OPTIMASI KLASIFIKASI IMBALANCED DATA PENYAKIT KELAINAN GENETIK MULTIFAKTORIAL DENGAN ALGORITMA ENSEMBLE

SKRIPSI

UIN SUNAN AMPEL
S U R A B A Y A

Disusun Oleh:

AHMAD BAGUS MAS'UDI

09020620018

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSTIAS ISLAM NEGERI SUNAN AMPEL
SURABAYA

2024

PERNYATAAN

PERTANGGUNG JAWABAN PENULISAN SKRIPSI

Bismillahirahmanirrahim

Yang bertanda tangan di bawah ini, saya:

Nama : Fuafatul Riza Nuriya

NIM : I71218049

Program Studi : Ilmu Politik

Yang Berjudul : Pengelolaan Dana Desa di Masa Pandemi Covid-19 di Desa Glodog Kecamatan Palang Kabupaten Tuban.

Menyatakan dengan sesungguhnya bahwa:

1. Skripsi ini tidak pernah dikumpulkan pada lembaga pendidikan manapun untuk mendapatkan gelar akademik apapun.
2. Skripsi ini adalah benar- benar hasil karya saya secara mandiri dan bukan merupakan plagiasi dari karya orang lain, kecualiy yang secara tertulis dikutip dalam naskah ini disebutkan dalam sumber kutipan dan daftar pustaka.
3. Apabila skripsi ini dikemudian hari terbukti atau dapat dibuktikan sebagai plagiasi, saya bersedia menanggung segala konsekuensi hukum yang terjadi.

Surabaya, 03 Agustus 2022

Yang menyatakan

Fuafatul Riza Nuriya

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi oleh

NAMA : AHMAD BAGUS MAS'UDI

NIM : 09020620018

JUDUL : OPTIMASI KLASIFIKASI IMBALANCED DATA PENYAKIT KELAINAN GENETIK MULTIFAKTORIAL DENGAN ALGORITMA ENSEMBLE

Ini telah diperiksa dan disetujui untuk diujikan.

Surabaya, 06 November 2023

Dosen Pembimbing 1	Dosen Pembimbing 2
Dwi Rolliawati, M.T	**Khalid, M.Kom**
NIP. 197909272014032001	NIP. 197906092014031002

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Ahmad Bagus Mas'udi ini telah dipertahankan
Di depan tim penguji skripsi
Di Surabaya, 04 Januari 2024

Mengesahkan,
Dewan Penguji

Penguji I	Penguji II
(Bayu Adhi Nugroho, Ph.D.)	(Subhan Nooriansyah, M.Kom.)
NIP. 197905182014031001	NIP. 199012282020121010
Penguji III	**Penguji IV**
(Dwi Rolliawati, M.T)	(Khalid, M. Kom)
NIP. 197909272014032001	NIP. 197906092014031002

Mengetahui,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

(Dr. Hj. Evi Hamdani, M.Pd)
NIP. 196507312000031002

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

UIN SUNAN AMPEL
S U R A B A Y A

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama	: AHMAD BAGUS MAS'UDI
NIM	: 09020620018
Fakultas/Jurusan	: SAINS DAN TEKNOLOGI / SISTEM INFORMASI
E-mail address	: ahmadbagusmasudi@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
■ Sekripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (.....)
yang berjudul :
OPTIMASI KLASIFIKASI IMBALANCED DATA PENYAKIT KELAINAN GENETIK MULTIFAKTORIAL DENGAN ALGORITMA ENSEMBLE

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 31 Januari 2024

Penulis

(AHMAD BAGUS MAS'UDI)

ABSTRAK

OPTIMASI KLASIFIKASI IMBALANCED DATA PENYAKIT KELAINAN GENETIK MULTIFAKTORIAL DENGAN ALGORITMA ENSEMBLE

Oleh:

Ahmad Bagus Mas'udi

Ketidakseimbangan data adalah masalah yang sering terjadi dalam klasifikasi. Ketidakseimbangan data terjadi ketika jumlah sampel pada salah satu kelas, jumlahnya jauh lebih sedikit atau jauh lebih besar dibanding kelas lain. Hal ini dapat menyebabkan model klasifikasi menjadi tidak akurat dan cenderung memprediksi kelas mayoritas. Untuk mengatasi masalah ketidakseimbangan data, metode resampling dapat digunakan. Sehingga penelitian ini bertujuan mengatasi permasalahan ketidakseimbangan data dalam klasifikasi data kelainan genetik multifaktorial. Penelitian ini menguji penerapan beberapa metode resampling, antara lain Random Undersampling, NearMiss, SMOTE, ADASYN, dan SMOTE-ENN, untuk meningkatkan kinerja model pada dataset yang tidak seimbang. Selain itu, diterapkan Principal Component Analysis (PCA) untuk seleksi fitur. Penerapan metode resampling tersebut diuji menggunakan algoritma klasifikasi Random Forest dan XGBoost. Dari hasil ujicoba, hasil terbaik pengujian didapatkan dalam penerapan SMOTE-ENN pada algoritma Random Forest dengan hasil accuracy sebesar 98%, spesificity sebesar 98.9%, sensitivity sebesar 94.9%, dan AUC sebesar 96.9%.

Kata Kunci: Imbalanced Data, SMOTE-ENN, Multifactorial genetic disorders, Klasifikasi.

ABSTRACT

OPTIMISATION OF IMBALANCED CLASSIFICATION OF MULTIFACTORIAL GENETIC DISORDER DISEASE DATA WITH ENSEMBLE ALGORITHM

By:

Ahmad Bagus Mas'udi

Data imbalance is a common problem in classification. Data imbalance occurs when the number of samples in one class is much smaller or much larger than the other class. This can cause the classification model to be inaccurate and tend to predict the majority class. To overcome the problem of data imbalance, resampling methods can be used. So this research aims to overcome the problem of data imbalance in the classification of multifactorial genetic disorder data. This research examines the application of several resampling methods, including Random Undersampling, NearMiss, SMOTE, ADASYN, and SMOTE-ENN, to improve model performance on imbalanced datasets. In addition, Principal Component Analysis (PCA) is applied for feature selection. The application of the resampling method was tested using Random Forest and XGBoost classification algorithms. From the test results, the best test results were obtained in applying SMOTE-ENN to the Random Forest algorithm with an accuracy of 98%, specificity of 98.9%, sensitivity of 94.9%, and AUC of 96.9%.

Keywords: Imbalanced Data, SMOTE-ENN, Multifactorial genetic disorders, Classification.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR PUSTAKA

Akter, S., Das, D., Haque, R. U., Quadery Tonmoy, M. I., Hasan, M. R., Mahjabeen, S., & Ahmed, M. (2022). AD-CovNet: An exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with COVID-19. Computers in Biology and Medicine, 146, 105657. https://doi.org/10.1016/j.compbiomed.2022.105657

Alamsyah, A. R. B., Anisa, S. R., Belinda, N. S., & Setiawan, A. (2022). SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data. Proceedings of The International Conference on Data Science and Official Statistics, 2021(1), 305–314. https://doi.org/10.34123/icdsos.v2021i1.240

Alzheimer. (2019, April 22). Statistik tentang Demensia - Alzheimer Indonesia. . https://alzi.or.id/statistik-tentang-demensia/

Ambarwari, A., Jafar Adrian, Q., & Herdiyeni, Y. (2020). Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4(1), 117–122. https://doi.org/10.29207/resti.v4i1.1517

Amei, W., Huailin, D., Qingfeng, W., & Ling, L. (2011). A Survey of Application-Level Protocol Identification Based on Machine Learning. 2011 International Conference on Information Management, Innovation Management and Industrial Engineering, 201–204. https://doi.org/10.1109/ICIII.2011.331

Andriansyah, D.-, & Eka Wulansari Fridayanthie. (2023). Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance. JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING, 6(2), 484–493. https://doi.org/10.31289/jite.v6i2.8373

Ankit, C. (2021, February 24). Random Forest Classifier and its Hyperparameters. Medium. https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

Annur, M. C. (2022, October 11). Kanker Payudara, Penyakit Kanker Paling Banyak Dialami Masyarakat Indonesia | Databoks. Databooks. https://databoks.katadata.co.id/datapublish/2022/10/11/kanker-payudara-penyakit-kanker-paling-banyak-dialami-masyarakat-indonesia

Anshul Saini. (2023, August 2). Gradient Boosting Algorithm: A Complete Guide for Beginners. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

