

SKRIPSI

**PENGGUNAAN METODE *XGBOOST* UNTUK KLASIFIKASI
STATUS OBESITAS DI INDONESIA**

Disusun dan diajukan oleh

YORIS ROMBE

H 121 15 303



**PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR**

2021

**PENGGUNAAN METODE XGBOOST UNTUK KLASIFIKASI
STATUS OBESITAS DI INDONESIA**

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada
Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu
Pengetahuan Alam Universitas Hasanuddin Makassar

YORIS ROMBE

H 121 15 303

**PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN**

MAKASSAR

2021

LEMBAR PERNYATAAN KEOTENTIKAN


Saya yang bertanda tangan dibawah ini menyatakan dengan sungguh-sungguh bahwa skripsi yang saya buat dengan judul:

**PENGGUNAAN METODE XGBOOST UNTUK KLASIFIKASI
STATUS OBESITAS DI INDONESIA**

adalah benar hasil karya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun.



Makassar, 1 November 2021


Voris Rombe

NIM. H 121 15 303

**PENGGUNAAN METODE XGBOOST UNTUK KLASIFIKASI
STATUS OBESITAS DI INDONESIA**

Disetujui oleh:

Pembimbing Utama



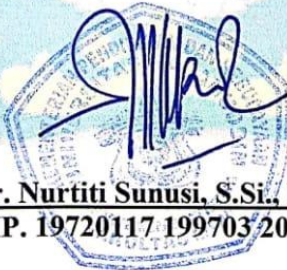
Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.
NIP. 19740713 199903 2 001

Pembimbing Pertama



Dr. Eng. Armin Lawi, M.Eng.
NIP.19720423 199512 1 001

Ketua Departemen Statistika



Dr. Nurtiti Sunusi, S.Si., M.Si.
NIP. 19720117 199703 200 2

Pada tanggal: 1 November 2021

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama : Yoris Rombe

NIM : H 121 15 303

Program Studi : STATISTIKA

Judul Skripsi : Penggunaan Metode XGBoost Untuk Klasifikasi Status Obesitas di Indonesia

“Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin”

DEWAN PENGUJI

Tanda Tangan

1. Ketua : Sri Astuti Thamrin, S.Si., M.Stat., Ph.D. (.....)

2. Sekretaris : Dr. Eng. Armin Lawi, M.Eng. (.....)

3. Anggota : Dr. Nirwan Ilyas, M.Si. (.....)

4. Anggota : Dr. Anna Islamiyati, S.Si., M.Si. (.....)

Ditetapkan di : Makassar

Tanggal : 1 November 2021

KATA PENGANTAR

Segala puji syukur senantiasa penulis panjatkan kepada Tuhan Yang Maha Esa sehingga penulisan skripsi ini dapat terselesaikan.

Puji Tuhan, skripsi dengan judul “Penggunaan Metode *XGBoost* Untuk Klasifikasi Status Obesitas Di Indonesia” yang disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin ini dapat dirampungkan. Dalam penulisan skripsi ini, penulis mampu melewati berbagai hambatan karena banyaknya dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang tak terhingga kepada orang tua penulis, Ibunda **Salfina Ropi Lantang** dan Ayahanda **Yopi Rombe** atas semua nasihat, pengorbanan, kerja keras, doa dan kasih sayang yang tulus kepada penulis.

Penghargaan dan ucapan terima kasih dengan penuh ketulusan juga penulis ucapkan kepada:

1. Ibu **Dr. Nurtiti, Sunusi, S.Si., M.Si.** selaku Ketua Departemen Statistika, yang telah membekali ilmu dan kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di jurusan Statistika.
2. Ibu **Sri Astuti Thamrin, M.Stat., Ph.D.** sebagai dosen pembimbing utama sekaligus ketua tim penguji atas ilmu yang beliau berikan selama proses perkuliahan, dan kesediaan membimbing dalam penyusunan skripsi ini.
3. Bapak **Dr.Eng. Armin Lawi, S.Si., M.Eng.** sebagai dosen pembimbing pertama sekaligus sekretaris tim penguji atas ilmu yang beliau berikan selama proses perkuliahan, dan bimbingan serta segala bentuk bantuan yang telah beliau berikan dalam penyusunan skripsi ini.
4. Bapak **Dr. Nirwan Ilyas, M.Si.** sebagai dosen penasehat akademik sekaligus ketua tim penguji atas ilmu yang beliau berikan selama proses perkuliahan dan memotivasi penulis baik diluar maupun dalam penyusunan skripsi ini.

5. Ibu **Dr. Anna Islamiyati, S.Si., M.Si.** selaku penguji yang selama seminar telah banyak memberikan kritik dan saran yang sangat berharga dalam perbaikan skripsi ke arah yang lebih baik.
6. Bapak **Andi Kresna Jaya, S.Si., M.Si.** dan Bapak **Siswanto, S.Si., M.Si.** serta **segenap dosen pengajar, staf departemen Statistika, dan staf Fakultas MIPA,** atas segala masukan, ilmu, dan kemudahan-kemudahan yang diberikan kepada penulis dalam berbagai hal selama menjadi mahasiswa di departemen Statistika.
7. Ketiga saudara penulis **Dwi, Ela** dan **Egi** serta Paman **Agus R** yang senantiasa membantu dan memberikan motivasi serta perhatian kepada penulis.
8. Sahabat-sahabatku **Masjidil Aqsha, Muhammad Fadil, Ihza Kurniawan, Ade Kurniawan, Ilham,** dan **Hilmi Abyan** atas kebersamaan, kepedulian, dan suka-duka yang telah kita lewati bersama. Semoga persahabatan kita terus terjalin dan tidak pernah usai.
9. **Liana Beatrix Larasani** atas segala bentuk bantuan yang diberikan baik berupa moril dan materil selama proses perkuliahan dan dalam penyusunan skripsi, serta motivasi yang selalu diberikan kepada penulis.
10. Keluarga besar Statistika 2015 atas segala bentuk dukungan dan bantuan selama proses perkuliahan. Terkhusus kepada teman seperjuangan dalam penyusunan skripsi, **Abdul Gafur Darussalam** dan **Muhammad Aris.**
11. Keluarga besar **Simetris 2015** terkhusus kepada **Andi Rahmat Kaswar, Nifal Gusti, Muhammad Anugerah,** dan **Ricky Risman Ali,** atas dedikasi serta semangatnya dalam menjalankan roda-roda organisasi sekaligus proses akademik.
12. Sahabatku **Rheza Paleva Parung, Stevan Tato Djimeng,** dan **Rian Rantetasak** yang selalu membantu dalam keadaan apapun.
13. **Tante Eka sekeluarga, Kakak Layuk** dan **kakak Novita Salinding sekeluarga** atas semua dukungan yang diberikan kepada penulis.
14. Seluruh pihak yang telah membantu dalam penyelesaian skripsi ini yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa masih banyak kekurangan dalam tugas akhir ini, untuk itu dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga tulisan ini memberika manfaat untuk pembaca.

