KOMBINASI METODE IMPUTASI *MEAN* DAN *MULTIPLE IMPUTATION BY CHAINED EQUATIONS* (MICE) UNTUK PENANGANAN DATA HILANG DAN PENINGKATAN EVALUASI KINERJA KLASIFIKASI PREDIKSI PENYAKIT DIABETES MELITUS

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
Sarjana Sains Bidang Studi Matematika

Oleh:
YULFITA TASYA
08011281924041

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SRIWIJAYA
2023

LEMBAR PENGESAHAN

KOMBINASI METODE IMPUTASI *MEAN* DAN *MULTIPLE IMPUTATION BY CHAINED EQUATIONS* (MICE) UNTUK PENANGANAN DATA HILANG DAN PENINGKATAN EVALUASI KINERJA KLASIFIKASI PREDIKSI PENYAKIT DIABETES MELITUS

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
Sarjana Sains Bidang Studi Matematika

Oleh

YULFITA TASYA
NIM.08011281924041

Indralaya, Januari 2023

Pembimbing Kedua

Dr. Yuli Andriani, S.Si., M.Si
NIP.197207021999032001

Pembimbing Utama

Dr. Anita Desiani, S.Si., M.Kom
NIP. 19771211 2003122002

Mengetahui,
Ketua Jurusan Matematika

Drs. Sugandi Yahdin, M.M
NIP. 19580727 198603 1003

PERNYATAAN KEASLIAN KARYA ILMIAH

Yang bertanda tangan di bawah ini:

Nama Mahasiswa	: Yulfita Tasya
NIM	: 08011281924041
Fakultas/Jurusan	: Matematika dan Ilmu Pengetahuan Alam/Matematika

Menyatakan bahwa skripsi ini adalah hasil karya saya sendiri dan karya ilmiah ini belum pernah diajukan sebagai pemenuhan persyaratan untuk memperoleh gelar kesarjanaan strata satu (S1) dari Universitas Sriwijaya maupun perguruan tinggi lain.

Semua informasi yang dimuat dalam skripsi ini yang berasal dari penulis lain baik yang dipublikasikan atau tidak telah diberikan penghargaan dengan mengutip nama sumber penulis secara benar. Semua isi dari skripsi ini sepenuhnya menjadi tanggung jawab saya sebagai penulis.

Demikianlah surat pernyataan ini saya buat dengan sebenarnya.

Indralaya, 22 Februari 2023

Penulis

Yulfita Tasya

NIM. 08011281924041

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

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

Nama	: Yulfita Tasya
NIM	: 08011281924041
Fakultas/Jurusan	: Matematika dan Ilmu Pengetahuan Alam/Matematika
Jenis Karya	: Skripsi

Demi pengembangan ilmu pengetahuan, saya menyetujui untuk memberikan kepada Universitas Sriwijaya "hak bebas royalti non-eksklusif (*non-exclusively royalty-free right*) atas karya ilmiah saya yang berjudul:

"Kombinasi Metode Imputasi *Mean* dan *Multiple Imputation by Chained Equations* (MICE) untuk Penanganan Data Hilang dan Peningkatan Evaluasi Kinerja Klasifikasi Prediksi Penyakit Diabetes Melitus"

Beserta perangkat yang ada (jika diperlukan). Dengan hak bebas royalti non-eksklusif ini Universitas Sriwijaya berhak menyimpan, mengalih media/memformatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir atau skripsi saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik hak cipta.

Demikian pernyataan ini saya buat dengan sesungguhnya.

