AUTHORS' BIODATA

Amir Ali is the Asset Reliability General Manager for Joint-Venture Malaysia of PETRONAS Carigali Sdn. Bhd. (PCSB). He holds BSc in Chemical Engineering with Minor in Business Administration from the University of Tennessee, Knoxville, USA. Graduated with a Dean list. A result-driven professional with more than 30-year experience of upstream oil and gas operations both in Malaysia and internationally. His passion in applying Artificial Intelligence and Machine Learning within his domain of expertise resulted in the deployment of DOMAIN successfully to Dulang field gas compression system. During the project, he stewarded the training of Engineers and Technician in Big Data Analytic and ideate the inclusion of Big Data Analytic course within Engineering Undergraduate program at UTP. DOMAIN, which stands for Digitalizing Operations and Maintenance via Artificial Intelligence is his brainchild and carry the testimony of impactful collaboration between UTP Cerdas brilliant academician mind and his vast experience in the industry. He was also the former Deputy Skill Group Advisor of SKG-16 (Process Technology and Operations) for PCSB.

Emelia Akashah Patah Akhir graduated PhD in Computing from the University of Brighton, United Kingdom. She is also an alumnus of the University of Surrey, United Kingdom, where she obtained her MSc in Information System. Before that, she holds her Diploma and First-Class BSc (Hons) in Computer Science. Upon completing her PhD, she returned to Malaysia and work as a lecturer in Universiti Teknologi PETRONAS. Currently, Dr Emelia works as a lecturer at Department of Computer & Information Sciences, Universiti Teknologi PETRONAS. Her research interest is mainly but not limited to business/IT alignment, sustainable supply chain, E-commerce, information systems, knowledge management, information architecture, information strategies, and big data governance. Dr. Amelia is also a member of the UK Association of Information.

Izzatdin Abdul Aziz is an Associate Professor and Head at the Centre for Research in Data Science (CeRDaS) in Universiti Teknologi PETRONAS (UTP), where he focuses on solving complex upstream oil and gas (O&G) industry problems from the viewpoint of computer science. He obtained his PhD in Information Technology from Deakin University, Australia working in the domain of hydrocarbon exploration. Prior to his PhD, he earned his Master's degree in Information Technology specializing in Computer Networks from the University of Sydney, Australia. Dr Izzatdin has secured and delivered numerous computing-related O&G projects mainly relating to seismic data processing and implementation of workflows

for processing hydrocarbon exploration dataset on cloud-based systems. He is working closely with O&G companies in delivering solutions for complex problems such as Offshore O&G pipeline corrosion rate prediction, O&G pipeline corrosion detection, securing data on clouds and designing and implementing Metocean prediction system. Additionally, he is also working on Big data transmission, security and optimization problems on High-performance Computing Clusters. His research is funded by the Petroleum Research Fund, PETRONAS and The Malaysian Ministry of Higher Education.

Jafreezal Jaafar is an Associate Professor and Dean at theFaculty of Science and Information Technology at Universiti Teknologi PETRONAS (UTP), Malaysia. He holds a PhD from University of Edinburgh, UK (2009), MAppSc in IT from RMIT University, Australia (2002) and BSc in Computer Science from Universiti Teknologi Malaysia (1998). He joined Department of Computer and Information Sciences, UTP since 1999, where he was a former Head (2012-2016). His main research areas are AI, Machine Learning and Data Analytics, with over 100 technical publications. Currently, Dr Jafreezal is also attached to Centre for Research in Data Science (CeRDaS), which focuses on the implementation of AI and Machine Learning in the oil and gas industry. He is also active in conducting professional training in AI and Big Data Analytics for public and industry sector. Based on his vast experience and expertise, he had become the member of Academy Science Malaysia SIG in Machine Learning, executive member for IEEE Computer Society Malaysia Chapter (2016-2018) and executive member for MyAIS (2017-2019).

Mohd Hilmi Hasan is a Senior Lecturer at the Computer and Information Sciences Department, and researcher at the Centre for Research in Data Science (CeRDaS) in Universiti Teknologi PETRONAS (UTP), Malaysia. He obtained his PhD in Information Technology from UTP in 2017 with a research in the area of artificial intelligence, specifically Interval Type-2 fuzzy. He also graduated with Master of Information Technology (eScience) from The Australian National University in 2004 and Bachelor of Technology (Hons.) Information Technology from UTP in 2002. Dr Mohd Himi has been teaching and being an active researcher at UTP since 2004, in which one of the courses that he coordinated was the Technopreneurship Team Project (TTP). His research interest covers the areas of artificial intelligence and data analytics. He has involved in many research projects awarded by the government agencies and industry that are related to data analytics, data engineering and predictive modelling.