Makassar, 1 November 2021

Penulis

**PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR
UNTUK KEPENTINGAN AKADEMIS**

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini:

Nama : Yoris Rombe
NIM : H121 15 303
Program Studi : Statistika
Departemen : Statistika
Fakultas : Matematika dan Ilmu Pengetahuan Alam
Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Prediktor Royalti Non-eksklusif** (*Non-exclusive Royalty-Free Right*) atas tugas akhir saya yang berjudul:

**PENGGUNAAN METODE XGBOOST UNTUK KLASIFIKASI
STATUS OBESITAS DI INDONESIA**

Beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan 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 Makassar pada tanggal 1 November 2021

Yang menyatakan

(Yoris Rombe)

ABSTRAK

Obesitas adalah suatu keadaan dimana terjadi penumpukan lemak tubuh yang berlebih, sehingga dapat membahayakan kesehatan. Beberapa faktor risiko yang menyebabkan obesitas diantaranya status perkawinan, pendapatan rumah tangga, wilayah domisili, aktivitas fisik serta asupan energi dan karbohidrat. Selain itu, faktor genetik, faktor psikologis, pola hidup yang kurang tepat, kebiasaan makan yang salah, stres dan faktor pemicu lain. Besarnya ketersediaan data serta pengetahuan di bidang medis yang terus meningkat berkontribusi pada pesatnya perkembangan di bidang ini. *Machine learning* yang tangguh dibutuhkan untuk memenuhi kebutuhan pengenalan pola data medis, termasuk data obesitas. Tujuan dari penelitian ini adalah untuk menentukan faktor-faktor yang mempengaruhi status obesitas di Indonesia. *XGBoost (Extreme Gradient Boosting)* merupakan metode klasifikasi yang sering digunakan, karena memiliki banyak keunggulan dibandingkan metode klasifikasi klasik. Algoritma *Adaptive Synthetic Nominal (ADASYN-N)* dapat digunakan untuk meningkatkan akurasi prediksi dari data yang tidak seimbang. Kedua metode itu akan diaplikasikan pada data Obesitas dari Riset Kesehatan Dasar Indonesia 2013. Hasil yang diperoleh menunjukkan bahwa jenis kelamin (perempuan) merupakan variabel atau fitur yang memiliki nilai *gain* tertinggi (37%). Hal ini menyimpulkan bahwa jenis kelamin (perempuan) memiliki pengaruh paling tinggi terhadap model dalam melakukan klasifikasi status obesitas di Indonesia.

Kata Kunci: *ADASYN-N, Feature Importance, informasi gain, obesitas, XGBoost*

ABSTRACT

Obesity is a condition due to excessive fat in the body, which can endanger health. Several risk factors that cause obesity in adult women are marital status, household income, domicile area, physical activity, energy and carbohydrate intake. Besides, genetic factors, psychological factors, improper lifestyle, bad eating habits, stress, and other trigger factors. The increasing availability of data and knowledge in the medical field has contributed to the rapid development in this field. Powerful machine learning is required to meet the pattern recognition needs of medical data, including obesity data. This study is aimed to determine the factors that influence obesity status in Indonesia. XGBoost (Extreme Gradient Boosting) is a classification method that is often used because it has many advantages over classical classification methods. Adaptive Synthetic Nominal Algorithm (ADASYN-N) can be used to improve the prediction accuracy of imbalanced data. Both methods will be applied to the Obesity data from the 2013 Indonesian Basic Health Research. The results show that gender (female) is a variable or feature that has the highest gain (37%). This concludes that gender (female) has the highest influence on the model in classifying obesity status in Indonesia.

Keywords: *ADASYN-N, Feature Importance*, information gain, obesity, *XGBoost*

DAFTAR ISI

HALAMAN SAMPUL.....	i
LEMBAR PERNYATAAN KEOTENTIKAN	ii
HALAMAN PENGESAHAN	iv
KATA PENGANTAR	v
PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS	viii
ABSTRAK.....	ix
ABSTRACT.....	x
DAFTAR ISI.....	xi
DAFTAR GAMBAR	xiv
DAFTAR TABEL.....	xv
DAFTAR LAMPIRAN.....	xvi
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan Penulisan	3
1.4 Batasan Masalah.....	3
BAB 2 TINJAUAN PUSTAKA	4
2.1 Obesitas	4
2.2 Penyaringan	4
2.3 Data Tidak Seimbang	4
2.4 <i>Adaptive Synthetic</i>	5
2.5 <i>Adaptive Synthetic Nominal</i>	8
2.6 Teknik <i>Ensemble</i>	9
2.7 <i>Boosting</i>	10
2.8 <i>Extreme Gradient Boosting</i>	11
2.8.1 Fungsi Objektif (<i>Objective Function</i>).....	11

2.8.2	<i>Decision Tree Ensemble</i>	12
2.8.3	<i>Additive Training</i>	14
2.8.4	<i>Model Complexity</i>	16
2.8.5	Nilai Struktur (<i>The Structure Score</i>).....	16
2.8.6	Tingkat Kepentingan Fitur (<i>Feature Importance</i>).....	17
2.9	Kinerja Klasifikasi.....	17
2.9.1	Matriks Konfusi (<i>Confusion Matrix</i>)	18
2.9.2	Kurva Operasi Penerima (KOP)	20
2.9.3	Luas Dibawah Kurva (LDK).....	20
BAB 3 METODOLOGI PENELITIAN		22
3.1	Sumber Data.....	22
3.2	Metode Analisis.....	24
3.2.1	Tahap Persiapan Data (<i>Filtering</i>).....	24
3.2.2	Tahap Metode Split Data	24
3.2.3	Penerapan ADASYN-N	24
3.2.4	Tahap Pemodelan.....	25
3.2.5	Evaluasi Kinerja Model	25
3.3	Diagram Alir	26
BAB 4 HASIL DAN PEMBAHASAN		27
4.1	Deskripsi Data	27
4.2	Penyaringan Data	27
4.3	Data Latih dan Data Uji.....	28
4.4	<i>XGBoost Tree</i>	28
4.5	Metode XGBoost.....	29
4.6	Penanganan Data Tidak Seimbang.....	32
4.7	Metode XGBoost dengan ADASYN-N	34
4.8	Perbandingan Model	37
4.9	Tingkat Kepentingan Fitur (<i>Feature Importance</i>).....	38
BAB V KESIMPULAN DAN SARAN		40
5.1	Kesimpulan.....	40
5.2	Saran.....	40

DAFTAR PUSTAKA	41
LAMPIRAN.....	43

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi proses <i>ensemble</i>	9
Gambar 2.2 Ilustrasi proses <i>boosting</i>	10
Gambar 2.3 Ilustrasi <i>regularization term</i>	12
Gambar 2.4 Ilustrasi CART	13
Gambar 2.5 Ilustrasi dua model <i>ensemble tree</i>	13
Gambar 3.1 Diagram Alir	27
Gambar 4.1 Plot model XGBoost Tree.....	29
Gambar 4.2 Plot Perbandingan Nilai LDK Model XGBoost.....	30
Gambar 4.3 Kinerja Klasifikasi Model XGBoost	31
Gambar 4.4 Plot KOP dan nilai LDK XGBoost	32
Gambar 4.5 Gambaran data tidak seimbang dan data setelah diseimbangkan dengan ADASYN-N pada data obesitas di Indonesia	33
Gambar 4.6 Plot Nilai LDK Model XGBoost menggunakan ADASYN-N	35
Gambar 4.7 Kinerja Klasifikasi Model XGBoost Menggunakan ADASYN-N	36
Gambar 4.8 Plot KOP XGBoost menggunakan ADASYN-N.....	37
Gambar 4.9 Perbandingan Kinerja Klasifikasi	38
Gambar 4.10 <i>Feature Importance</i> Model XGBoost	39