Indralaya, 22 Februari 2023

Penulis

Yulfita Tasya

NIM. 08011281924041

HALAMAN PERSEMBAHAN

Kupersembahkan skripsi ini untuk:

Yang Maha Kuasa Allah Subhanahu Wa Ta'ala,

Kedua orang tuaku tersayang,

Satu-satunya saudariku,

Keluarga besarku,

Semua guru dan dosenku,

Sahabat-sahabatku,

Almamaterku

Motto

"Asa yang kecil ini terus memupuk diri, bertumbuh, dan berkembang seraya berharap asa ini bukan hanya sekadar asa"

KATA PENGANTAR

Puji syukur kehadirat Allah Subhanahu wa Ta'ala yang telah memberikan rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Kombinasi Metode Imputasi *Mean* dan *Multiple Imputation by Chained Equations* (MICE) untuk Penanganan Data Hilang dan Peningkatan Kinerja Klasifikasi Prediksi Penyakit Diabetes Melitus" sebagai salah satu syarat memperoleh gelar sarjana sains bidang studi Matematika di Fakultas MIPA Universitas Sriwijaya.

Penulis menyadari bahwa proses pembuatan skripsi ini merupakan proses pembelajaran yang sangat berharga serta tak lepas dari kekurangan dan keterbatasan. Dengan segala hormat dan kerendahan hati, penulis mengucapkan terima kasih dan penghargaan kepada Kedua orang tuaku tercinta, Bapak **Sidharta Gautama** dan Ibu **Yuniar Kursusi**, yang tak pernah lelah merawat, mendidik, menuntun, memberi nasehat, semangat serta doa untuk penulis. Terima kasih atas segala perjuangan, pengorbanan, serta kasih sayang hingga detik ini dan sampai kapanpun. Penulis juga mengucapkan terima kasih dan penghargaan kepada:

1. Bapak **Drs. Sugandi Yahdin, M.M** selaku Ketua Jurusan Matematika FMIPA Universitas Sriwijaya yang telah memberikan arahan dan motivasi kepada penulis selama proses perkuliahan dan Ibu **Dr. Dian Cahyawati Sukanda, M.Si** selaku Sekretaris Jurusan Matematika FMIPA Universitas Sriwijaya yang telah mengarahkan urusan akademik kepada penulis.

2. Ibu **Dr. Anita Desiani, S.Si., M.Kom** selaku dosen pembimbing utama yang telah bersedia memberikan waktu, tenaga, pikiran, nasehat, dan motivasi untuk memberikan bimbingan dan pengarahan selama proses pembuatan skripsi, kompetisi, dan perjalanan perkuliahan ini dan Ibu **Dr. Yuli Andriani, S.Si., M.Si** selaku dosen pembimbing pendamping yang telah bersedia memberikan waktu, tenaga, pikiran, nasehat, dan motivasi untuk memberikan bimbingan dan pengarahan selama proses pembuatan skripsi serta proses perkuliahan dengan penuh pengertian dan kesabaran.
3. Bapak **Dr. Bambang Suprihatin, S.Si., M.Si** dan Bapak **Drs. Endro Setyo Cahyono, M.Si** selaku dosen pembahas dan penguji yang telah memberikan tanggapan, kritik, dan saran yang sangat bermanfaat untuk perbaikan dan penyelesaian skripsi ini.
4. Ibu **Indrawati, S.Si., M.Si** selaku dosen pembimbing akademik yang telah membimbing dan mengarahkan urusan akademik penulis.
5. Seluruh **Dosen di Jurusan Matematika FMIPA** yang telah memberikan ilmu, nasihat, motivasi, serta bimbingan selama proses perkuliahan.
6. Pak **Irwansyah** selaku admin dan Ibu **Hamidah** selaku pegawai tata usaha Jurusan Matematika FMIPA yang telah membantu penulis selama perkuliahan.
7. **Seluruh guru** yang telah memberikan ilmu yang bermanfaat hingga mengantarkan penulis pada pendidikan ini.