Norshakirah Abdul Aziz is a Senior Lecturer at Computer and Information Sciences Department at Universiti Teknologi PETRONAS, Malaysia. Currently, she is a researcher at UTP Centre of Research in Data Sciences (CeRDaS) and a Data Governance Leader at High-Performance Cloud Computing Data Centre (HPCCC). She holds a Diploma in Management Technology, a bachelor's degree in Information Technology (IT), an MSc Degree in IT, and a Ph.D. in e-SCM Integration & Collaboration. She has a total of 19 years of experience gained from academic institutions and industry under her belt. Her industry working experience is related to Business Intelligence, E-Business, and IT Project Management. Her research scope covers Business Intelligence, Data Analytics, Data Governance, and Digital Addiction. She is also a qualified Technologist Specialist (Ts), and HRDF Certified Trainer, and a present EXCO Committee Member for Data Science Association, Malaysia.

Said Jadid Abdulkadir received the B.Sc. degree in computer science from Moi University, the M.Sc. degree in computer science from Universiti Teknologi Malaysia, and the PhD degree in information technology Universiti Teknologi PETRONAS. He is currently a Senior Lecturer with the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS.His current research interests include machine learning, data clustering and predictive analytics. Said Jadid is currently serving as the editor-in-chief for the Journal of Applied Artificial Intelligence, Chief Editor for Platform A Journal of Engineering, and reviewer for Artificial Intelligence Review, the IEEE ACCESS, and Knowledge-Based Systems.

BIBLIOGRAPHY

[1] B. Anderson, M. Genty, D. Hart, & E. Thanhardt, "Using Data Science to Understand Tape-Based Archive Workloads," in *XSEDE'15*, Jul. 2015, doi: 10.1145/2792745.2792776

[2] M. Maoz, "How IT should deepen big data analysis to support customer-centricity." Gartner.com. 2013. [Online]. Available: https://www.gartner.com/en/documents/2531116

[3] T. Erl, W. Khattak, & P. Buhler, *Big Data Fundamentals: Concepts, Drivers & Techniques*: Pearson, 2016.

[4] DAMA-International, *DAMA-DMBOK: Data Management Body of Knowledge*, 2nd Ed, Technics Publication, 2017.

[5] Kaggle. "Supermarket sales,". Kaggle.com. 2021. [Online]. Available: https://www.kaggle.com/aungpyaeap/supermarket-sales.

[6] T.H. Davenport, P. Barth, & R. Bean, in "How big data is different," *MIT Sloan Management Review*, vol. 54, no.1, pp. 43-46, 2012.

[7] D. Delen & H. Demirkan, *Data, information and analytics as services*: Elsevier, vol. 55, pp. 359-363, 2013.

[8] L.T. DeCarlo, "On the meaning and use of kurtosis.", *Psychological Methods*, vol. 2, no. 3, pp. 292–307, 1997, doi: 10.1037/1082-989X.2.3.292

[9] P.H. Westfall, "Kurtosis as peakedness, 1905–2014. RIP," *The American Statistician*, vol. 68, no.3, pp. 191-195, 2014.

[10] N. Mishra & S. Silakari, "Predictive analytics: a survey, trends, applications, opportunities & challenges," *International Journal of Computer Science and Information Technologies*, vol. 3, no. 3, pp. 4434-4438, 2012.

[11] S. Earley, "Big data and predictive analytics: What's new?," *IT Professional,* vol. 16, no. 1, pp. 13-15, 2014, doi: 10.1109/MITP.2014.3

[12] K. Lepenioti, A. Bousdekis, D. Apostolou, & G. Mentzas, "Prescriptive analytics: a survey of approaches and methods," in *the International Conference on Business Information Systems*, 2018.

[13] B. Gill, B. C. Borden, & K. Hallgren, "A conceptual framework for data-driven decision making," In *Mathematica Policy Research. Seattle, WA*: Bill & Melinda Gates Foundation, Jun. 02, 2014.

[14] M. Zaitsava, E. Marku, & M. C. Di Guardo, "Is data-driven decision-making driven only by data? When cognition meets data," *European Management Journal*, 2022, doi: 10.1016/j.emj.2022.01.003

[15] A. Scappini, *80 Fundamental Models for Business Analysts: Descriptive, Predictive, and Prescriptive Analytics Models with Ready-To-Use Excel Templates*: CreateSpace Independent Publishing Platform, 2016.

[16] E. Alpaydın, *Introduction to Machine Learning*, Cambridge: MIT Press, 2020.