DAFTAR TABEL

Tabel 2.1 Matriks Konfusi	18
Tabel 2.2 Kurva Operasi Penerima	20
Tabel 2.3 Nilai Luas Dibawah Kurva	20
Tabel 3.1 Variabel Penyusun Model	22
Tabel 4.1 Gambaran umum tentang data obesitas di Indonesia.....	27
Tabel 4.2 Matriks konfusi XGBoost	30
Tabel 4.3 Gambaran data setelah penerapan ADASYN-N	34
Tabel 4.4 Matriks konfusi XGBoost menggunakan ADASYN-N.....	35

DAFTAR LAMPIRAN

Lampiran 1 Konfigurasi hyperparameter model XGBoost.....	43
Lampiran 2 Hasil pengujian model XGBoost yang dibangun dari konfigurasi hyperparameter.....	44
Lampiran 3 Hasil pengujian model XGBoost dengan ADASYN-N yang dibangun dari konfigurasi hyperparameter	46
Lampiran 4 Contoh perhitungan jarak MVDM	48
Lampiran 5 Bentuk pohon dan <i>Feature importance</i> model 09 XGBoost dengan ADASYN-N berdasarkan nilai <i>gain</i>	53
Lampiran 6 Contoh Pembentukan Model Pohon <i>XGBoost</i>	58

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Obesitas adalah suatu keadaan dimana terjadi penumpukan lemak tubuh yang berlebih, sehingga berat badan seseorang jauh diatas normal dan dapat membahayakan kesehatan. Obesitas terjadi karena ketidakseimbangan energi yang masuk dengan energi yang keluar (Dewi, 2015).

Prevalensi obesitas di negara maju dan berkembang mengalami peningkatan. Di negara maju prevalensi obesitas pada laki-laki dan perempuan pada tahun 2004 berkisar antara 23,2% di Jepang dan 66,3% di Amerika (Low dkk., 2009). Sementara di negara berkembang pada tahun 2000-2001 berkisar antara 13,4% di Indonesia sampai 72,5% di Arab Saudi (Low dkk., 2009). Hasil Riset Kesehatan Dasar (Riskesdas) tahun 2013 di Indonesia menunjukkan bahwa prevalensi obesitas berdasarkan nilai indeks massa tubuh (IMT) pada laki-laki dewasa (umur ≥ 18 tahun) sebesar 19,7% dan pada perempuan dewasa (umur ≥ 18 tahun) sebesar 32,9%.

Beberapa faktor risiko yang menyebabkan terjadi obesitas pada wanita dewasa yaitu status perkawinan, pendapatan rumah tangga, wilayah domisili, aktivitas fisik serta asupan energi dan karbohidrat (Diana dkk., 2013). Selain itu, faktor genetik, faktor psikologis, pola hidup yang kurang tepat, kebiasaan makan yang salah, stres dan faktor pemicu lain (Norton, 1996). Oleh karena itu diperlukan metode untuk menentukan faktor-faktor yang mempengaruhi status obesitas seseorang.

Seiring berkembangnya teknologi komputasi dan algoritma pembelajaran mesin (*machine learning*) berimplikasi pada pesatnya teknologi di berbagai bidang termasuk bidang penelitian medis. Teknologi *Machine Learning* (ML) memanfaatkan komputer untuk melakukan proses belajar dari data dan

menghasilkan prediksi. Prediksi dengan tingkat akurasi yang tinggi memudahkan para peneliti melakukan evaluasi atas suatu eksperimen dengan cepat dan tepat.

Machine learning yang tangguh dibutuhkan untuk memenuhi kebutuhan pengenalan pola data medis. Salah satu metode yang umum digunakan dalam ML adalah *decision tree* (DT). DT mampu mengekstrak informasi dari suatu kumpulan data menjadi pengetahuan yang intuitif dan mudah dipahami (Barros, 2011). Keunggulan *Decision Tree* dibandingkan algoritma pembelajaran lain di antaranya adalah ketahanan terhadap *noise*, biaya komputasi yang rendah untuk menghasilkan model, serta mampu menangani fitur yang berlebih (Rokach & Maimon, 2005).

Kendala *Decision Tree* dalam ketersediaan data latih dengan nilai prediksi yang lemah diatasi dengan cara *ensemble*. Metode *ensemble* merupakan algoritma pembelajaran yang dibangun dari beberapa model pengklasifikasi atau prediksi. *Ensemble* juga dapat digunakan dalam menangani permasalahan yang umum terjadi dalam algoritma pembelajaran mesin (Dietterich, 2000a), di mana nilai optimal yang dapat dicapai sebuah algoritma berada dalam rentang nilai tertentu yang dibatasi. Pada penelitian yang dilakukan oleh Dietterich (2000) disimpulkan bahwa klasifikasi dan prediksi dengan algoritma *ensemble* yang ditangani dengan baik secara umum dapat menghasilkan akurasi dan stabilitas yang lebih tinggi dari penggunaan satu algoritma saja. Cara yang umum dilakukan dalam metode *ensemble* adalah *boosting* dan *bagging*.

Konsep *ensemble* dengan *boosting* bekerja dengan cara melatih kelompok model secara sekuensial dan kemudian menggabungkan keseluruhan model tersebut untuk melakukan prediksi, model yang dihasilkan belajar dari kesalahan model sebelumnya (Zhou, 2012). Salah satu metode *boosting* yang sering digunakan adalah *gradient boosting* dimana metode ini menggunakan pendekatan *boosting* secara *gradient descent*. Mekanisme *gradient descent* dilakukan dalam rangka evaluasi untuk membuat model berikutnya. *Gradient boosting* pertama kali diperkenalkan oleh (Friedman, 2001), algoritma ini kemudian dikembangkan lebih lanjut, salah satu bentuk implementasinya adalah XGBoost atau *Extreme Gradient Boosting* oleh Chen dan Geustrin (2016). Algoritma ini merupakan perluasan dari

algoritma *Gradient Boosting Machine* (GBM) klasik dan hanya digunakan untuk data yang memiliki label dalam proses latihnya. Algoritma ini amat populer dalam kompetisi ML yang diadakan oleh Kaggle.

Extreme Gradient Boosting (XGBoost) merupakan suatu metode pada *machine learning* dimana XGBoost merupakan algoritma regresi dan klasifikasi dengan metode ensemble. XGBoost ini juga merupakan suatu varian dari algoritma *Tree Gradient Boosting* yang dikembangkan dengan optimasi 10 kali lebih cepat dibandingkan *Gradient Boosting* (Chen & Guestrin, 2016).