Apriliah, W., Kurniawan, I., Baydhowi, M., Haryati, T., Informasi Kampus Kabupaten Karawang, S., Teknik dan Informatika, F., Bina Sarana Informatika, U., Banten No, J., & Karawang Barat, K. (2021). SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest (Vol. 10, Issue 1). http://sistemasi.ftik.unisi.ac.id

Attenberg, J., & Ertekin, Ş. (2013). Class Imbalance and Active Learning. In Imbalanced Learning (pp. 101–149). Wiley. https://doi.org/10.1002/9781118646106.ch6

Awalliantoni. (n.d.). Pengaruh Kelainan Gen ALDH pada Intoleransi terhadap alkohol.

Board, F. S. (2017). Artificial intelligence and machine learning in financial services Market developments and financial stability implications. www.fsb.org/emailalert

Breiman, L. (2001). Random forests. Machine learning. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2016). Master Machine Learning Algorithms: discover how they work and implement them from scratch.

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University - Computer and Information Sciences, 34(7), 4514–4523. https://doi.org/10.1016/j.jksuci.2020.10.013

Cancer. (2022, February 3). Cancer. WHO. https://www.who.int/news-room/fact-sheets/detail/cancer

Chawla, N. V. , B. K. W. , & H. L. O. (2002). Handling Imbalance Data Prediksi Churn menggunakan metode SMOTE dan KNN Based on Kernel. E-Proceeding of Engineering, 4(117), 1–15.

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. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Darujati, C., & Bimo Gumelar, A. (2012). PEMANFAATAN TEKNIK SUPERVISED UNTUK KLASIFIKASI TEKS BAHASA INDONESIA (Vol. 16, Issue 1).

Das, S., & Nene, M. J. (2017). A survey on types of machine learning techniques in intrusion prevention systems. 2017 International Conference on Wireless

Communications, Signal Processing and Networking (WiSPNET), 2296–2299. https://doi.org/10.1109/WiSPNET.2017.8300169

Dementia. (2023, March 15). WHO. https://www.who.int/news-room/fact-sheets/detail/dementia

Diabetes. (2022, April 5). Diabetes. WHO. https://www.who.int/news-room/fact-sheets/detail/diabetes

Diniari B. E. (2018, July 3). Mengenal Gen, DNA,RNA, dan Kromosom | Biologi Kelas 12. Ruangguru.

Edu C. (2022). Multifactorial Inheritance and Birth Defects – *Children's Hospital* of Philadelphia. Chop Edu. https://www.chop.edu/conditions-diseases/multifactorial-inheritance-and-birth-defects

Emerging Risk Factors Collaboration. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. The Lancet, 375(9733), 2215–2222.

Fidler, D. J. (2005). The Emerging Down Syndrome Behavioral Phenotype in Early Childhood Implications for Practice (Vol. 18, Issue 2).

Friedman, J. H. (2001). 999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE 1. In The Annals of Statistics (Vol. 29, Issue 5).

G, S. G. C., & -, B. S. (2020). Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction. International Journal of Advanced Computer Science and Applications, 11(9). https://doi.org/10.14569/IJACSA.2020.0110920

Garnia, E., Rizal Riyadi, D., Akbar Pasuma, R., Sri Astuti, P., & Sangga Buana Bandung, U. (2023). Dominant Macro Factors on the Performance of Sharia Stocks in the Period of Two Presidencies in Indonesia. Indonesian Journal of Banking and Financial Technology (FINTECH), 1(1), 25–44. https://doi.org/10.55927/fintech.v1i1.2715

Ghazal, T. M., Al Hamadi, H., Umar Nasir, M., Atta-Ur-Rahman, Gollapalli, M., Zubair, M., Adnan Khan, M., & Yeob Yeun, C. (2022). Supervised Machine Learning Empowered Multifactorial Genetic Inheritance Disorder Prediction. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/1051388

Goldberg, D. E. (1986). The Genetic Algorithm Approach: Why, How, and What Next? In Adaptive and Learning Systems (pp. 247–253). Springer US. https://doi.org/10.1007/978-1-4757-1895-9_17

Gorunescu, F. (2011). Data mining: concepts and techniques. Springer.

