Journal of Statistical Theory and Applications
https://doi.org/10.1007/s44199-022-00048-y

REVIEW

Ingredients for Responsible Machine Learning: A Commented Review of *The Hitchhiker's Guide to Responsible Machine Learning*

Fernando Marmolejo-Ramos[1] · **Raydonal Ospina**[2] · **Enrique García-Ceja**[3] · **Juan C. Correa**[4]

Received: 19 July 2022 / Accepted: 2 September 2022

Abstract

In *The hitchhiker's guide to responsible machine learning*, Biecek, Kozak, and Zawada (here BKZ) provide an illustrated and engaging step-by-step guide on how to perform a machine learning (ML) analysis such that the algorithms, the software, and the entire process is interpretable and transparent for both the data scientist and the end user. This review summarises BKZ's book and elaborates on three elements key to ML analyses: inductive inference, causality, and interpretability.

Keywords Machine learning · Predictive statistics · Inference · Causality

Abbreviations

ML	Machine learning
BKZ	Biecek, Kozak, and Zawada
COVID-19	Coronavirus disease 2019

✉ Fernando Marmolejo-Ramos
fernando.marmolejo-ramos@unisa.edu.au

Raydonal Ospina
raydonal@de.ufpe.br

Enrique García-Ceja
e.g.mx@ieee.org

Juan C. Correa
juan.correan@cesa.edu.co

[1] Centre for Change and Complexity in Learning, University of South Australia, Adelaide, SA 5001, Australia

[2] CASTLab, Department of Statistics, Universidade Federal de Pernambuco, Recife, Pernambuco 51280-000, Brazil

[3] Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, 64849 Monterrey, Nuevo León, Mexico

[4] CESA Business School, Bogotá, Bogotá, DC 110231, Colombia

Published online: 15 September 2022

DGP	Data generating process
EDA	Exploratory data analysis
PD	Partial dependence
ALE	Accumulated local effects
1R	One rule algorithm
PCR	Polymerase chain reaction
K-NN	k-Nearest neighbors algorithm

1 Introduction

Complex, varied, and big data sets are being amassed rapidly in different fields thanks to digitisation. In the field of health sciences, for example, such data sets have been emerging due to the COVID-19 pandemic [77]. Making sense of such types of data requires powerful and sophisticated computational, mathematical, and statistical tools. Machine learning (ML) is a favourite approach to deal with those data sets as it consists of computer algorithms tuned to automatically find patterns in data [33]. One of the major criticisms of ML, though, is that the algorithms' internal workings are not tailored to human understanding. Biecek et al. [11] provide a concise, accessible, and engaging tutorial on how to carry out ML analyses that use powerful algorithms in a way that allows both the data scientist and the end user to interpret the workings of the ML analytical process (see also Murdoch et al. [63]). Following canonical book reviews, we summarise and briefly comment on BKZ's book. BKZ's book is rich in concepts relating to statistical learning, statistical modelling, and computational statistics, that could be further commented on. However, we chose to elaborate on three concepts that the reader should keep in mind while reading BKZ's book because they are key to ML and any other form of data analysis: inductive inference, causality, and interpretability.

2 A Commented Summary of the Book

BKZ's book presents the way three fictional data scientists—Bit, Beta, and DALEX—undertake an ML analysis of a COVID-19 data set. While Bit is eager to have quick results, Beta is more cautious and diligent in undertaking further checks and inspecting more predictive models. DALEX is a robot (akin to a friendly version of a Dalek!) that demands explanations of the models built and prompts Bit and Beta to provide these at key steps during the model building. The conceptual foundations underlying these data scientists' analytical pipeline are grounded in proposals found in Breiman [14] and James et al. [46]. Those conceptual foundations are further developed in more detail in a book by one of the authors Biecek et al. [10].

Bit and Beta are tasked to come up with a predictive model able to determine the risk of death in case of an infection and suggest the age order in which people need to be vaccinated. That is, the data scientists have to sort patients by their individual risks. Bit and Beta thus commence reading up on the topic of COVID-19 to familiarise themselves with the terminology and related aspects. Also, as no data are

given to them, they start to find a comparable and representative data set with which to build the predictive model. This step is crucial in that the data set Bit and Beta use will be the data generating process (DGP) substantiating any statistical model such that any subsequent explanation and prediction is directly dependent on the DGP.