Oleh karena itu, XGBoost akan digunakan dalam tugas akhir ini untuk menentukan faktor-faktor yang mempengaruhi status obesitas di Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang, maka rumusan masalah penelitian ini adalah faktor-faktor apa saja yang mempengaruhi status obesitas di Indonesia dengan menggunakan metode XGBoost. Penggunaan metode XGBoost karena metode ini merupakan metode *ensemble* yang mampu mengklasifikasi variabel optimasi yang lebih cepat dibandingkan dengan metode *Boosting* lainnya.

1.3 Tujuan Penulisan

Tujuan dari penulisan tugas akhir ini berdasarkan rumusan masalah adalah untuk menentukan faktor-faktor yang mempengaruhi status obesitas di Indonesia dengan menggunakan XGBoost.

1.4 Batasan Masalah

Batasan masalah pada tugas akhir ini adalah sebagai berikut:

1. Data yang digunakan adalah data obesitas Indonesia dari Riset Kesehatan Dasar (RISKEDAS) 2013.
2. Dalam proses mengatasi ketidakseimbangan kelas data digunakan metode *oversampling ADASYN-N*.

BAB 2

TINJAUAN PUSTAKA

2.1 Obesitas

Obesitas merupakan keadaan patologis karena penimbunan lemak berlebihan yang diperlukan untuk fungsi tubuh. Penderita obesitas adalah seseorang yang timbunan lemak bawah kulitnya terlalu banyak. Perbandingan normal antara lemak tubuh dengan berat badan adalah sekitar 12-35% pada wanita dan 18-23% pada pria. Obesitas merupakan salah satu faktor resiko penyebab terjadinya penyakit degeneratif seperti diabetes mellitus, penyakit jantung koroner, dan hipertensi (Diana dkk., 2013). Obesitas berhubungan dengan pola makan, terutama bila makan makanan yang mengandung tinggi kalori, tinggi garam, dan rendah serat. Selain itu terdapat faktor lain yang mempengaruhi seperti faktor demografi, faktor sosiokultur, faktor biologi dan faktor perilaku. Obesitas juga dapat disebabkan oleh faktor genetik atau faktor keturunan.

2.2 Penyaringan

Penyaringan (*filtering*) adalah proses pemilihan subset dari *probe* yang tersedia untuk pengecualian atau penyertaan dalam analisis. Banyaknya data biasanya terdapat nilai yang tidak diketahui maknanya. *Filtering* dilakukan untuk memilih data yang mempunyai nilai atau memilih data sesuai dengan kebutuhan.

2.3 Data Tidak Seimbang

Data tidak seimbang terjadi ketika ada satu atau lebih kelas yang mendominasi keseluruhan data sebagai kelas mayor dan kelas lainnya merupakan kelas minor. Data tidak seimbang akan menghasilkan suatu akurasi prediksi klasifikasi yang baik terhadap kelas mayor, sedangkan pada kelas minor akurasi yang dihasilkan jelek. Sulitnya mendapatkan model prediksi yang baik dan bermakna pada kelas minor karena adanya ketidakcukupan informasi (Yap dkk., 2014).

Kondisi ketidakseimbangan kelas data dapat dilihat pada jumlah amatan pada kategori kelas minor dan kategori kelas mayor yang tidak proporsional, dimana kategori kelas minor memiliki jumlah amatan yang jauh lebih sedikit dibandingkan dengan kategori kelas mayor yang memiliki jumlah amatan yang jauh lebih banyak.

Beberapa alasan buruknya kinerja dari algoritma metode klasifikasi biasa yang digunakan pada data tidak seimbang adalah bahwa tujuan metode klasifikasi dalam meminimumkan galat secara keseluruhan tidak dapat tercapai karena kelas minoritas hanya sedikit memberikan kontribusi. Selain itu, asumsi yang digunakan pada metode klasifikasi biasa adalah sebaran data pada semua kelas sama atau seimbang dan diasumsikan pula bahwa galat yang berasal dari semua kelas memiliki biaya yang sama.

2.4 Adaptive Synthetic

Adaptive Synthetic (ADASYN) merupakan metode pendekatan untuk *sampling* pada kasus data yang tidak seimbang yang diajukan oleh (Haibo He dkk., 2008). Ide utama dari ADASYN adalah menggunakan bobot distribusi untuk data pada kelas minoritas berdasarkan pada tingkat kesulitan pemahamannya, sehingga data sintesis dihasilkan dari kelas minoritas yang sulit untuk dipahami dibandingkan dengan data minoritas yang lebih mudah untuk dipahami. ADASYN meningkatkan pemahaman dengan dua cara. Pertama, mengurangi bias yang diakibatkan oleh ketidakseimbangan kelas dan kedua secara adaptif menggeser batas keputusan klasifikasi terhadap kesulitan data.

Berikut prosedur menangani data tidak seimbang dengan metode ADASYN:

1. Input:

- a. Data latih D_{tr} dengan m sampel $\{x_i, y_i\}, i = 1, \dots, m$ dimana x_i adalah *sampel* dalam fitur ruang dimensi X dan $y_s \in Y = \{1, \dots, C\}$ adalah label identitas kelas dengan jumlah *sampel* terbanyak. Tentukan m_s dan m_l sebagai jumlah *sampel* kelas minoritas dan

jumlah *sampel* kelas mayoritas. Oleh karena itu, $m_{sc} \leq m_l$ dan $\sum m_{sc} + m_l = m$.

2. Prosedur

- a. Menghitung *degree of class imbalance* menggunakan persamaan berikut:

$$d_c = \frac{m_{sc}}{m_l} \quad (2.1)$$

dimana $d \in [0,1]$.

- b. Jika $d_c < d_{th}$ maka (d_{th} adalah penetapan *threshold* untuk derajat toleransi maksimum dari rasio *imbalance class*):

- i. Menghitung jumlah *sampel* data sintesis yang perlu dibangkitkan untuk kelas minoritas ke- c dengan persamaan berikut:

$$G_c = (m_l - m_{sc}) \times \beta \quad (2.2)$$

dimana $\beta \in [0,1]$ adalah parameter yang digunakan untuk menetapkan *level balance* yang diinginkan setelah pembangkitan data sintesis. $\beta = 1$ berarti data sepenuhnya seimbang setelah proses pembangkitan.

- ii. Untuk setiap *sampel* $x_i \in X_{sc}$ (X_{sc} adalah kelas minoritas ke- c), temukan *k-nearest neighbors* berdasarkan pada *Euclidean Distance* pada ruang dimensi n , dan kalkulasi rasio r_i yang didefinisikan oleh persamaan berikut:

$$r_i = \frac{\Delta_i}{K}, i = 1, \dots, m_{sc} \quad (2.3)$$

dimana Δ_i adalah jumlah *sampel* pada *nearest neighbor* yang termasuk kelas mayoritas atau termasuk semua kelas

kecuali kelas minoritas yang dievaluasi, oleh karena itu $r_i \in [0,1]$.

- iii. Normalisasi r_i dengan persamaan berikut, sehingga \hat{r}_i membentuk distribusi kerapatan (*density distribution*):

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_{sc}} r_i} \quad (2.4)$$

$$\sum \hat{r}_i = 1 \quad (2.5)$$

- iv. Menghitung jumlah dari *sampel* data sintesis yang perlu dihasilkan pada *sampel* minoritas menggunakan persamaan berikut:

$$g_i = \hat{r}_i \times G_c \quad (2.6)$$

dimana G_c adalah total jumlah *sampel* data sintesis yang perlu dihasilkan untuk kelas minoritas ke- c yang dijelaskan pada persamaan (2.6).