8. Satu-satunya saudariku tersayang, **Miswarita Putri Gautama**, yang selalu mendoakan dan memberikan perhatian kepada penulis, beserta keluarga besar yang selalu mendukung penulis.
9. **Alumni Mahasiswa Matematika 2018 dengan NIM 08011181823006** yang senantiasa menemani, memberikan semangat, dan mendoakan penulis sehingga dapat menyelesaikan skripsi ini.
10. **Semua sahabat seperjuangan** dalam masa perkuliahan dan proses skripsi. Terima kasih sudah menjadi orang-orang baik di sekeliling penulis yang selalu mendukung, membantu dengan tulus, dan memberi energi positif.
11. **Keluarga Matematika 2019, Tim PHP2D Himastik Unsri 2020, BPH Himastik Akselerasi, BPH COIN Periode 2020/2021, BPH IKMS Kabinet Nebula,** dan rekan-rekan selama perkuliahan.
12. Kakak-kakak tingkat angkatan 2016, 2017, dan 2018 serta adik-adik tingkat angkatan 2020, 2021 dan 2022, terima kasih atas segala kebaikan dan bantuan.
13. Semua pihak yang tidak dapat penulis sebutkan satu per satu. Semoga segala kebaikan yang diberikan mendapatkan balasan terbaik dari Allah SWT.

Semoga skripsi ini dapat menambah pengetahuan dan bermanfaat bagi mahasiswa/mahasiswi Jurusan Matematika Fakultas dan Ilmu Pengetahuan Alam Universitas Sriwijaya dan semua pihak yang memerlukan.

Indralaya, Januari 2023

Penulis

COMBINATION OF MEAN IMPUTATION METHODS AND MULTIPLE IMPUTATION BY CHAINED EQUATIONS (MICE) FOR HANDLING MISSING DATA AND IMPROVING PERFORMANCE EVALUATION OF DIABETES MELLITUS CLASSIFICATION PREDICTION

By

YULFITA TASYA
NIM. 08011281924041

ABSTRACT

Pima Indians Diabetes 2020 dataset is one of the datasets that contains missing data. Missing data can cause some statistical information to be lost due to the small sample size and can cause overfitting problems in the training data. One way to deal with missing data can be done by imputing data. This study aims to improve classification performance on Pima Indians Diabetes 2020 dataset by applying a combination of Single Imputation using the Mean imputation method on attributes containing missing data less than or equal to 10% and Multiple Imputation using MICE on attributes containing more than 10% missing data. 10%. The results of missing data imputation were tested using the Multi Layer Perceptron (MLP) and Support Vector Machine (SVM) methods to find out the increase in classification performance evaluation. Before handling missing data, the results of the classification performance evaluation obtained an accuracy of 78.947%, a precision of 78.554%, and a recall of 76.616%, after handling missing data using the Mean and MICE methods, the results of the classification performance evaluation obtained an accuracy of 84.221%, a precision of 82.462%, and a recall of 82.462%. Accuracy, precision and recall values increased by 5.274%, 3.908% and 5.846% respectively. It can be concluded that the prediction of missing data using the Multi Layer Perceptron (MLP) and Support Vector Machine (SVM) methods can improve the performance evaluation of the prediction classification of diabetes mellitus.

Keywords: Missing Data, Mean, MICE, MLP, SVM, and Classification

KOMBINASI METODE IMPUTASI *MEAN* DAN *MULTIPLE IMPUTATION BY CHAINED EQUATIONS* (MICE) UNTUK PENANGANAN DATA HILANG DAN PENINGKATAN EVALUASI KINERJA KLASIFIKASI PREDIKSI PENYAKIT DIABETES MELITUS