[17] R. Mitchell, J. Michalski, & T. Carbonell, *An Artificial Intelligence Approach*: Springer, 2013.

[18] T. M. Mitchell & T. M. Mitchell, *Machine Learning* (Vol. 1). New York: McGraw-hill, 1997.

[19] S. Russell & P. Nowig, *Artificial Intelligence: A modern approach (Vol. 90)*: Pearson Education, 2002.

[20] G.Moisen, "Classification and regression trees. In S. E. Jorgensen & B. Fath (Eds.)," *Encyclopedia of Ecology*. Oxford: UK Elsevier, 2008, vol. 1, pp. 582-588.

[21] C. Douglas, E. A. Montgomery, G. Geoffrey Vining, *Introduction to Linear Regression Analysis*: Wiley, 2021.

[22] R. E. Wright, "Logistic regression," In *Reading and Understanding Multivariate Statistics*. Washington: American Psychological Association, 1995, pp. 217-244.

[23] E. Ostertagová, "Modelling using Polynomial Regression," *Procedia Engineering*, no. 48, pp. 500-506, 2012, doi:10.1016/j.proeng.2012.09.545

[24] B. Thompson, "Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply here: A Guidelines Editorial," *Educational and Psychological Measurement,* vol. 55, no. 4, pp. 525-534, 1995, doi:10.1177/0013164495055004001

[25] A. E. Hoerl, R.W. Kannard, & K.F. Baldwin, "Ridge regression: some simulations," *Communications in Statistics,* vol. 4, no. 2, pp. 105-123,1975, doi:10.1080/03610927508827232

[26] M. Arashi, A.M.E. Saleh, & B.G. Kibria, *Theory of ridge regression estimation with applications*: John Wiley & Sons, 2019.

[27] J. Ranstam & J. A. Cook, "LASSO regression," *BJS (British Journal of Surgery),* vol. 105, no. 10, pp. 1348-1348, 2018, doi:10.1002/bjs.10895

[28] C. Hans, "Elastic Net Regression Modeling With the Orthant Normal Prior," *Journal of the American Statistical Association, vol. 106*, no. 496, pp. 1383-1393,2011. [Online]. Available: http://www.jstor.org/stable/23239545

[29] J. Feng, H. Xu, S. Mannor, & S. Yan, "Robust logistic regression and classification," in *the Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014, vol. 1, pp. 253-261

[30] D. Berrar, "Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology," *ABC of Bioinformatics*, Amsterdam, The Netherlands: Elsevier Science Publisher, 2018, pp. 403-412.

[31] S. L. Ting, W. H. Ip, & A. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification?" *International Journal of Software Engineering and its Applications,* vol. 5, no. 3, pp. 37-46,2011.

[32] M. Granik & V. Mesyura, "Fake news detection using naive Bayes classifier," *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017, pp.900-903.

[33] J. T. Laaksonen & E. Oja, "Classification with learning k-nearest neighbours," *Proceedings of International Conference on Neural Networks (ICNN'96),* vol. 3, no. 3 pp. 1480-1483, 1996.

[34] P. H. Swain & H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics,* vol. 15, no. 3, pp. 142-147, 1977, doi:10.1109/TGE.1977.6498972

[35] X. Huang, L. Shi, & J.A.K. Suykens, "Support Vector Machine Classifier With Pinball Loss," *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no.* 5, pp. 984-997, 2014, doi:10.1109/TPAMI.2013.178

[36] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing,* vol. 26, no. 1, pp. 217-222, 2005, doi:10.1080/01431160412331269698

[37] B. Schmidt & A. Hildebrandt, "Deep learning in next-generation sequencing," *Drug Discov Today,* vol. 26, no. 1, pp. 173-180, 2021, doi:10.1016/j.drudis.2020.10.002

[38] S. Niebler, A. Muller, T. Hankeln, & B. Schmidt, "RainDrop: Rapid activation matrix computation for droplet-based single-cell RNA-seq reads," *BMC Bioinformatics, vol. 21*, no. 1, p. 274,2020, doi:10.1186/s12859-020-03593-4

[39] J. Ding, A. Condon, & S.P. Shah, "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models," *Nat Commun,* vol. 9, no. 1, p. 2002, 2018, doi:10.1038/s41467-018-04368-5

[40] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J. R. MaguireHartl, C., ... & M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature genetics*, vol. 43, no. 5, pp. 491-49, 2011.