- v. Membangkitkan data G_i untuk setiap tetangga sekitar. Pertama, ambil sampel minoritas untuk tetangga sekitar, x_i . Kemudian, pilih secara acak sampel minoritas lain yang berada pada tetangga sekitar, x_{zi} . Sampel sintetik baru akan dihitung menggunakan:

$$s_i = x_i + (x_{zi} - x_i)\lambda \quad (2.7)$$

Pada persamaan diatas, λ adalah bilangan acak antara 0-1, s_i adalah sampel sintetik baru, x_i dan x_{zi} adalah dua sampel minoritas dalam lingkungan ketetanggan yang sama.

2.5 Adaptive Synthetic Nominal

Sebagai bentuk penyempurnaan dari Adaptive Synthetic ADASYN yang hanya diperuntukkan untuk data numerik, terdapat metode pendekatan untuk sampling pada kasus data tidak seimbang untuk data nominal yang disebut dengan *Adaptive Synthetic Nominal* (ADASYN-N) (Fithriasari dkk., 2020). Seluruh peubah yang digunakan pada penelitian ini adalah variabel kategorik, sehingga perhitungan jarak antar kelas minornya dilakukan dengan rumus Modified Value Difference Metric (MVDM) yaitu (Cost & Salzberg, 1993):

$$\Delta(x, y) = w_x w_y \sum_{i=1}^N \delta(x_{ai}, y_{bi}) \quad (2.8)$$

Dimana:

$\Delta(x, y)$: jarak antara amatan x dan y

w_x : bobot antara x (dapat diabaikan)

w_y : bobot amatan y (dapat diabaikan)

N : banyaknya variabel penjelas

$\delta(x_{ai}, y_{bi})$: jarak antara amatan x kategori a dan y kategori b pada variabel i

dengan perhitungan jarak antar amatan x kategori a dan y kategori b pada variabel ke- i dilakukan melalui formula VDM (Cost & Salzberg, 1993):

$$\delta(x_{ai}, y_{bi}) = \sum_{j=1}^S \left| \frac{c_{aj}}{c_b} - \frac{c_{bj}}{c_b} \right|^k \quad (2.9)$$

dimana:

S : banyaknya kelas pada variabel respon

c_{aj} : banyaknya kategori a pada kelas ke- j

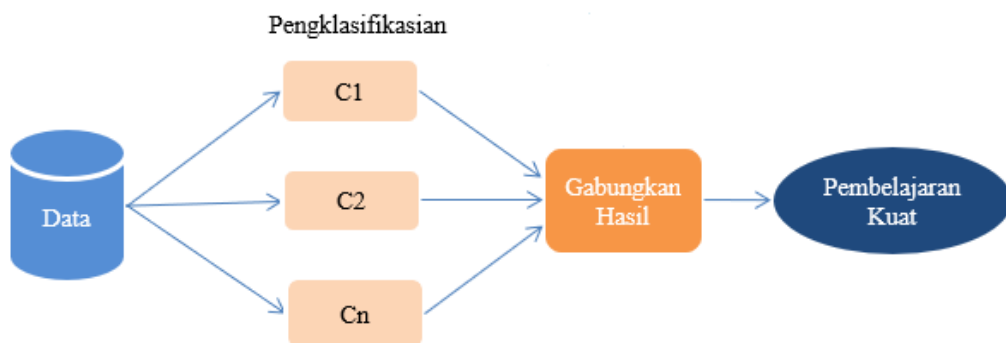
c_{bj} : banyaknya kategori b pada kelas ke- j

- C_a : banyaknya kategori a terjadi
- C_b : banyaknya kategori b terjadi
- k : konstanta (biasanya bernilai 1)

Algoritma ADASYN hampir sama dengan ADASYN-N. Perbedaannya terletak pada perhitungan jarak pada ADASYN-N menggunakan MVDM dan data baru dibangkitkan dengan melakukan duplikasi berdasarkan jumlah sampel sintetik yang akan dibangkitkan setiap sampel minoritas (Fithriasari dkk., 2020).

2.6 Teknik *Ensemble*

Teknik *ensemble* merupakan metode algoritma pembelajaran yang dibangun dari beberapa model pengklasifikasian atau prediksi untuk selanjutnya digunakan untuk mengklasifikasi data baru berdasarkan bobot prediksi yang dihasilkan sebelumnya (Dietterich, 2000b). Secara umum proses *ensemble* tergambar pada ilustrasi Gambar 2.1. Proses pengklasifikasian dilakukan dengan mengkombinasikan dengan cara-cara tertentu (umumnya berdasarkan bobot dan voting). Metode ini sudah sejak lama dikembangkan dan saat ini terus diupayakan untuk dihasilkan metode *ensemble* yang lebih baik. Ini dikarenakan metode *ensemble* dianggap dapat menghasilkan akurasi yang lebih baik dibandingkan penggunaannya hanya satu pengklasifikasi.



Gambar 2.1 Ilustrasi proses *ensemble*

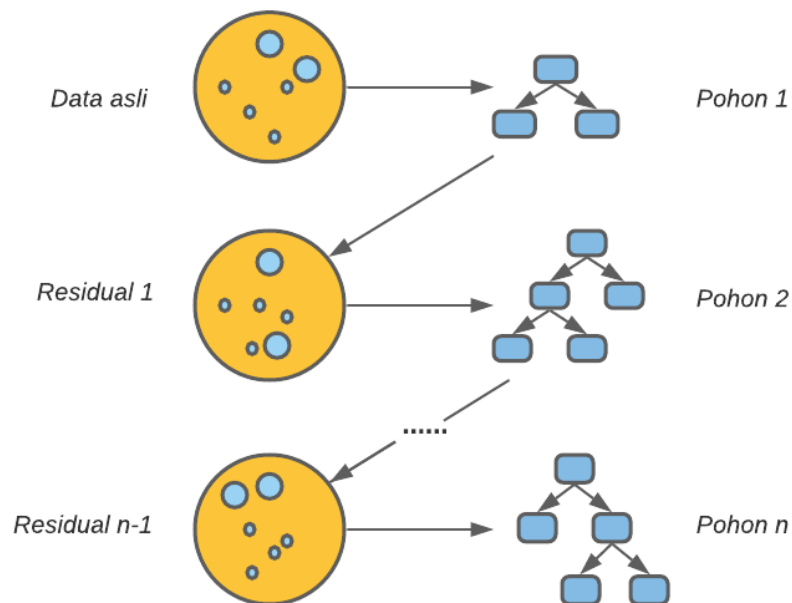
Untuk mendapatkan gabungan model terbaik dengan metode *ensemble* memungkinkan untuk menggunakan beberapa model pengklasifikasi yang berbeda

jenis. Pemilihan metode pengklasifikasi dapat ditentukan berdasarkan kebutuhan dan jenis data yang diolah. Teknik *ensemble* yang umum digunakan adalah *boosting* dan *bagging*.

