

Application of Adaptive Synthetic Nominal and Extreme Gradient Boosting Methods in Determining Factors Affecting Obesity: A Case Study of Indonesian Basic Health Research Survey 2013*

Yoris Rombe¹, Sri Astuti Thamrin^{2‡}, and Armin Lawi^{3,4}

^{1,2}Department of Statistics, Universitas Hasanuddin, Indonesia

³Department of Mathematics, Universitas Hasanuddin, Indonesia

⁴Institut Teknologi Bacharuddin Jusuf Habibie, Indonesia

‡Corresponding author: tuti@unhas.ac.id

Copyright © 2022 Yoris Rombe, Sri Astuti Thamrin, dan Armin Lawi. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Obesity is the accumulation of excessive body fat and can be harmful to health. According to recent studies, several factors that contribute to the increasing prevalence of obesity in Indonesia include poor diet, lack of consumption of vegetables and fruits, high consumption of fast food, area of residence, and lack of physical activity. In addition, psychological factors, high consumption of alcohol and cigarettes, cultural differences, and stress factors also trigger obesity. The rapid development of the medical field cannot be separated from the availability of data that is increasingly easy to access and increasing knowledge in the medical field. This makes machine learning increasingly needed for pattern recognition from very large medical data, including obesity data. In this study, the factors that influence obesity status in Indonesia will be determined. In order to achieve this, Extreme Gradient Boosting (XGBoost) was used. This method is one of the classification methods that has better scalability and more efficient over its previous methods. Besides that, to overcome the imbalanced data, Adaptive Synthetic Nominal Algorithm (ADASYN-N) is used in order to balance the data and improve its prediction accuracy. Both the ADASYN-N and XGBoost methods will be applied to obesity data from the Indonesian Basic Health Research Survey in 2013. This study shows that female is more at risk in determining obesity status in Indonesia based on the highest gain value (37%). In addition, age 35-54 years, strenuous activity, and eating vegetables for 6 days are also risk factors of obesity.

Keywords: ADASYN-N, feature important, gain, obesity, XGBoost.

* Received: Jan 2021; Reviewed: Jan 2021; Published: Aug 2022

Pendahuluan

Obesitas adalah akumulasi lemak tubuh yang berlebihan dan dapat membahayakan kesehatan. Obesitas sangat terkait dengan banyak penyakit kronis yang memiliki efek negatif jangka panjang bagi penderita. Obesitas disebabkan oleh ketidakseimbangan antara asupan energi dan produksi energi dalam tubuh (Thamrin et al., 2021).

Prevalensi obesitas berdasarkan nilai indeks massa tubuh (IMT) pada pria dewasa (≥ 18 tahun) dan pada wanita dewasa (≥ 18 tahun) masing-masing sebesar 19,7% dan 32,9% berdasarkan hasil Riset Kesehatan Dasar (RISKESDAS) tahun 2013 di Indonesia. Faktor yang diketahui berkontribusi terhadap peningkatan prevalensi obesitas di Indonesia antara lain pola makan yang buruk, kurangnya konsumsi sayur dan buah, tingginya konsumsi makanan cepat saji, wilayah tempat tinggal, dan kurangnya aktivitas fisik (Oddo et al., 2019). Selain itu, faktor psikologis, tingginya konsumsi alkohol dan rokok, perbedaan budaya, dan faktor stress juga memicu obesitas (Sari & Rosha, 2016). Berdasarkan hal ini, penting untuk diidentifikasi faktor-faktor apa sajakah yang dapat memengaruhi status obesitas seseorang dengan menggunakan metode pendekatan yang lebih efisien. Salah satu pendekatan yang menjanjikan adalah dengan menerapkan teknologi *machine learning* (ML).

Machine learning (ML) adalah metode analisis data yang menerapkan sistem komputer untuk mengidentifikasi pola kompleks dalam sejumlah besar data sehingga komputer bisa menjalankan tugas tanpa diprogram secara eksplisit (Morgenstern et al., 2021). Seiring dengan perkembangan pesat teknologi komputer dan algoritma ML dalam dekade terakhir telah menghasilkan peningkatan besar dalam akurasi dan waktu pengoperasian ML. Diperkirakan bahwa ML akan semakin memainkan peran penting dalam penelitian medis (Alkhalaf et al., 2022).