[41] H. B. Barlow, "Unsupervised Learning," *Neural Computation,* vol. 1, no. 3, pp. 295-311, 1989, doi:10.1162/neco.1989.1.3.295

[42] Z. Ghahramani, "Unsupervised learning," in *Summer school on machine learning*, Berlin, Heidelberg: Springer, 2003, pp. 72-112

[43] S. Miloudi, Y. Wang, & W. Ding, "A Gradient-Based Clustering for Multi-Database Mining," *IEEE Access,* vol. 9, pp. 11144-11172, 2021 doi:10.1109/ACCESS.2021.3050404

[44] S. Park, S. Han, S. Kim, D. Kim, S. Park, S. Hong, & M. Cha, "Improving Unsupervised Image Clustering With Robust Learning," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2021.

[45] A. Ahmad & L. Dey, (2007). "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503-527, 2007, doi: 10.1016/j.datak.2007.03.016

[46] K. G. Derpanis, "Mean Shift Clustering. Lecture Notes in Computer Vision", 32, 2005.

[47] J. Yang, S. Rahardja, & P. J. P. R. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognit*, vol. 115, p. 107874, 2021.

[48] S. Anand, S. Mittal, O. Tuzel, & P. Meer, "Semi-Supervised Kernel Mean Shift Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 6, pp. 1201-1215, 2014, doi:10.1109/TPAMI.2013.190

[49] S. Teki, K. V. Sriharsha, & M. Nandimandalam, "A Diabetic Prediction System Based on Mean Shift Clustering," *Ingénierie des Systèmes d Inf.,* vol. 26, pp. 231-235, 2021.

[50] D. Comaniciu, & P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 24, no. 5, pp. 603–619, 2002, doi:10.1109/34.1000236

[51] H. Sadewo, Y. Satria, & H. Burhan, "Application of Mean Shift Clustering to optimize matching problems in ridesharing for maximize the total number of match," *Journal of Physics: Conference Series*, vol. 1821, p. 012019, 2021.

[52] T. Genish & C. Balaji, "A Novel Skull Stripping Tool Based on Mean Shift Clustering and Mathematical Morphology," *Turkish Journal of Physiotherapy Rehabilitation,* vol. 32, no. 2, pp. 1235-1241, 2021.

[53] P. K. Saha, & D. Logofatu, "Efficient Approaches for Density-Based Spatial Clustering of Applications with Noise," in *The IFIP International Conference on Artificial Intelligence Applications and Innovations*, 2021.

[54] B. Hassan, R. Zineb, L. Amine, & L. Elhoussine, "A Complexity Survey on Density based Spatial Clustering of Applications of Noise Clustering Algorithms," in *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 12 Issue 2, 2021, doi: 14569/IJACSA.2021.0120283

[55] V. Mistry, U. Pandya, A. Rathwa, H. Kachroo, & A. Jivani, "AEDBSCAN—Adaptive Epsilon Density-Based Spatial Clustering of Applications with Noise," In *Progress in Advanced Computing and Intelligent Engineering*: Springer, pp. 213-226, 2021.

[56] L. van der Maaten, E. Postma, & H. Herik, "Dimensionality Reduction: A Comparative Review," *Journal of Machine Learning Research,* vol. 10, 2009.

[57] X. Xie, "Principal component analysis," *Wiley interdisciplinary reviews*, 2019.

[58] H. Abdi & L. J. Williams, "Principal component analysis," *WIREs Computational Statistics,* vol. 2, no. 4, pp. 433-459, 2010, doi: 10.1002/wics.101

[59] S. Wold, K. Esbensen, P. J. C. Geladi, & i. l. systems, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37-52, 1987.

[60] H. H. Harman, *Modern factor analysis*: University of Chicago press, 1976.

[61] P. Kline, *An easy guide to factor analysis*: Routledge, 2014.

[62] D. Zhang, X. Jing, Y., & J. Yang, "Linear discriminant analysis," In *Biometric Image Discrimination Technologies: Computational Intelligence and Its Applications Series*: IGI Global, 2006, pp. 41-64

[63] Y. Li, B. Liu, Y. Yu, H. Li, J. Sun, & J. Cui, "3E-LDA: Three Enhancements to Linear Discriminant Analysis," *ACM Transactions on Knowledge Discovery from Data,* vol. 15, no. 4, 2021, doi:10.1145/3442347

[64] A. Patcha & J. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448-3470, 2007.