2.7 Boosting

Boosting adalah teknik *ensemble* yang kuat. Teknik ini melibatkan penggabungan *weak learners* atau umumnya juga disebut *base learners*, untuk membentuk *strong learners* sehingga menghasilkan model yang lebih baik (Shakya, 2019). Berbeda dengan teknik *bagging* yang prosesnya berjalan secara paralel, teknik *boosting* melatih *weak learners* secara sekuensial dan lalu menggabungkannya. Contoh algoritma dengan *boosting* yaitu *adaboost*, *gradient boosting*, dan *XGBoost*.

Ilustrasi teknik *ensemble* dengan *boosting* dapat dilihat pada Gambar 2.2 berikut ini:



Gambar 2.2 Ilustrasi proses *boosting*

2.8 Extreme Gradient Boosting

Extreme Gradient Boosting atau *XGBoost* dikembangkan oleh (Chen & Guestrin, 2016). *XGBoost* merupakan salah satu metode *boosting* yaitu kumpulan DT yang pembangunan pohon berikutnya akan bergantung pada pohon sebelumnya. Pohon pertama dalam *XGBoost* akan lemah dalam melakukan klasifikasi dengan inisialisasi *probability* yang ditentukan oleh peneliti dan kemudian akan dilakukan *update* bobot pada setiap pohon yang dibangun sehingga menghasilkan kumpulan pohon klasifikasi yang kuat.

2.8.1 Fungsi Objektif (*Objective Function*)

Kita perlu mendefinisikan *objective function* untuk mengukur seberapa baik model tersebut sesuai dengan data latih (Hanif, 2019). Karakteristik penting dari *objective function*, terdiri dari dua bagian yaitu *training loss* dan *regularization term* seperti pada persamaan 2.10 berikut ini:

$$obj(\theta) = L(\theta) + \Omega(\theta) \quad (2.10)$$

dimana L adalah fungsi *training loss*, Ω adalah fungsi *regularization term*, dan θ adalah parameter berupa model terkait (Zhang & Zhan, 2017). *Training loss* mengukur seberapa prediktif model tersebut sehubungan dengan data latih. Fungsi *training loss* secara umum dapat ditulis seperti pada persamaan 2.11 berikut ini:

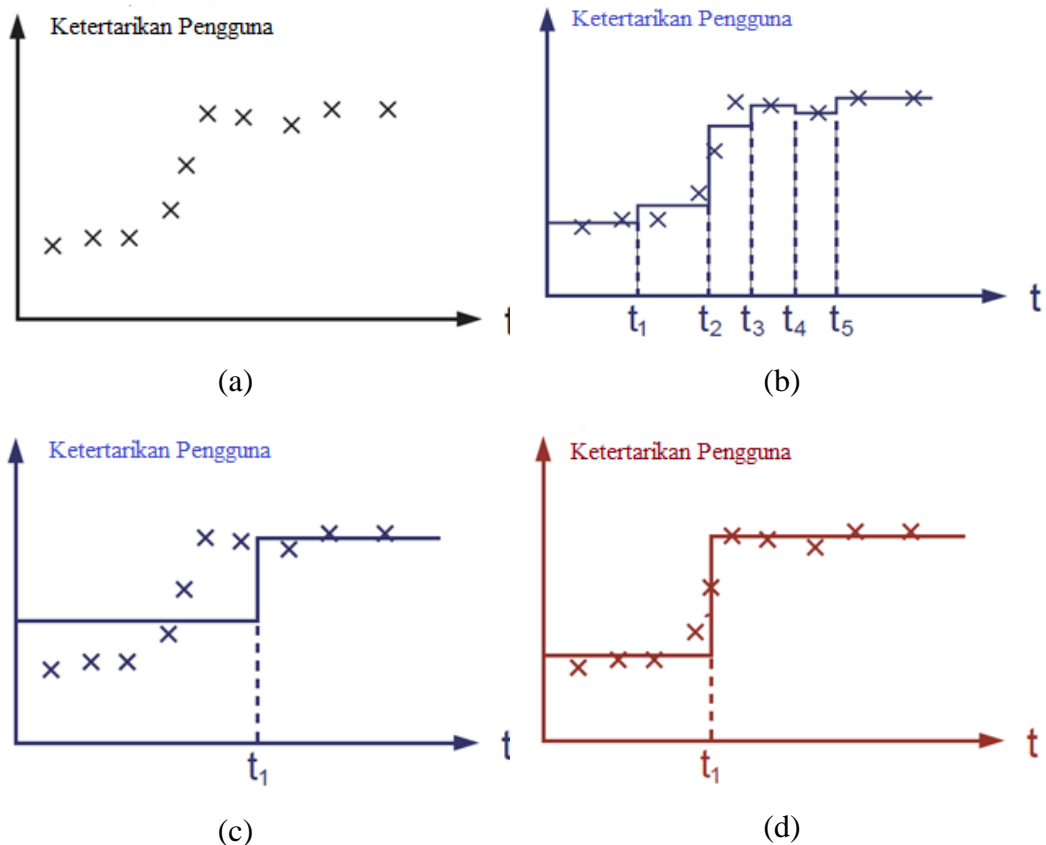
$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) \quad (2.11)$$

dimana y_i adalah nilai aktual yang dianggap benar dan \hat{y}_i adalah hasil prediksi model terkait sedangkan n adalah jumlah iterasi nilai input untuk model terkait. Formula umum yang biasa digunakan dari pengukuran *training loss* adalah *cross entropy loss* yang dapat dilihat pada persamaan 2.12 berikut ini:

$$L(\theta) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

(2.12)

Regularization term mengontrol kompleksitas model, yang membantu kita untuk menghindari *overfitting*. Gambar 2.3 merupakan ilustrasi *regularization term* dengan permasalahan keterkaitan user pada suatu topik terhadap waktu.



Gambar 2.3 Ilustrasi *regularization term*

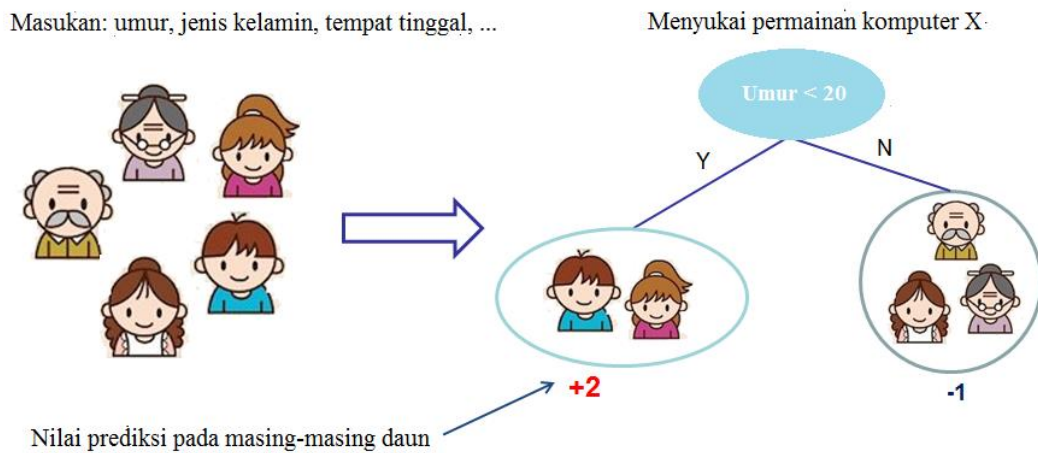
Gambar 2.3 (a) adalah output model perhitungan *user interest* dengan input berupa waktu. Gambar 2.3 (b) menunjukkan model terlalu *overfitting* sedangkan Gambar 2.3 (c) menunjukkan *training loss* yang tinggi. Gambar 2.3 (d) merupakan jawaban yang benar dimana prinsip umumnya adalah kita menginginkan model yang sederhana dan prediktif. Pertukaran antara keduanya juga disebut sebagai *bias-variance tradeoff* dalam ML.