Beberapa metode dalam ML telah digunakan para peneliti, diantaranya yang sering digunakan adalah pohon keputusan atau *decision tree* (DT). Pohon keputusan merupakan algoritma yang menggunakan fungsi untuk mengklasifikasikan data dalam bentuk yang mirip seperti pohon. Pohon keputusan mudah untuk diterapkan karena konsepnya yang cukup sederhana. Selain itu, pohon keputusan juga bisa digunakan untuk mengklasifikasikan data numerik dan data kategorik. Namun, pohon keputusan sangat rentan terhadap *overfitting* dan *underfitting* ketika jumlah data sedikit, akibatnya dapat membatasi generalisasi dan ketahanan model yang dihasilkan (Charbuty & Abdulazeez, 2021; Song & Lu, 2015).

Kendala penggunaan pohon keputusan dikaitkan dengan ketersediaan data dapat diatasi dengan teknik *ensemble*. Metode *ensemble* menggabungkan beberapa model pengklasifikasi untuk mengurangi *error* dan mendapatkan akurasi yang lebih baik. *Bagging* dan *boosting* merupakan teknik *ensemble* yang umum digunakan saat ini. Teknik *ensemble* dengan *bagging* bekerja dengan cara melatih model secara paralel, sedangkan teknik *ensemble* dengan *boosting* melatih model secara sekuensial (Jukic et al., 2020; Zhou, 2012). Salah satu bentuk implementasi dari *boosting* yang populer adalah *Extreme Gradient Boosting* (XGBoost) (Chen & Guestrin, 2016).

XGBoost merupakan pengembangan dari metode *Gradient Boosting*. Metode XGBoost memiliki skalabilitas yang lebih baik yang mampu melakukan optimasi yang lebih cepat daripada metode *Gradient Boosting* (Chen & Guestrin, 2016). XGBoost dapat menghasilkan kinerja model regresi dan klasifikasi dengan metode *ensemble* yang baik dengan syarat informasi pada kelas datanya seimbang. Untuk menyeimbangkan informasi pada kelas data dapat digunakan beberapa metode *sampling* misalnya metode *Synthetic Minority Over-Sampling Technique* (SMOTE) (Chawla et al., 2002; Thamrin et al., 2021), *Adaptive Synthetic* (ADASYN) (Haibo He et al., 2008), dan *Adaptive Synthetic Nominal* (ADASYN-N) (Fithriasari et al., 2020;

Rahayu et al., 2017). Pada studi ini akan digunakan metode ADASYN-N. Metode ini dapat menghasilkan data sintesis yang berasal dari kelas minoritas yang memiliki tingkat kesulitan belajar lebih tinggi dari data kelas minoritas itu sendiri. Hal ini dilakukan dengan memberikan bobot pada distribusi data di kelas minoritas. Oleh karena itu, pada studi ini akan digunakan metode ADASYN-N dan XGBoost untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap status obesitas di Indonesia berdasarkan data Survei RISKESDAS Indonesia 2013.

1. Metodologi

1.1 Sumber Data

Data yang digunakan dalam studi ini adalah data sekunder, yaitu data obesitas yang berasal dari Survei RISKESDAS Indonesia 2013. Gugus data obesitas ini tersedia di situs <https://www.litbang.kemkes.go.id/layanan-permintaan-data-riiset/> yang dapat diakses melalui proses permintaan data ke Kementerian Kesehatan. Adapun variabel yang digunakan dalam studi ini meliputi 25 variabel yang bertipe kategorik dan 69 atribut, dengan jumlah responden 722.329. Variabel status obesitas ini terdiri dari dua karakteristik yaitu obesitas dan tidak obesitas. Semua data hilang atau "NA" dikeluarkan dari analisis.

1.2 Penyaringan Data

Penyaringan (*filtering*) merupakan tahapan proses yang dilakukan untuk memilih *subset* dari *probe* yang akan disertakan dalam analisis. Penyaringan ini bertujuan untuk memilih data yang mempunyai nilai. Variabel status obesitas dan variabel hamil disaring dengan menggunakan fungsi *filter*. Setelah variabel status obesitas dan variabel hamil disaring, jumlah responden yang dianalisis yaitu 704.977.

Selanjutnya untuk seleksi variabel digunakan pendekatan *Chi-Square* (χ^2) dengan taraf nyata (α) sebesar 5%. Hasil pengujian hipotesis menunjukkan bahwa kedua kelas rata-ratanya sama, artinya tidak ada perbedaan signifikan. Dari hasil penyaringan variabel, diperoleh sebanyak 11 variabel yang paling informatif ($p\text{-value} < 0,05$).