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

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

Han, J. , & K. M. (2006). Concepts and Techniques. In Data Mining (2nd ed.). Diane Cerra.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. Neurocomputing, 70(1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Husin, N. (2023). Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN). *Jurnal Esensi Infokom : Jurnal Esensi Sistem Informasi Dan Sistem* Komputer, 7(1), 75–84. https://doi.org/10.55886/infokom.v7i1.608

Indrawati, A., Subagyo, H., Sihombing, A., & Wagiyah, S. A. (2017). Analyzing the Impact of Resampling Method for Imbalanced Data Text in Indonesian Scientific Articles Categorization. Education.

Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. Decision Analytics, 2(1), 1. https://doi.org/10.1186/s40165-014-0010-2

Karn, A. L., Romero, C. A. T., Sengan, S., Mehbodniya, A., Webber, J. L., Pustokhin, D. A., & Wende, F.-D. (2022). Fuzzy and SVM Based Classification Model to Classify Spectral Objects in Sloan Digital Sky. IEEE Access, 10, 101276–101291. https://doi.org/10.1109/ACCESS.2022.3207480

Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 2(10), 3003–3008.

Kaur, P., & Gosain, A. (2018). Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise (pp. 23–30). https://doi.org/10.1007/978-981-10-6602-3_3

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review . GESTS International Transactions on Computer Science and Engineering, 30.

Kuhn, M. (2008). Building Predictive Models in R Using the **caret** Package. Journal of Statistical Software, 28(5). https://doi.org/10.18637/jss.v028.i05

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. IOP Conference Series: Materials Science and Engineering, 1099(1), 012077. https://doi.org/10.1088/1757-899X/1099/1/012077

Lakshmi, J. V. N., & Sheshasaayee, A. (2015). Machine learning approaches on map reduce for Big Data analytics. 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 480–484. https://doi.org/10.1109/ICGCIoT.2015.7380512

Li, S., & Zhang, X. (2020). Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. Neural Computing and Applications, 32(7), 1971–1979. https://doi.org/10.1007/s00521-019-04378-4

Ma'rifat, T. N., Purwanto, S., & Windarwati, S. (2017). PERCEPTION ON HALAL TRACEABILITY ON CHICKEN MEAT SUPPLY CHAIN. Agroindustrial Technology Journal, 01, 33–41. https://doi.org/10.21111/atj.v1i1.1838

Melissa, I., & Oetama, R. S. (2013). Analisis Data Pembayaran Kredit Nasabah Bank Menggunakan Metode Data Mining. Jurnal ULTIMA InfoSys, 4(1), 18–27. https://doi.org/10.31937/si.v4i1.238

Meng, H., Wang, S., Gao, C., & Liu, F. (2021). Research on Recognition Method of Railway Perimeter Intrusions Based on Φ-OTDR Optical Fiber Sensing Technology. IEEE Sensors Journal, 21(8), 9852–9859. https://doi.org/10.1109/JSEN.2020.3043193

Mqadi, N. M., Naicker, N., & Adeliyi, T. (2021). Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss. Mathematical Problems in Engineering, 2021, 1–16. https://doi.org/10.1155/2021/7194728

Nababan, C. E., & Simamora, E. (2023). Bootstrap Estimation of Confidence Intervals of Multiple Regression Model Parameters in the Presence of Multicollinearity Using Principal Component Analysis. Formosa Journal of Applied Sciences, 2(1), 185–202. https://doi.org/10.55927/fjas.v2i1.2851

Nawang Wulan, K., Ramkita, N., & M. Muhartono. (n.d.). Sarang Semut (Myrmecodia pendans) sebagai Antikanker.

Nuklianggraita, T. N., Adiwijaya, A., & Aditsania, A. (2020). On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier. JURNAL INFOTEL, 12(3), 89–96. https://doi.org/10.20895/infotel.v12i3.485

Nurhopipah, A., & Magnolia, C. (2022). PERBANDINGAN METODE RESAMPLING PADA IMBALANCED DATASET UNTUK KLASIFIKASI KOMENTAR PROGRAM MBKM. Jurnal Publikasi Ilmu Komputer Dan Multimedia, 1(2), 10–22.