Oleh

YULFITA TASYA
NIM.08011281924041

ABSTRAK

Dataset Pima Indians Diabetes Tahun 2020 merupakan salah satu *dataset* yang mengandung data hilang**.** Data yang hilang dapat menyebabkan beberapa informasi statistik hilang karena ukuran sampel menjadi kecil dan dapat menyebabkan masalah *overfitting* dalam *training* data**.** Salah satu cara untuk mengatasi data hilang dapat dilakukan dengan melakukan imputasi data. Penelitian ini bertujuan untuk meningkatkan kinerja klasifikasi pada *dataset* Pima Indians Diabetes Tahun 2020 dengan menerapkan kombinasi dari Imputasi Tunggal menggunakan metode imputasi *Mean* pada atribut yang mengandung data hilang kurang dari atau sama dengan 10% dan Imputasi Ganda menggunakan MICE pada atribut yang mengandung data hilang lebih dari 10%. Hasil dari imputasi data hilang diuji coba menggunakan metode *Multi Layer Perceptron* (MLP) dan *Support Vector Machine* (SVM) untuk mengetahui peningkatan evaluasi kinerja klasifikasi. Sebelum dilakukan penanganan data hilang, hasil evaluasi kinerja klasifikasi diperoleh akurasi sebesar 78,947%, presisi sebesar 78,554%, serta *recall* sebesar 76,616%, setelah dilakukan penanganan data hilang menggunakan metode *Mean* dan MICE, hasil evaluasi kinerja klasifikasi memperoleh akurasi sebesar 84,221%, presisi sebesar 82,462%, serta *recall* sebesar 82,462%. Nilai akurasi, presisi, dan *recall* masing-masing meningkat sebesar 5,274%, 3,908%, dan 5,846%. Dapat disimpulkan bahwa prediksi data hilang menggunakan metode *Multi Layer Perceptron* (MLP) dan *Support Vector Machine* (SVM) dapat meningkatkan evaluasi kinerja klasifikasi prediksi penyakit diabetes melitus.

Kata Kunci : Data Hilang, *Mean*, MICE, MLP, SVM, dan Klasifikasi

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang

Data hilang merupakan data yang mengandung informasi yang tidak lengkap berupa nilai atribut yang hilang karena berbagai faktor (Liu *et al*., 2016). Data yang hilang dapat menyebabkan beberapa informasi statistik hilang karena ukuran sampel menjadi kecil dan dapat menyebabkan masalah *overfitting* dalam *training data* (Li *et al*., 2020). Data hilang dapat menurunkan hasil kinerja dalam mengenali pola pada masalah klasifikasi (Pedersen *et al*., 2017). *Dataset* Pima Indians Diabetes Tahun 2020 merupakan salah satu *dataset* yang mengandung banyak data hilang. *Dataset* Pima Indians Diabetes Tahun 2020 dapat diperoleh dari *website kaggle* yang dapat diakses melalui laman (https://www.kaggle.com/datasets/oguz-kaanmavice/pimaindians-diabetes-with-null-values). *Dataset* ini berisi 9 fitur untuk mendiagnosa pasien penyakit diabetes antara lain *Pregnant*, *Glucose*, *Diastolic Blood Pressure (Diastolic* BP), *Skin Fold*, *Serum Insulin*, *Body Mass Indeks* (BMI), *Diabetes Predigree*, *Age* dan *Class* dimana kelas merupakan label yang berisi kategori *tested positive* (positif) dan *tested negative* (negatif). Terdapat 5 dari 9 fitur yang mengalami data hilang yaitu pada fitur *Glucose*, *Diastolic* BP, *Skin Fold*, *Serum Insulin*, BMI dengan persentase yang bervariasi mulai dari 0,7% sampai dengan 48,7% dari keseluruhan data.

Beberapa penelitian dilakukan untuk menanggulangi permasalahan data hilang pada *dataset* Pima Indians Diabetes Tahun 2020 diantaranya adalah Azrar

(2018) menggunakan imputasi *Mean* dengan mengukur nilai evaluasi dengan hasil akurasi masih dibawah 75%. Bodinga *et al*. (2022) melakukan penanganan data hilang pada *dataset* Pima Indians Diabetes Tahun 2020 dengan penghapusan kolom pada fitur yang mengandung data hilang dan menghasilkan akurasi masih dibawah 75%. Barale & Shirke (2016) melakukan penanganan data hilang pada dataset *dataset* Pima Indians Diabetes Tahun 2020 dengan menggunakan metode *K-Nearest Neighbor* (KNN) dan menghasilkan akurasi masih dibawah 75%.