1.3 Adaptive Synthetic Nominal

Salah satu metode *sampling* yang dapat digunakan untuk mengatasi ketidakseimbangan kelas data adalah *Adaptive Synthetic* (ADASYN). Pada metode ADASYN digunakan pembobotan distribusi untuk data pada kelas minoritas yang berbeda. Hal ini dilakukan berdasarkan pada tingkat kesulitan kelas minoritas dalam belajar. Oleh karena itu, data sintesis dapat dihasilkan dari kelas minoritas yang lebih sulit untuk belajar daripada data minoritas yang lebih mudah untuk belajar. ADASYN juga dapat mengurangi bias yang disebabkan oleh ketidakseimbangan kelas dan mengubah batas-batas keputusan klasifikasi terhadap kesulitan data secara adaptif (Haibo He et al., 2008).

Metode *Adaptive Synthetic* (ADASYN) hanya dapat digunakan pada data bertipe numerik. Pada kasus kelas data tidak seimbang yang bertipe nominal dapat digunakan metode ADASYN-N. Metode ini merupakan pengembangan dari metode ADASYN (Fithriasari et al., 2020; Rahayu et al., 2017). Pada studi ini, teknik *Modified Value Difference Metric* (MVDM) digunakan untuk menghitung jarak antar kelas minoritasnya. Teknik ini digunakan karena tipe pengukuran variabel yang digunakan pada studi ini adalah kategorik.

1.4 Extreme Gradient Boosting

Salah satu bentuk implemenstasi dari *boosting* yang berupa kumpulan DT, yang pembangunan DT berikutnya bergantung pada DT sebelumnya adalah *Extreme Gradient Boosting* (XGBoost). Sama halnya dengan *Gradient Boosting*, XGBoost menggunakan DT sebagai *base learner* dan membangun ekspansi aditif dari *objective function* untuk meminimalkan *loss function*. Namun, XGBoost memiliki skalabilitas yang lebih baik yang mampu melakukan optimasi lebih cepat daripada *Gradient Boosting* (Chen & Guestrin, 2016).

Untuk melatih model, perlu didefinisikan *objective function* untuk mengukur seberapa baik model terhadap data latih. Fungsi objektif ini memiliki dua karakteristik penting. Karakteristik tersebut yaitu *training loss* dan *reguralization term*, yang dapat dilihat pada persamaan 1.

$$obj(\theta) = L(\theta) + \Omega(\theta), \quad (1)$$

dimana L adalah *training loss function* yang mengukur seberapa prediktif model terhadap data latih dan Ω adalah *reguralization term* yang mengontrol kompleksitas model agar terhindar dari *overfitting*. Secara umum *training loss function* didefinisikan seperti pada persamaan (2) berikut ini.

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i), \quad (2)$$

dimana y_i adalah nilai aktual dan \hat{y}_i adalah hasil prediksi model. Pada studi ini, diterapkan *cross entropy loss* seperti pada persamaan (3) ini.

$$L(\theta) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \quad (3)$$

Adapun variasi *loss function* yang digunakan untuk mengontrol kompleksitas *tree* dapat dilihat pada persamaaan (4) dan (5) berikut ini.

$$L_{xgb} = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2, \quad (5)$$

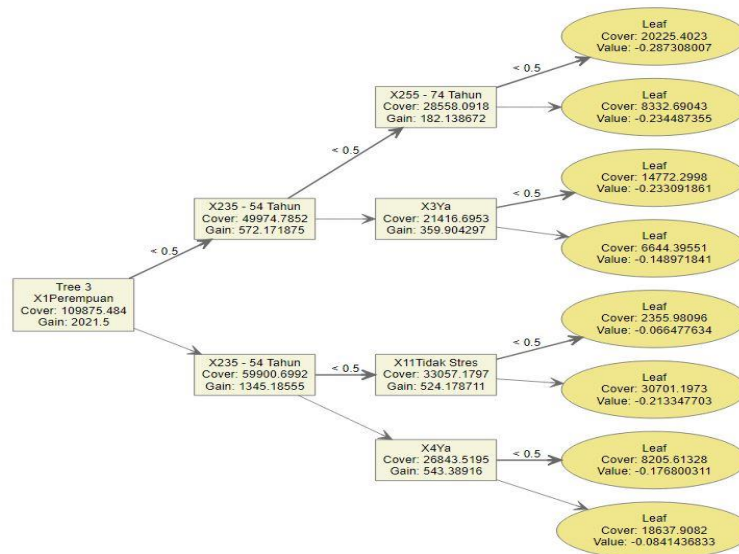
dimana T adalah jumlah *leaf* dari *tree* dan w adalah nilai output dari *leaf*. Nilai γ mengontrol nilai minimum dari *loss reduction gain* dalam membagi *node*. Akibatnya, semakin besar nilai γ maka *tree* yang terbentuk semakin sederhana.