[65] F. T. Liu, K. M. Ting, & Z. Zhou, -H. , "Isolation forest*," Paper presented at the 2008 eighth ieee international conference on data mining*, pp. 413-422, 2008, doi: 10.1109/ICDM.2008.17

[66] M. Maoz, "How IT should deepen big data analysis to support customer-centricity", 2013. [Online.] Available: https://www.gartner.com/en/documents/2531116

[67] P. D. Santo, "On the Alleged Use of Keplerian Telescopes in Naples in the 1610s,", 2021. [Online.] Available: https://arxiv.org/abs/2101.06465

[68] M. Dobre, "Henry More and William Petty: Revisiting an Early Modern Polemic," *Early Science and Medicine,* vol. 23, pp. 244-264, 2018, doi:10.1163/15733823-00233P03

[69] J. E. Ciecka & G. R. Skoog, "Life Expectancies and Annuities: A Modern Look at an Old Fallacy", *Mathematics Magazine, 91*, 3, pp. 163-170, 2018, doi:10.1080/0025570X.2018.1439256

[70] J. Yang, Y. Wang, Y. Yu, J. Xiao, & X. Wang, "Huygens' synchronization experiment revisited: luck or skill?" *European Journal of Physics*, vol. 39, no. 5, p. 055004, 2018, doi:10.1088/1361-6404/aac98e

[71] R. Geen, S. Bordoni, D. Battisti, & K. Hui, "Monsoons, ITCZs and the Concept of the Global Monsoon," *Reviews of Geophysics,* vol. 58, 2020, doi:10.1029/2020RG000700

[72] H. Grossman, "W. Playfair, the Earliest Theorist of Capitalist Development," *The Economic History Review,* vol. 18, no. 1/2, pp. 65-83, 1948, doi:10.2307/2590263

[73] K. Ramasubramanian, & A. Singh, *Machine Learning Using R*: Apress, 2017.

[74] M. Badri, "Comparison of two approaches for web-based 3d visualization of smart building sensor data," Master's thesis, University of Oulu, 2017 [Online]. Available: http://jultika.oulu.fi/files/nbnfioulu-201706022485.pdf

[75] R. Clary, "The royal school of mines: Henry de la beche's convergence of professionalization and public advocacy," *Earth Sciences History,* vol. 39, pp. 291-304, 2020, doi:10.17704/1944-6187-39.2.291

[76] M. Eckardt & J. Mateu, "Partial and Semi-Partial Statistics of Spatial Associations for Multivariate Areal Data," *Geographical Analysis,* vol. 53, no. 4, pp. 818-835, 2021, doi: 10.1111/gean.12266

[77] S. Rendgen, *The Minard System: The Complete Statistical Graphics of Charles-Joseph Minard*. Princeton Architectural Press, 2018.

[78] D. C. Hoaglin, "John W. Tukey and Data Analysis," *Statistical Science,* vol. 18, no. 3, pp. 311-318, 2003. [Online]. Available: http://www.jstor.org/stable/3182748

[79] C. Perin, J. -D. Fekete, & P. Dragicevic, "Jacques Bertin's legacy in information visualization and the reorderable matrix," *Cartography and Geographic Information Science*, vol. 46, no. 2, pp. 176-181, 2019, doi:10.1080/15230406.2018.1470942

[80] M. Bécue-Bertaut, *Textual Data Science with R*: Chapman and Hall/CRC, 2018

[81] N. LePan, "Visualizing the history of pandemics," 2020. [Online]. Available: https://www.visualcapitalist.com/history-of-pandemics-deadliest/

INDEX

A

accuracy, 13, 18, 21
analytical techniques, 7
anomaly detection, 58, 62

B

bad actor, 92
big data
    analysis (BDA), 26
    analytics, 4, 7
    lifecycle, 8-10
business case, 9, 82

C

completeness, 13, 21-22

D

data
    driven decision making (DDDM), 40
    acquisition and filtering, 9,
    aggregation, 96
    analysis, 10
    cleansing, 5, 94
    extraction, 9
    identification, 9
    preparation, 94
    management association (DAMA), 12
    visualisation, 3, 10, 80
    quality assessment, 11, 20
    dimension, 19
    management, 11
dataset, 13-14
descriptive
    analytics, 5, 27
    statistics, 28
diagnostic analytics, 6

E

elasticNet, 53

F

fourth industrial revolution (IR 4.0), 64

H

histogram, 33
historical data, 35, 85

I

isolation forest, 62

K

K-Mean algorithm, 58

L

lasso regression, 52-53
Linear Regression (LR), 50
logistic regression, 50, 55

M

machine learning, 4, 48, 54

N

normalisation, 77

P

perspective analytics, 6
predictive analytics, 6, 34-35, 47
python, 23

R
regression
    learning 49
    model, 50
ridge regression, 52

S
skewness, 29, 32
standard deviation (SD), 30
supervised learning, 48-49

T
time series analysis, 35, 36
timeliness, 13, 16

U
uniqueness, 13, 15, 21
unsupervised learning, 57-58

V
validity, 13, 17