Teknik *preprocessing* yang tepat dapat meningkatkan kinerja klasifikasi, salah satunya adalah penanganan data hilang dengan imputasi data (Symeonidis *et al*., 2018). Imputasi data adalah teknik menangani nilai yang hilang dengan cara mengganti data hilang dengan nilai tertentu melalui teknik statistik atau pembelajaran mesin (Abidin & Ismail, 2018). Cara mengatasi data hilang dengan imputasi data terdiri dari metode Imputasi Tunggal dan Imputasi Ganda. Imputasi Tunggal adalah metode yang menggantikan secara langsung data yang hilang menggunakan satu nilai dari suatu variabel (Lee *et al*., 2022). Metode yang dapat dilakukan untuk mengatasi data hilang pada tipe data numerik adalah dengan menggunakan metode imputasi *Mean* (Zhang, 2016).

Imputasi Tunggal menggunakan metode imputasi *Mean* sebaiknya digunakan pada atribut dengan data hilang tidak lebih dari 10% (Desiani *et al*. 2021). Atribut dengan data hilang diatas 10% dikhawatirkan akan menghasilkan nilai perkiraan yang bias karena mengabaikan varian dari populasi atau sampel yang ada (Eekhout *et al*. 2014; Pedersen *et al*. 2017). Beberapa penelitian menggunakan Imputasi Tunggal dengan metode imputasi *Mean* diantaranya

adalah Silva-Ramírez *et al*. (2015) melakukan imputasi data hilang pada dataset penyakit jantung menggunakan metode imputasi *Mean* dan menghasilkan akurasi masih dibawah 70%. Silva-Ramírez *et al*. (2015) melakukan imputasi data hilang pada dataset penyakit kardiomiopati menggunakan metode imputasi *Mean* dan menghasilkan akurasi masih dibawah 70%. Souto *et al*. (2015) melakukan imputasi data hilang pada dataset penyakit otak menggunakan metode imputasi *Mean* dan menghasilkan akurasi masih dibawah 60%.

Imputasi Ganda adalah metode yang diusulkan untuk mengkompensasi kekurangan dari metode Imputasi Tunggal (Lee *et al*., 2022). Imputasi Ganda bekerja dalam tiga langkah, pertama menentukan nilai setiap data hilang dengan menggunakan model statistik, kedua membuat kumpulan data dengan nilai baru yang dianalisis menggunakan prosedur statistik standar, ketiga hasil analisis akhir ditentukan menjadi analisis statistik keseluruhan berdasarkan kesalahan standar (Ginkel *et al*., 2020). Pendekatan nilai hilang dengan Imputasi Ganda merupakan alternatif untuk prediksi data hilang dalam jumlah besar (Desiani *et al*., 2021). Salah satu metode yang dapat digunakan untuk Imputasi Ganda adalah *Multiple Imputation by Chained Equations* (MICE).

MICE adalah adaptasi dari metode Imputasi Ganda yang bekerja dengan mengubah imputasi masalah keserangkaian estimasi dimana setiap variabel diprediksi menggunakan model penduga regresi pada variabel lainnya (Wulff & Ejlskov, 2017). Beberapa penelitian lain menggunakan MICE menunjukkan hasil kinerja yang baik, diantaranya adalah penelitian yang dilakukan oleh Tran *et al*. (2018) menerapkan MICE pada data penyakit hepatitis dengan menggabungkan

dengan metode *Multilayer Perceptron* (MLP) menghasilkan nilai akurasi sebesar 83,8%. Penelitian lain dilakukan oleh Rafsunjani & Safa (2019) mengenai sistem tekanan udara dengan menggabungkan MICE dan *Random Forest* (RF) menghasilkan akurasi sebesar 94,79%. Selain itu, Mera-Gaona *et al*. (2021) menerapkan MICE untuk data penyakit jantung dengan menghasilkan nilai akurasi sebesar 85,8%. Kelebihan dari MICE adalah dapat menghasilkan kesalahan standar yang lebih masuk akal dibandingkan dengan pendekatan Imputasi Tunggal karena menggunakan perkiraan berulang dengan model yang berbeda (Saffari *et al*., 2022).