With the data at hand, Bit and Beta create a training data set and a test data set (in BKZ's book, these are the `COVID_spring` and `COVID_summer` data sets, respectively). The former is used to build the model and the latter is used to validate the model. BKZ's book briefly comments on a key aspect; a true validation is done on a separate new data set. Note that 'true validation' is different from cross-validation (sometimes called rotation estimation or out-of-sample testing). Different from a true validation approach, in cross-validation the original data set is split in such a way that a large chunk of the data (say, 80%) is used to train the model, and the remaining data is used to test the model [82]. Cross-validation, or any form of model assessment, is at the core of model building in that it enables examination of the stability of the model's estimations [92]. Anecdotally, cross-validation predates bootstrapping [25], an influential technique used in statistical modelling [34], and these two techniques can be used in conjunction in ML analyses [84].

Bit and Beta move onto exploring (via exploratory data analysis [EDA] techniques [85]) and cleansing the data. When these steps are cleared, they are ready to consider statistical algorithms suitable to the data and the research problem at hand. This is the stage where the predictive power of some (binomial) classification algorithms is assessed via DALEX. At this point the reader realises that DALEX is a robot that embodies the `DALEX` R package, a package designed for assessing and explaining predictive models [9]. Bit, Beta, and DALEX first try a regression tree that uses the variables 'age' and 'cardiovascular diseases' (that these variables were used by the regression tree is not surprising as these variables were also highlighted during the EDA phase of the data analysis). The data scientists then try an algorithm that is an improvement on regression trees: random forests. The results are better and after some optimisation of the hyperparameters (i.e. tuning) the diagnostic ability of the binary classifier improves even more. BKZ explain how to optimise hyperparameters and evaluate the importance of the data set's variables. The examination of the variables is furthered via partial dependence (PD) and accumulated local effects (ALE). We will not expand on hyperparameter optimisation, variable importance, PD, and ALE as BKZ already do this in their book. Regarding classification algorithms, it is important to note that although classification trees and random forests provide good visuals of decision trees, there are other algorithms that can assist in classification tasks. There are, for example, the one rule (1R) [42] and the Boruta [52] algorithms. A logistic regression algorithm could also be considered as it has been shown this method is more interpretable than, yet similarly accurate to, more complex ML algorithms [19, 54, 57, 64]. Note that it is indeed possible to combine classification algorithms in order to inform a final model. For example, Cardona et al. [16] used the Boruta and 1R algorithms for selecting variables to be used in a logistic regression model. In the case of numeric dependent variables, techniques such as distributional regression trees and forests [74] and transformation forests [44] could be used (these are implemented in the `disttree` and `trtf` R packages, respectively).

Bit, Beta, and DALEX inspect their models further through Shapley values (a concept from cooperative game theory), break-down plots, and ceteris paribus plots (a.k.a. what-if plots). Once again, these concepts are clearly explained in BKZ's book but other sources such as Biecek and Burzykowski Molnar [10] and [60] are recommended. Once the three data scientists are satisfied with the results of the further assessment of the models and the results of some individual risk analyses, they are thus finally ready to deploy the model. The three data scientists create an application that allows any individual to estimate the probability of severe condition and death after being diagnosed with COVID-19 depending on age, gender (male or female), presence/absence of cardiovascular disease, presence/absence of cancer, presence/absence of kidney disease, presence/absence of diabetes, and presence/absence of other diseases (the app lives at https://crs19.pl/). In the app's page, it is made explicit that the model is built using a sample of 50,000+ cases in Poland who gave a positive PCR (polymerase chain reaction) test for COVID-19. Other important information about the data set, variables, and models is provided therein.

In a nutshell, BKZ's book argues that responsible ML consists of preparing the data, understanding it, proposing an ensemble of models to parse the data (based on the research question), carefully auditing the models, and finally deploying the models. Thus, BKZ's book sets an example of what good practices and principles in explainable ML should look like [6]. As mentioned earlier, BKZ's book is rich in concepts that cut across, mostly, the fields of statistical and computer sciences. We chose three concepts central to data science in general (including ML) and we consider them in turn.

3 Inductive Inference

Inductive inference [2, 3, 22] can be understood as an ML procedure [7, 12, 70, 78] or algorithm [29, 37] that assumes a specific type of relationship between hypotheses about the data and propositions that go beyond the data (and these include predictions about future data, general conclusions about all possible data, and the DGP) [20].