Nurmasani, A., & Pristyanto, Y. (2021). ALGORITME STACKING UNTUK KLASIFIKASI PENYAKIT JANTUNG PADA DATASET IMBALANCED CLASS. In Jurnal Pseudocode (Vol. 1). www.ejournal.unib.ac.id/index.php/pseudocode

Nussbaumer, S., Bonnabry, P., Veuthey, J. L., & Fleury-Souverain, S. (2011). Analysis of anticancer drugs: A review. In Talanta (Vol. 85, Issue 5, pp. 2265–2289). Elsevier B.V. https://doi.org/10.1016/j.talanta.2011.08.034

Nyoman, N., Pinata, P., Sukarsa, M., Kadek, N., & Rusjayanthi, D. (n.d.). Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python.

Pahlevi, R. (2021, November 22). Jumlah Penderita Diabeter Indonesia Terbesar Kelima di Dunia | Databoks. Databooks. https://databoks.katadata.co.id/datapublish/2021/11/22/jumlah-penderita-diabetes-indonesia-terbesar-kelima-di-dunia

Prasetya, J. (2022). Penerapan Klasifikasi Naive Bayes dengan Algoritma Random Oversampling dan Random Undersampling pada Data Tidak Seimbang Cervical Cancer Risk Factors. Leibniz: Jurnal Matematika, 2(2), 11–22. https://doi.org/10.59632/leibniz.v2i2.173

Pristyanto, Y. (2019). PENERAPAN METODE ENSEMBLE UNTUK MENINGKATKAN KINERJA ALGORITME KLASIFIKASI PADA IMBALANCED DATASET. In Jurnal TEKNOINFO (Vol. 13, Issue 1). https://archive.ics.uci.edu/ml/datasets/User+Knowledge

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015). Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. 2015 IEEE International Conference on Information Reuse and Integration, 197–202. https://doi.org/10.1109/IRI.2015.39

Puri, A., & Kumar Gupta, M. (2022). Improved Hybrid Bag-Boost Ensemble With K-Means-SMOTE–ENN Technique for Handling Noisy Class Imbalanced Data. The Computer Journal, 65(1), 124–138. https://doi.org/10.1093/comjnl/bxab039

Rahman, R. M., & Afroz, F. (2013). Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. Journal of Software Engineering and Applications, 06(03), 85–97. https://doi.org/10.4236/jsea.2013.63013

Ramadhan, N. G. (2021). Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus. Scientific

Journal of Informatics, 8(2), 276–282. https://doi.org/10.15294/sji.v8i2.32484

Ramadhanti, D. V., Santoso, R., & Widiharih, T. (2023a). PERBANDINGAN SMOTE DAN ADASYN PADA DATA IMBALANCE UNTUK KLASIFIKASI RUMAH TANGGA MISKIN DI KABUPATEN TEMANGGUNG DENGAN ALGORITMA K-NEAREST NEIGHBOR. Jurnal Gaussian, 11(4), 499–505. https://doi.org/10.14710/j.gauss.11.4.499-505

Ramadhanti, D. V., Santoso, R., & Widiharih, T. (2023b). PERBANDINGAN SMOTE DAN ADASYN PADA DATA IMBALANCE UNTUK KLASIFIKASI RUMAH TANGGA MISKIN DI KABUPATEN TEMANGGUNG DENGAN ALGORITMA K-NEAREST NEIGHBOR. Jurnal Gaussian, 11(4), 499–505. https://doi.org/10.14710/j.gauss.11.4.499-505

Rhomadhona, H., & Permadi, J. (2019). Klasifikasi Berita Kriminal Menggunakan NaÃ¯ve Bayes Classifier (NBC) dengan Pengujian K-Fold Cross Validation. Jurnal Sains Dan Informatika, 5(2), 108–117. https://doi.org/10.34128/jsi.v5i2.177

Rodríguez-Gómez, F., Romero-Gil, V., Bautista-Gallego, J., Garrido-Fernández, A., & Arroyo-López, F. N. (2012). Multivariate analysis to discriminate yeast strains with technological applications in table olive processing. World Journal of Microbiology and Biotechnology, 28(4), 1761–1770. https://doi.org/10.1007/s11274-011-0990-1

Rujito, L. (2010). Rujito, L. (2010). Konseling genetik, strategi mengontrol penyakit genetik di Indonesia. Mandala of Health, 4(1).