2.8.2 Decision Tree Ensemble

Decision Tree merupakan metode pembelajaran yang melibatkan grafik seperti pohon untuk memodelkan data yang bersifat kontinu atau kategorik.

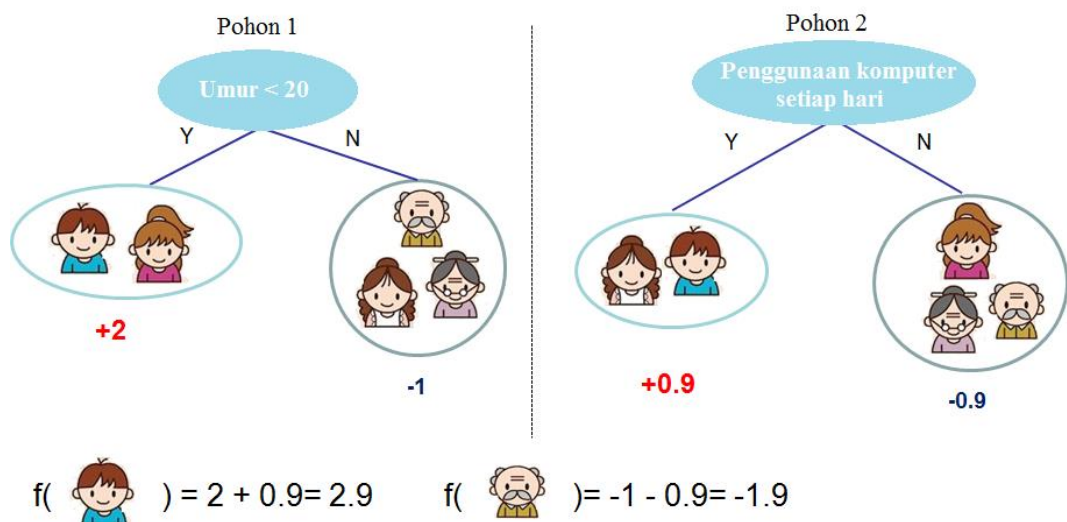
Decision Tree juga terdiri dari serangkaian pertanyaan biner yang ketika dijawab akan menghasilkan prediksi tentang data. Model *tree ensemble* terdiri dari sekumpulan *classification and regression trees* (CART).

Gambar 2.4 adalah contoh sederhana dari CART yang mengklasifikasikan apakah seseorang akan menyukai permainan komputer.



Gambar 2.4 Ilustrasi CART

Anggota keluarga diklasifikasikan menjadi *leaf* yang berbeda, dan menetapkan skor pada *leaf* yang sesuai yang dapat disebut CART (Gambar 2.5).



Gambar 2.5 Ilustrasi dua model *ensemble tree*

Skor prediksi dari masing-masing *tree* disimpulkan untuk mendapatkan skor akhir. Fakta pentingnya adalah bahwa kedua pohon berusaha saling melengkapi. Secara matematis, dapat ditulis dalam persamaan 2.13:

$$\hat{y}_i = \sum_{k=1}^t f_k(x_i), f_k \in F \quad (2.13)$$

dimana t adalah jumlah *tree*, f adalah fungsi dalam himpunan F , dan F adalah himpunan semua CART yang mungkin. *Objective function* dengan mengetahui *training loss* dan *decision tree ensemble* membentuk persamaan 2.14:

$$obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^t \Omega(f_k) \quad (2.14)$$

Dikarenakan model pada XGBoost terdiri dari sekumpulan *tree* maka fungsi *regularization term* (Ω) mengukur kompleksitas setiap *tree* berdasarkan jumlah *node* pada *tree*, *depth*, dan skor *leaf* sehingga kompleksitas model secara keseluruhan dapat diukur.

2.8.3 Additive Training

Menurut (Chen & Guestrin, 2016) sangat sulit untuk mempelajari semua *tree* sekaligus. Sebagai gantinya menggunakan strategi aditif, yaitu memperbaiki apa yang telah dipelajari dengan menambahkan satu *tree* baru sekaligus. *Objective function* yang digunakan pada strategi aditif seperti pada persamaan 2.15:

$$obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i) \quad (2.15)$$

Pada *additive training*, formula prediksi $\hat{y}_i^{(t)}$ disederhanakan seperti pada persamaan 2.16:

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (2.16)$$

dan untuk *regularization term* disederhanakan menjadi 2.17:

$$\sum_{k=1}^t \Omega(f_k) = \text{constant} + \Omega(f_t) \quad (2.17)$$

sehingga diperoleh *Objective function* seperti pada persamaan 2.18:

$$\text{obj}(t) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{constant} \quad (2.18)$$

Jika menggunakan *cross entropy loss* sebagai *loss function* maka *objective function* menjadi persamaan 2.19:

$$\text{obj}(t) = \sum_{i=1}^n -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] + \sum_{i=1}^t \Omega(f_i) \quad (2.19)$$

Dalam kasus umum, untuk mengoptimalkan nilai *objective function* digunakan *taylor expansion* dari *loss function* hingga *second order* menjadi persamaan 2.20:

$$\begin{aligned} \text{obj}(t) = \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \\ + \text{constant} \end{aligned} \quad (2.20)$$

Dengan g_i dan h_i didefinisikan sebagai persamaan 2.21 dan persamaan 2.22:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (2.21)$$

$$h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)}) \quad (2.22)$$

Setelah kita menghapus semua *constant*, *objective function* menjadi persamaan 2.23:

$$obj(t) = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (2.23)$$

Keuntungan penting dari definisi ini adalah nilai *objective function* hanya bergantung pada g_i dan h_i . Ini adalah bagaimana XGBoost mendukung *loss function*.