1.5 Hasil dan Pembahasan

Pada studi ini digunakan proporsi data latih dan data uji masing-masing 80% dan 20%. Dari 80% (563.982) data latih tersebut terdapat 79.292 data di kelas obesitas dan 484.690 data di kelas tidak obesitas. Kemudian, data uji yang digunakan adalah 20% (140.995) data yang terdiri dari 19.823 data di kelas obesitas dan 121.172 data di kelas tidak obesitas.

Sebelum penerapan metode XGBoost, dilakukan pembentukan XGBoost *tree* menggunakan algoritma DT sebagai *base learner*. Parameter *n round* digunakan dalam menentukan jumlah *tree* dalam model. Selanjutnya, data latih ini, dibagi menjadi sejumlah *subset* sesuai dengan jumlah *n round* yang ditentukan. Satu *tree* akan dihasilkan dari setiap *subset* data. Pada akhirnya, ada sejumlah *n round* *tree* yang terbentuk dalam model. Ilustrasi dari salah satu *tree* yang dihasilkan dalam model

XGBoost disajikan pada Gambar 1. Berdasarkan Gambar 1, pada *tree* ke-3 fitur jenis kelamin (X_1) memiliki nilai *gain* tertinggi, sehingga menjadi *rootnode*.



Gambar 1: Model XGBoost Tree

Dalam penerapan algoritma ADASYN-N, dilakukan pembagian data terlebih dahulu dengan proporsi 80% data latih yang telah melalui proses ADASYN-N dan 20% data uji yang berasal dari data asli survei RISKESDAS 2013 di Indonesia. Sebesar 100%, data sintesis dibangkitkan. Selisih data minoritas dan data mayoritas pada data latih, yaitu 405.398. Data baru ini selanjutnya ditambahkan pada data asli sehingga berjumlah 969.380. Jumlah data minoritas sebelum dan sesudah diterapkan ADASYN-N pada data latih masing-masing 79.292 dan 484.690. Pada data hasil ADASYN-N, perbandingan antara kelas mayoritas dan kelas minoritas menjadi seimbang yaitu 50,00% pada kelas obesitas dan 50,00% pada kelas normal. Perbandingan jumlah data obesitas tanpa ADASYN-N dan dengan ADASYN-N disajikan pada Tabel 1.

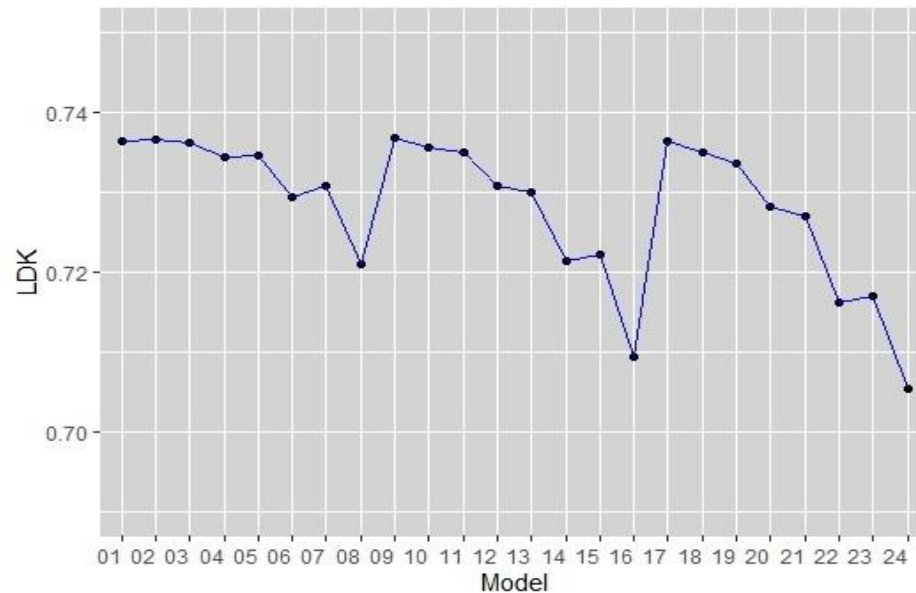
Tabel 1 menyajikan distribusi jumlah data tiap kelas obesitas sebelum digunakan ADASYN-N. Sebelum digunakan ADASYN-N, jumlah data kelas obesitas (14,05%) jauh lebih sedikit daripada data kelas tidak obesitas (85,95%). Hal ini menunjukkan bahwa data tidak seimbang. Untuk mengatasi permasalahan tersebut digunakan ADASYN-N. Dari Tabel 1 terlihat distribusi jumlah data kelas obesitas dan tidak obesitas telah seimbang.