Pada penelitian ini akan mengimputasi data hilang pada *dataset* Pima Indians Diabetes Tahun 2020 dengan menerapkan kombinasi dari Imputasi Tunggal menggunakan metode imputasi *Mean* untuk atribut yang mengandung data hilang kurang dari atau sama dengan 10% dan Imputasi Ganda menggunakan MICE untuk atribut yang mengandung data hilang diatas 10%. Pengaruh dari imputasi data hilang akan diuji menggunakan beberapa algoritma klasifikasi yaitu *Multi Layer Perceptron* (MLP) dan *Support Vector Machine* (SVM). Hasil klasifikasi algoritma tersebut akan diukur akurasi, presisi, dan *recall* untuk evaluasi kinerja dari imputasi data yang dilakukan.

1.2 Perumusan Masalah

Rumusan masalah dalam penelitian ini adalah bagaimana mengatasi data hilang menggunakan Imputasi Tunggal dengan metode imputasi *Mean* dan Imputasi Ganda dengan metode imputasi MICE pada *dataset* Pima Indians

Diabetes Tahun 2020 untuk meningkatkan kinerja klasifikasi dengan mengukur akurasi, presisi, dan *recall*.

1.3 Pembatasan Masalah

Permasalahan pada penelitian ini hanya terbatas pada :

1. Penanganan data hilang dilakukan dengan metode Imputasi *Mean* pada data yang mengandung data hilang kurang dari atau sama dengan 10% dan metode Imputasi MICE pada data yang mengandung data hilang lebih dari 10%.
2. Hasil evaluasi kinerja klasifikasi disimpulkan berdasarkan nilai akurasi, presisi dan *recall* yang diperoleh.

1.4 Tujuan

Tujuan dari penelitian ini adalah untuk meningkatkan kinerja dataset Pima Indians Diabetes Tahun 2020 dengan mengaplikasikan kombinasi metode imputasi *Mean* dan MICE pada data hilang dilihat dari keakuratan prediksi penyakit diabetes melitus sesuai nilai akurasi yang diperoleh.

1.5 Manfaat

Manfaat dari penelitian ini adalah:

1. Memiliki *dataset* yang lengkap tanpa mengandung data hilang yang dapat digunakan untuk klasifikasi prediksi gangguan diabetes melitus.
2. Menambah referensi dibidang kesehatan dalam memprediksi penyakit diabetes melitus.

Daftar Pustaka

Abidin, N. Z., & Ismail, A. R. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *9*(6), 442–447.

Alaoui, S. S., Farhaoui, Y., & Aksasse, B. (2018). Classification algorithms in data mining. *International Journal of Tomography and Simulation*, *6*(1), 1–6. https://www.researchgate.net/profile/B-Aksasse/publication/326866871_Classification_algorithms_in_Data_Mining/links/5b9785ae4585153a5329962d/Classification-algorithms-in-Data-Mining.pdf

Azifah, N., Pauzi, M., Wah, Y. B., Deni, S. M., & Khatijah, S. (2021). Comparison of single and MICE imputation methods for missing values. *Science & Technology*, *29*(2), 979–998.

Azrar, A. (2018). Data mining models comparison for diabetes prediction. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *9*(8), 320–323.

Bodinga, B. A., Abdulsalam, M. A., Buhari, B. A., & Mansur, M. (2022). On the analysis of some machine learning algorithms for the prediction of diabetes. *International Journal of Advanced Networking and Applications*, *14*(01), 5294–5299. https://doi.org/10.35444/ijana.2022.14109

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, *14*(113), 13–21.