2.8.4 Model Complexity

Kita perlu mendefinisikan formula dari *regularization term* $\Omega(f)$. Untuk melakukannya, pertama pada *regularization term* fungsi $f_t(x)$ dapat didefinisikan sebagai persamaan 2.24:

$$f_t(x) = w_{q(x)}, w \in R^t, q: R^d \rightarrow \{1, 2, \dots, t\} \quad (2.24)$$

dimana w adalah *vector* skor dari *leaf*, q adalah fungsi yang menempatkan setiap input data ke *leaf* yang sesuai, t adalah jumlah *leaf*. Pada XGBoost kompleksitas didefinisikan sebagai persamaan 2.25:

$$\Omega(f) = \gamma t + \frac{1}{2} \lambda \sum_{j=1}^t w_j^2 \quad (2.25)$$

2.8.5 Nilai Struktur (*The Structure Score*)

Setelah merumuskan kembali model *tree*, kita dapat menulis *objective value* dengan t *tree* sebagai persamaan 2.26:

$$obj(t) \approx \sum_{i=1}^n \left[g_i w_q(x_i) + \frac{1}{2} h_i w_q^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (2.26)$$

$$= \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \quad (2.27)$$

2.8.6 Tingkat Kepentingan Fitur (*Feature Importance*)

Salah satu keuntungan menggunakan *gradient descent* adalah setelah *boosted tree* dibangun, *importance score* untuk setiap atribut mudah ditentukan. *Importance score* ini dapat digunakan untuk menunjukkan besarnya kegunaan atau pengaruh setiap fitur dalam pembuatan *boosted tree* dalam model. *Importance score* dihitung secara eksplisit untuk setiap fitur dalam kumpulan data, sehingga memungkinkan fitur diberi peringkat dan dibandingkan satu sama lain. Digunakan fungsi *xgb.importance* yang telah disediakan oleh *package xgboost* untuk menampilkan *feature importance* setiap fitur.

Salah satu parameter *xgb.importance* adalah *importance type*, yaitu parameter yang menentukan cara *importance score* dihitung. Pada metode ini digunakan *gain* yang merupakan rata-rata *gain of splits* yang menggunakan fitur tersebut dalam *tree*. *Gain* dapat didefinisikan sebagai persamaan 2.28:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (2.25)$$

2.9 Kinerja Klasifikasi

Setelah hasil prediksi diperoleh selanjutnya adalah melakukan pengukuran kinerja algoritma. Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting. Secara umum, pengukuran kinerja algoritma dilakukan dengan cara membandingkan antara hasil prediksi algoritma klasifikasi dengan nilai target variabel data latih sebagai data sebenarnya (Davis & Goadrich, 2006).

2.9.1 Matriks Konfusi (*Confusion Matrix*)

Matriks konfusi merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya matriks konfusi berisi informasi tentang klasifikasi aktual dan prediksi yang dilakukan oleh sistem klasifikasi. Kinerja dari sistem tersebut umumnya dievaluasi menggunakan data dalam matriks.

Pada pengukuran kinerja menggunakan matriks konfusi sebagaimana yang diperlihatkan pada Tabel 2.1, terdapat 4 (empat) istilah yang digunakan sebagai representasi dari proses klasifikasi. Keempat istilah tersebut adalah Positif Benar (*True Positive*) yang disingkat dengan PB, Negatif Benar (*True Negative*) yang disingkat dengan NB, Positif Salah (*False Positive*) yang disingkat dengan PS dan Negatif Salah (*False Negative*) yang disingkat dengan NS. PB merupakan data positif yang terdeteksi benar, NB merupakan data negatif yang terdeteksi salah, PS merupakan data negatif yang terdeteksi benar dan NS merupakan data positif yang terdeteksi salah (Davis & Goadrich, 2006).

Tabel 2.1 Matriks Konfusi

Matriks Konfusi		Prediksi	
		Positif	Negatif
Aktual	Positif	Positif Benar (PB)	Negatif Salah (NS)
	Negatif	Positif Salah (PS)	Negatif Benar (NB)

Positif Benar (PB) adalah banyaknya sampel dari kelas positif dan diklasifikasikan ke dalam kelas positif (*correctly classified*). Negatif Benar (NB) adalah banyaknya sampel dari kelas negatif namun diklasifikasikan ke dalam kelas positif (*incorrectly classified*). Negatif Salah (NS) adalah banyaknya sampel dari kelas positif namun diklasifikasikan ke dalam kelas negatif (*incorrectly classified*). Negatif Benar (NB) adalah banyaknya sampel dari kelas negatif dan diklasifikasikan ke dalam kelas negatif (*correctly classified*).

Berdasarkan nilai PB, NB, NB dan NS dapat diperoleh nilai metrik yang biasa dipergunakan untuk menghitung kinerja dari model klasifikasi seperti akurasi, sensitivitas, spesifikasi, dan LDK sebagai berikut (Davis & Goadrich, 2006):

a. Akurasi (*Accuracy*)

Akurasi adalah nilai ketepatan model dalam memprediksi data dengan perbandingan data aktualnya dan sebagai pengukur model untuk menentukan seberapa akurat dalam melakukan prediksi.

$$Akurasi = \frac{PB + NB}{(PB + NB + PS + NS)} \quad (2.28)$$

b. Sensitivitas (*Sensitivity*)

Sensitivitas (*Sensitivity*) adalah untuk mengevaluasi seberapa besar keberhasilan suatu model dalam memprediksi kelas positif yang diklasifikasikan. Sensitivitas didapatkan dengan menghitung perbandingan antara jumlah data untuk satu kelas tertentu yang diprediksi dengan benar dibagi jumlah total kelas tersebut.

$$Sensitivitas = \frac{PB}{PB + NS} \quad (2.29)$$

c. Spesifikasi (*Specificity*)

Spesifikasi (*specivicity*) adalah untuk mengevaluasi seberapa besar keberhasilan suatu model dalam memprediksi kelas negatif yang diklasifikasikan.

$$Spesifikasi = \frac{NB}{NB + PS} \quad (2.30)$$

2.9.2 Kurva Operasi Penerima (KOP)

Kurva Operasi Penerima (KOP) merupakan plot dari Laju Positif Benar (LPB) dibandingkan dengan Laju Positif Salah (LPS) dengan memvariasikan ambang batas metrik ini sebagaimana yang diperlihatkan pada Tabel 2.2 (Davis & Goadrich, 2006).

Tabel 2.2 Kurva Operasi Penerima

Metrik	Formula	Ekuivalent
Laju Positif Benar (LPB)	$\frac{PB}{PB + NS}$	Sensitivitas
Laju Positif Salah (LPS)	$\frac{PS}{NB + PS}$	1- Spesifikasi

2.9.3 Luas Dibawah Kurva (LDK)

Luas Dibawah Kurva (LDK) merupakan area dibawah kurva KOP yang menggambarkan probabilitas dengan variabel sensitivitas dan spesifikasi. Nilai LDK memiliki rentang antara 0.5 sampai dengan 1. Interpretasi nilai LDK dapat dilihat pada Tabel 2.3 (Gorunescu, 2011).

Tabel 2.3 Nilai Luas Dibawah Kurva

Nilai LDK	Keterangan
0.9 – 1	Tingkat akurasi sangat tinggi
0.8 – 0.9	Tingkat akurasi tinggi
0.7 – 0.8	Tingkat akurasi sedang
0.6 – 0.7	Tingkat akurasi lemah
0.5 – 0.6	Tingkat akurasi salah

Rumus LDK sebagai berikut:

$$LDK = \frac{\sum_{i=1}^{n^+} \sum_{j=1}^{n^-} 1_{f(x_i^+) > f(x_j^-)}}{n^+ n^-} \quad (2.31)$$

dimana:

- $f(\cdot)$: Nilai suatu fungsi
- x^+ : Sampel positif
- x^- : Sampel negatif
- n^+ : Jumlah sampel positif
- n^- : Jumlah sampel negatif.