Tabel 1: Perbandingan distribusi kelas data obesitas dan tidak obesitas tanpa ADASYN-N dan dengan ADASYN-N pada data obesitas dari RISKESDAS 2013.

Status Obesitas	Tanpa ADASYN-N (%)	ADASYN-N (%)
Obesitas	79.292 (14,05)	484.690 (50,00)
Tidak obesitas	484.690 (85,95)	484.690 (50,00)
Jumlah	563.982 (100,00)	969.380 (100,00)

Pada klasifikasi data obesitas dengan metode XGBoost, proporsi data latih yang digunakan adalah 80% (969.380) data, yang terdiri dari 484.690 data kelas obesitas dan 484.690 data kelas tidak obesitas. Kemudian, proporsi data uji yang digunakan adalah 20% (140.995) data yang terdiri dari 19.823 data kelas obesitas dan 121.172

data kelas tidak obesitas. Perbandingan nilai luas dibawah kurva (LDK) untuk masing-masing model XGBoost diperlihatkan pada Gambar 2. Model 9 memiliki nilai LDK tertinggi (73,67%) (Gambar 2). Setelah menentukan model XGBoost dengan hyperparameter terbaik, dilakukan evaluasi kinerja terhadap model tersebut. Matriks konfusi untuk ketepatan klasifikasi model dapat dilihat pada Tabel 2.



Gambar 2: Plot Nilai LDK Model XGBoost dengan ADASYN-N pada data obesitas

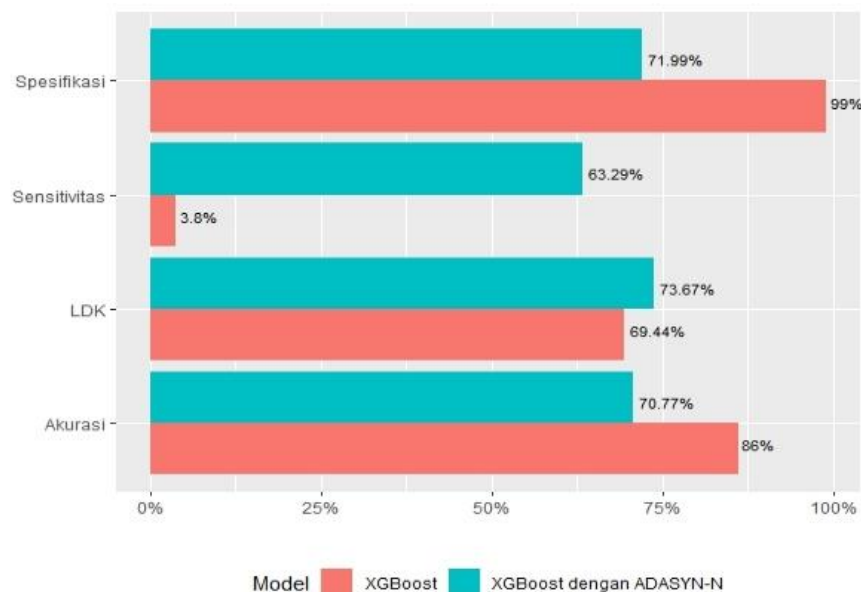
Berdasarkan Tabel 2, hasil klasifikasi metode XGBoost dari data uji kemudian dievaluasi menggunakan metrik akurasi, sensitivitas, spesififikasi, dan LDK. Perbandingan hasil evaluasi kinerja model XGBoost dengan ADASYN-N dan tanpa ADASYN-N ditunjukkan pada Gambar 3. Pada Gambar 3 itu terlihat bahwa model 9 dari XGBoost dengan ADASYN-N memiliki nilai LDK paling tinggi (73,67%). Hal ini menunjukkan bahwa kinerja yang lebih baik diberikan oleh XGBoost dengan ADASYN-N dibandingkan dengan model XGBoost tanpa ADASYN-N.