Chicco, D., Tötsch, N., & Jurman, G. (2021). The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, *14*, 1–22. https://doi.org/10.1186/s13040-021-00244-z

Darmawahyuni, A., Nurmaini, S., & Firdaus, F. (2019). Coronary heart disease interpretation based on Deep Neural Network. *Computer Engineering and Applications Journal*, *8*(1), 1–12. https://doi.org/10.18495/comengapp.v8i1.288

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340–341*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033

Desiani, A., Dewi, N. R., Fauza, A. N., Rachmatullah, N., Arhami, M., &

Nawawi, M. (2021). Handling missing data using combination of Deletion Technique, Mean, Mode and Artificial Neural Network imputation for heart disease dataset. *Science and Technology Indonesia*, *6*(4), 303–312. https://doi.org/10.26554/sti.2021.6.4.303-312

Dinh, D. T., Huynh, V. N., & Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, *571*, 418–442. https://doi.org/10.1016/j.ins.2021.04.076

Dzulkalnine, M. F., & Sallehuddin, R. (2019). Missing data imputation with fuzzy feature selection for diabetes dataset. *SN Applied Sciences*, *1*(4), 1–12. https://doi.org/10.1007/s42452-019-0383-x

Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, *67*(3), 335–342. https://doi.org/10.1016/j.jclinepi.2013.09.009

Erler, N. S., Rizopoulos, D., Rosmalen, J. van, Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, *35*(17), 2955–2974. https://doi.org/10.1002/sim.6944

Ginkel, J. R. Van, Linting, M., Rippe, R. C. A., Voort, A. Van Der, Ginkel, J. R. Van, Linting, M., Rippe, R. C. A., & Van, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data rebutting existing misconceptions about multiple imputation as a method for. *Journal of Personality Assessment*, *102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680

Gorade, S. M., Deo, A., & Purohit, P. (2017). A study some data mining classification techniques. *International Research Journal of Engineering and Technology (IRJET)*, *4*(1), 210–215. https://doi.org/10.21884/ijmter.2017.40-31.zt9tv

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining*, *9*(3), 139–154. https://doi.org/10.1002/sam.11312

Huang, J., Mao, B., Bai, Y., Zhang, T., & Miao, C. (2020). An integrated Fuzzy C-Means method for missing data imputation using Taxi GPS Data. *MDPI Journal Sensors*, *20*, 1–19.

J.Roiger, R. (2016). Data mining a tutorial-based primer. In *A Chapman & Hall Book*.https://medium.com/@arifwicaksanaa/pengertian-use-case-a7e576e1b-6bf

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2019). Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, *00*(00), 1–13. https://doi.org/10.1080/23789689.2019.1600960

Kaiser, J. (2014). Dealing with missing values in data. *Journal of Systems Integration*, *1*, 42–51. https://doi.org/10.20470/jsi.v5i1.178

Khan, S. I., Sayed, A., & Hoque, L. (2020). SICE : an improved missing data imputation technique. *Journal of Big Data*, *7*(37), 2–21. https://doi.org/10.1-186/s40537-020-00313-w

Lee, D., Woo, S., Jung, M., & Heo, T. (2022). Evaluation of odor prediction model performance and variable importance according to various missing imputation methods. *MDPI Applied Sciences*, *12*, 1–19.

