat dalam mengatasi masalah stunting.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan, maka didapatkan rumusan masalah pada penelitian ini sebagai berikut:

a) Fitur-fitur apa saja yang memberikan pengaruh signifikan pada algoritma *Random Forest* dengan dataset Stunting Kota Samarinda menggunakan seleksi fitur *ANOVA*?
b) Seberapa meningkat akurasi yang didapat algoritma *Random Forest* dalam mengklasifikasi data stunting Kota Samarinda dengan menggunakan seleksi fitur *ANOVA* dan metode *oversampling SMOTE*?

1.3 Tujuan Penelitian

Berikut adalah tujuan dari penelitian ini:

a) Mengidentifikasi atribut apa saja yang berpengaruh terhadap akurasi model *Random Forest* dalam Analisis Data Stunting di Kota Samarinda
b) Mengimplementasi dan mengevaluasi algoritma *Random Forest* untuk klasifikasi penyakit Stunting di Kota Samarinda, dengan penerapan seleksi fitur *ANOVA* dan metode *balancing SMOTE.* Kinerja model dievaluasi *menggunakan K-Fold Cross Validation* dan *confusion matrix*, untuk mengukur akurasi guna memastikan efektivitas model dalam klasifikasi Stunting.

1.4 Manfaat Penelitian

Diharapkan penelitian ini dapat memberikan manfaat dan pengetahuan kepada berbagai pihak, khususnya:

a) Penulis
   Untuk menambah pengetahuan dan wawasan penulis dalam mengimplementasikan metode klasifikasi dengan algoritma *Random Forest* dengan seleksi fitur *ANOVA* untuk mengingkatkan akurasi menggunakan algoritma *Random Forest* pada dataset stunting Kota Samarinda.
b) Pembaca
   Penelitian ini bermanfaat bagi pembaca yang menekuni bidang data mining, membantu mengasah kemampuan analisis data, penggunaan *Random Forest,* serta implementasi *ANOVA* dan *SMOTE* pada dataset stunting di Samarinda. Hasilnya juga dapat dijadikan referensi untuk penelitian di masa mendatang.

1.5 Batasan Masalah

Agar masalah yang dibahas tidak menjadi lebih luas lagi, maka penulis membatasi masalah penelitian sebagai berikut:

a) Data yang digunakan adalah data penyakit stunting kota samarinda pada tahun 2023.
b) Melakukan Penerapan seleksi fitur *ANOVA* pada algoritma *Random Forest* untuk memilih fitur pada data stunting berdimensi tinggi.
c) Melakukan penerapan *oversampling SMOTE* pada algoritma *Random Forest* untuk menyeimbangkan kelas pada data stunting.
d) Penelitian ini akan menggunakan atribut yang telah di seleksi dari dataset, yaitu *(i)* Nama, *(ii)* JK, *(iii)* Berat, *(iv)* Tinggi, *(v)* LiLA, *(vi)* BB/U, *(vii)* ZS BB/U, *(viii)* TB/U, *(ix)* ZS TB/U, *(x)* BB/TB, *(xii)* ZS BB/TB, *(xiii)* Naik Berat Badan, *(xiv)* Jml Vit A, *(xv)* Tanggal Pengukuran.