Tabel 2: Matriks konfusi XGBoost menggunakan ADASYN-N pada data obesitas

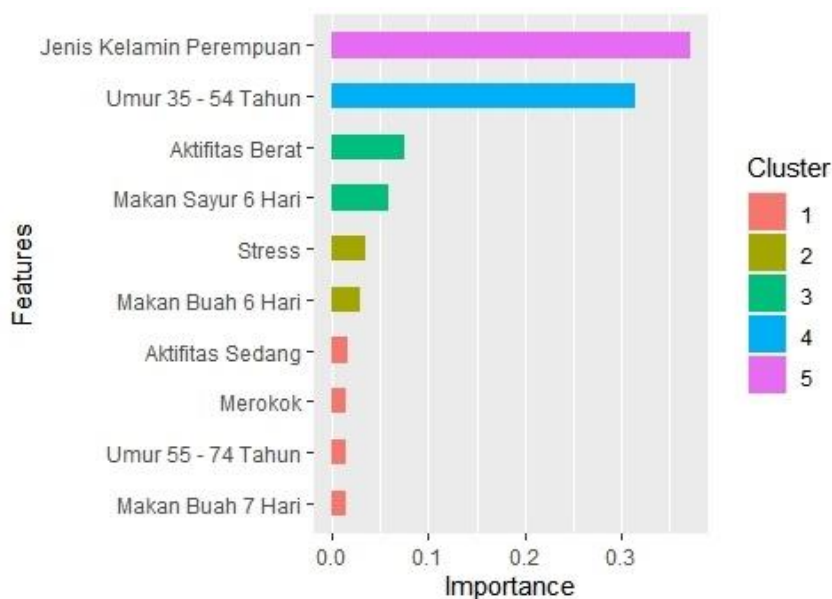
Matriks Konfusi		Prediksi	
		Obesitas	Tidak Obesitas
Aktual	Obesitas	12.546	7.277
	Tidak obesitas	33.932	87.240

Berdasarkan Gambar 3, akurasi pengklasifikasian model XGBoost tanpa ADASYN-N sebesar 86%, namun sensitivitas sangat rendah sebesar 3,8%. Hal ini menandakan model tidak mengenali kelas minoritas (obesitas) dengan baik. Kemudian, nilai LDK dari model XGBoost tanpa ADASYN-N sebesar 69,44%. Hal ini menyebabkan prediksi model hanya mengarah pada kelas mayoritas (tidak obesitas), akibatnya informasi dari kelas minoritas cenderung diabaikan karena dianggap sebagai *noise*. Oleh karena itu, status obesitas dari responden akan cenderung terklasifikasi ke kelas tidak obesitas. Untuk mengatasi masalah itu, maka digunakan metode ADASYN-N. Setelah dilakukan ADASYN-N, nilai sensitivitas mengalami peningkatan (3,8% ke 63,29%). Pada studi ini lebih difokuskan pada nilai sensitivitas, karena pada kasus ini kelas obesitas akan lebih menonjol daripada kelas tidak obesitas. Nilai LDK pada model XGBoost setelah menangani ketidakseimbangan data

menggunakan ADASYN-N juga meningkat dari 69.44% menjadi 73,67%. Dengan kata lain, XGBoost dengan ADASYN-N menghasilkan kinerja yang lebih baik dibandingkan dengan model XGBoost tanpa ADASYN-N.



Gambar 3: Perbandingan kinerja klasifikasi model XGBoost tanpa ADASYN-N dan dengan ADASYN-N pada data Obesitas dari Survei RISKESDAS 2013



Gambar Error! No text of specified style in document.: Feature Importance Model XGBoost data Obesitas dari Survei RISKESDAS 2013

Selanjutnya berdasarkan model terbaik (model 9) yang diperoleh, maka dilakukan penghitungan tingkat kepentingan fitur (*feature importance*). Hasil penghitungan tingkat kepentingan fitur ini disajikan pada Gambar 4. Dari Gambar 4 tersebut dapat diketahui bahwa gain tertinggi (0,37) dimiliki oleh fitur jenis kelamin (X1). Kemudian, fitur penting lainnya berdasarkan nilai gain ini secara berurutan adalah umur 35-54

tahun, aktivitas berat, makan sayur 6 hari, stress, makan buah 6 hari, aktivitas sedang, merokok, umur 55-74 tahun dan makan buah 7 hari.