Sasada, T., Liu, Z., Baba, T., Hatano, K., & Kimura, Y. (2020). A Resampling Method for Imbalanced Datasets Considering Noise and Overlap. Procedia Computer Science, 176, 420–429. https://doi.org/10.1016/j.procs.2020.08.043

Scikit Learn. (n.d.). Standard Scaler. Retrieved January 9, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Setyohadi, D. B., Kristiawan, F. A., & Ernawati. (2017). PERBAIKAN PERFORMANSI KLASIFIKASI DENGAN PREPROCESSING ITERATIVE PARTITIONING FILTER ALGORITHM. *TELEMATIKA*, *14*(1), 12–20.

Shaikh, S., Daudpota, S. M., Imran, A. S., & Kastrati, Z. (2021). Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models. *Applied Sciences*, *11*(2), 869. https://doi.org/10.3390/app11020869

Sole. (2023, March 22). *The Role of Undersampling in Tackling Imbalanced Datasets in Machine Learning*. Train In Data.

Solomon, D. D., Sonia, Kumar, K., Kanwar, K., Iyer, S., & Kumar, M. (2023a). Extensive Review on the Role of Machine Learning for Multifactorial Genetic Disorders Prediction. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-023-09996-9

Solomon, D. D., Sonia, Kumar, K., Kanwar, K., Iyer, S., & Kumar, M. (2023b). Extensive Review on the Role of Machine Learning for Multifactorial Genetic Disorders Prediction. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-023-09996-9

Solomon, D. D., Sonia, Kumar, K., Kanwar, K., Iyer, S., & Kumar, M. (2023c). Extensive Review on the Role of Machine Learning for Multifactorial Genetic Disorders Prediction. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-023-09996-9

Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016). A review of machine learning techniques using decision tree and support vector machine. *2016 International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–7. https://doi.org/10.1109/ICCUBEA.2016.7860040

Soomro NI, Qureshi NA, & Bakhtiar SM. (2022). *Precision medicine for multifactorial disorders precision medicine and multifactorial diseases view project probiotic biopolymeric based coating to control food borne pathogens view project*.

Sunata, H. (2020). Komparasi Tujuh Algoritma Identifikasi Fraud ATM Pada PT. Bank Central Asia Tbk. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *7*(3), 441–450. https://doi.org/10.35957/jatisi.v7i3.471

Susanti, Z., Sirait, P., & Panjaitan, E. S. (2023). Peningkatan Kinerja Random Forest Melalui Seleksi Fitur Secara Pca Untuk Mendeteksi Penyakit Diabetes Tahap Awal. *Jurnal Sains Dan Teknologi*, *4*(3), 51–56. https://doi.org/10.55338/saintek.v5i1.1093

Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications*, *36*(4), 8610–8615. https://doi.org/10.1016/j.eswa.2008.10.032

Thupae, R., Isong, B., Gasela, N., & Abu-Mahfouz, A. M. (2018). Machine Learning Techniques for Traffic Identification and Classifiacation in SDWSN: A Survey. *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 4645–4650. https://doi.org/10.1109/IECON.2018.8591178

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241–266. https://doi.org/10.14257/ijbsbt.2013.5.5.25

Trivusi. (2023, March 12). *Gradient Boosting: Pengertian, Cara Kerja, dan Kegunaannya*. Trivusi. https://www.trivusi.web.id/2023/03/algoritma-gradient-boosting.html?m=1

Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. *Proceedings of the 24th International Conference on Machine Learning*, 935–942. https://doi.org/10.1145/1273496.1273614

Vercellis, C. (2009). *Business Intelligence :Data Mining and Optimization for Decision Making*. John Wiley & Sons, Ltd.

Wang, H., & Liu, X. (2021). Undersampling bankruptcy prediction: Taiwan bankruptcy data. *PLOS ONE*, *16*(7), e0254030. https://doi.org/10.1371/journal.pone.0254030

XGBoost. (n.d.). XGBoost. Retrieved January 9, 2024, from https://xgboost.readthedocs.io/en/stable/parameter.html

Yusa, M., Utami, E., & Luthfi, E. T. (2016). Analisis Komparatif Evaluasi Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes. *Jurnal Buana Informatika*, *7*(4). https://doi.org/10.24002/jbi.v7i4.770

Zidek, K., Pitel, J., & Hosovsky, A. (2017). Machine learning algorithms implementation into embedded systems with web application user interface. *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, 000077–000082. https://doi.org/10.1109/INES.2017.8118532