Li, L., Du, B., Wang, Y., Qin, L., & Tan, H. (2020). Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowledge-Based Systems*, *194*, 1–13. https://doi.org/10.1016/j.knos-ys.2020.105592

Liu, D., Liang, D., & Wang, C. (2016). A novel three-way decision model based on incomplete information system. *Knowledge-Based Systems*, *91*(July), 32–45. https://doi.org/10.1016/j.knosys.2015.07.036

Majid, A. M., & Utomo, W. H. (2021). Application of discretization and adaboost method to improve accuracy of classification algorithms in predicting diabetes mellitus. *ICIC Express Letters, Part B: Applications*, *12*(12), 1177–1184. https://doi.org/10.24507/icicelb.12.12.1177

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Mera-Gaona, M., Neumann, U., Vargas-Canas, R., & López, D. M. (2021). Evaluating the impact of multivariate imputation by MICE in feature selection. *Plos One*, *16*, 1–28. https://doi.org/10.1371/journal.pone.0254720

Noei, M., & Abadeh, M. S. (2019). A genetic asexual reproduction optimization algorithm for imputing missing values. *International Conference on Computer and Knowledge Engineering (ICCKE)*, 214–218. https://doi.org/10.1109/ICCKE48569.2019.8964808

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*,

157–166. https://doi.org/10.2147/CLEP.S129785

Rafsunjani, S., & Safa, R. S. (2019). An empirical comparison of missing value imputation techniques on APS failure prediction. *I.J. Information Technology and Computer Science*, *2*, 21–29. https://doi.org/10.5815/ijitcs.2-019.02.03

Saffari, S. E., Volovici, V., Ong, M. E. H., Goldstein, B. A., Vaughan, R., Dammers, R., Steyerberg, E. W., & Liu, N. (2022). Proper use of multiple imputation and dealing with missing covariate data. *World Neurosurgery*, *161*, 284–290. https://doi.org/10.1016/j.wneu.2021.10.090

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE. *American Journal of Epidemiology*, *179*(6), 764–774. https://doi.org/10.1093/aje/kwt312

Silva-Ramírez, E. L., Pino-Mejías, R., & López-Coello, M. (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, *29*, 65–74. https://doi.org/10.1016/j.asoc.2014.09.0-52

Souto, M. C. P. D., Jaskowiak, P. A., & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. *B-MC Bioinformatics*, *16*(1), 1–9. https://doi.org/10.1186/s12859-015-0494-3

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*, 298–310. https://-doi.org/10.1016/j.eswa.2018.06.022

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining*, *10*(6), 363–377. https://doi.org/10.100-2/sam.11348

Tran, C. T., Zhang, M., Andreae, P., Xue, B., & Bui, L. T. (2018). An effective and efficient approach to classification with incomplete data. *Knowledge-Based Systems*, *154*, 1–16. https://doi.org/10.1016/j.knosys.2018.05.013

Venkata Vara Prasad, D., Venkataramana, L., Balasubramanian, P., Priyankha, B., Rajagopal, S., & Dattuluri, R. (2019). An efficient pre-processing method for improved classification of diabetics using decision tree and artificial neural network. *AIP Conference Proceedings*, *2161*(October). https://doi.org/10.10-63/1.5127648

Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2015). The use of

sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, *53*, 73–80. https://doi.org/10.1016/j.jbi.2014.09.003

Wulff, J. N., & Ejlskov, L. (2017). Multiple imputation by chained equations in praxis: Guidelines and review. *Electronic Journal of Business Research Methods*, *15*(1), 41–56.

Xiao, J., & Bulut, O. (2020). Evaluating the performances of missing data handling methods in ability estimation from sparse data. *Educational and Psychological Measurement*, *80*(5), 932–954. https://doi.org/10.1177/0013164420911136

Yahdin, S., Desiani, A., Gofar, N., Agustin, K., & Rodiah, D. (2021). Application of the Relief-f algorithm for feature selection in the prediction of the relevance education background with the graduate employment of the Universitas Sriwijaya. *Computer Engineering and Applications Journaland Applications Journal*, *10*(2), 71–80. https://comengapp.unsri.ac.id/index.php-/comengapp/article/view/369%0Ahttps://comengapp.unsri.ac.id/index.php/comengapp/article/download/369/228

Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*, *4*(1), 1–8. https://doi.org/10.3978/j.issn.2305-583-9.2015.12.38