2. Simpulan

Penentuan faktor-faktor yang memengaruhi status obesitas dari data hasil Survei RISKESDAS Indonesia 2013 telah dilakukan menggunakan metode ADASYN-N dan XGBoost. Metode ADASYN-N mampu mengatasi ketidakseimbangan kelas data obesitas. Selanjutnya, berdasarkan model XGBoost menggunakan ADASYN-N, jenis kelamin merupakan faktor risiko utama penyebab seseorang mengalami obesitas berdasarkan data hasil Survei RISKESDAS 2013. Faktor lain yang berpengaruh adalah umur 35 – 54 tahun, aktivitas berat, dan makan sayur 6 hari. Studi ini memberikan bukti bahwa dengan mengetahui faktor risiko maka pemerintah melalui instansi terkait perlu memikirkan dan mendesain kebijakan dan intervensi untuk mengatasi masalah obesitas di Indonesia. Intervensi yang menangani faktor risiko yang lebih besar secara spesifik perlu dipertimbangkan untuk dilakukan dengan disagregasi variasi faktor risiko untuk setiap wilayah.

Ucapan Terima Kasih. Penulis kedua dan sekaligus penulis korespondensi mengucapkan terima kasih kepada Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi atas bantuan dana yg diberikan untuk melakukan studi ini melalui Hibah PDUPT kontrak No. 752/UN4.22/PT.01.03/2021. Selain itu, terima kasih juga kepada Kementerian Kesehatan melalui Badan Penelitian dan Pengembangan Masyarakat yang memberikan akses ke data RISKESDAS Indonesia.

Daftar Pustaka

- Alkhalaf, M., Yu, P., Shen, J., & Deng, C. (2022). A review of the application of machine learning in adult obesity studies. *Applied Computing and Intelligence*, 2(1): 32–48. <https://doi.org/10.3934/aci.2022002>
- Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2(01): 20–28. <https://doi.org/10.38094/jastt20165>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Fithriasari, K., Hariastuti, I., & Wening, K. S. (2020). Handling Imbalance Data in Classification Model with Nominal Predictors. *International Journal of Computing Science and Applied Mathematics*, 6(1): 33. <https://doi.org/10.12962/j24775401.v6i1.6643>
- Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. <https://doi.org/10.1109/IJCNN.2008.4633969>
- Jukic, S., Saracevic, M., Subasi, A., & Kevric, J. (2020). Comparison of Ensemble Machine Learning Methods for Automated Classification of Focal and Non-Focal Epileptic EEG Signals. *Mathematics*, 8(9): 1481. <https://doi.org/10.3390/math8091481>

- Morgenstern, J. D., Rosella, L. C., Costa, A. P., de Souza, R. J., & Anderson, L. N. (2021). Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology. *Advances in Nutrition*, 12(3): 621–631. <https://doi.org/10.1093/advances/nmaa183>
- Oddo, V. M., Maehara, M., & Rah, J. H. (2019). Overweight in Indonesia: an observational study of trends and risk factors among adults and children. *BMJ Open*, 9(9): e031198. <https://doi.org/10.1136/bmjopen-2019-031198>
- Rahayu, S., Adji, T. B., & Setiawan, N. A. (2017). *Analisis Perbandingan Metode Over-Sampling Adaptive Synthetic-Nominal (ADASYN-N) dan Adaptive Synthetic-kNN (ADSYN-kNN) untuk Data dengan Fitur Nominal-Multi Categories*. 5.
- Sari, K., & Rosha, B. Ch. (2016). Several dominants risk factors related to obesity in urban childbearing age women in Indonesia. *Health Science Journal of Indonesia*, 6(1Jun): 63–68. <https://doi.org/10.22435/hsji.v6i1Jun.4494>
- Song, Y., & Lu, Y. (2015). *Decision tree methods: applications for classification and prediction*. 27(2): 7. <http://dx.doi.org/10.11919/j.issn.1002-0829.215044>
- Thamrin, S. A., Arsyad, D. S., Kuswanto, H., Lawi, A., & Nasir, S. (2021). Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. *Frontiers in Nutrition*, 8: 669155. <https://doi.org/10.3389/fnut.2021.669155>
- Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms* (0 ed.). <https://doi.org/10.1201/b12207>