Inductive inference aims to provide the best predictions and identify the best model for inferential purposes (variable selection, hypothesis testing, etc.) that allow the generation of scientific knowledge and interpretation. A key premise, though, is that simple models are preferable [18, 94]. Inductive inference requires assumptions for the application of statistical tests; however, from an ML perspective, an algorithm, by definition, is a set of finite steps that become an inductive inferential process in itself [79, 80, 91]. That is, any assumption check built into statistical testing is stripped by inferential processes carried out by algorithms [36, 69].

The language used to describe patterns in the data, sample size [53], computational complexity of problems [67] in approximating concepts [30, 72], and poor pattern identification methods further adds a layer of complexity to inductive inference. The way those domains are described can induce biases in inductive inferential reasoning [47] (an example of this can be found in several probabilistic problems) [5, 15, 48, 50, 71, 83].

In the specific case of interpretation of results obtained via ML, it has been argued that ML researchers tend to incur the illusion of probabilistic proof by contradiction, which consists of the erroneous belief that a null hypothesis becomes improbable because a significant result has been obtained [27, 28]. This illusion is, however, difficult to eradicate in the use of inductive inference. Given that BKZ's book embraces an ML approach, it does not stress the importance of the verification of hypotheses, attention to the limits of extrapolating results [40, 45, 81], and securing corrective measures [26, 59]. We strongly believe that these are aspects in inductive inference to which future work in ML should give serious consideration.

4 Causality

In statistics and ML literature, causality or causal inference (i.e., deciding whether a variable X causes Y or vice versa) is one of the most debated topics in the academic community. The possibility of making causal inferences represents an ideal mechanism for any scientist trying to uncover natural laws, and traditional approaches to uncovering these laws favour controlled experiments [38, 41, 62]. Besides controlled experiments, more recent data-driven perspectives suggest other techniques for causal inference purposes in experimental and non-experimental contexts [13, 75]. BKZ's book's position regarding this topic is evident: predictive models are mentioned without implying any connection to causality or causal inference. Such a pedagogical position, we believe, not only mirrors the infancy that describes the current stage of the literature on ML and causality, but also exploits the data of COVID-19 to illustrate how different ML models can be used in R and how they provide several approaches to the same problem: modelling individual mortality risk after COVID-19 infection.

In our view, even though the topic of causality was not covered in BKZ's book, the reader is encouraged to understand that this topic cannot be ignored. Regardless of existent contrasting views on the possible ways to make causal inferences out of ML models, there will always be relevant spaces for discussing these classic concerns in statistical reasoning. For example, Bontempi and Flauder [13] proposed a supervised ML approach to infer the existence of a directed causal link between two variables in multivariate settings with $n > 2$ variables. By the same token, the idea of discriminating cause from effect with observational non-experimental data is well introduced by Mooij et al. [62]. Since then, another branch of the literature presents interesting insights about the way researchers can learn causality from data [38, 65, 93]. In line with the working paper of Schölkopf [75], we also believe that the hard open problems of ML and AI are intrinsically related to causality and that this is another central topic requiring more attention from ML researchers.

5 Interpretability (explainable ML and AI)

Pedagogical efforts like the one provided by BKZ are undoubtedly helpful in an era where several institutions leverage 'black-box' ML models for high-stakes decisions (e.g. healthcare and criminal justice) [73]. The utility of these efforts is evident when

it comes to illustrating how ML models work in general and how they reach their predictions in particular. In our view, the use of COVID-19 data makes BKZ's book a clear and updated reference and highlights their unique intended goal: finding a balance between technicalities and possible pedagogical illustrations through funny adventures of comic characters. In just 54 pages, the book does not pretend to dive deep into the inner workings of the methods. Nonetheless, it provides a good sample of appropriate references and serves as an intuitive starting point for beginners. A more expert audience might find helpful other sources that invest more pages for similar purposes without the pedagogical resource of comics [49].

BKZ make the distinction between two types of ML interpretable methods: global model-based and instance-based. Examples of both types of methods are presented. One thing to note is that the primary focus of the book is on explainable methods for *supervised* ML and *tabular data*. Given recent advances in algorithms that work on more unstructured data such as text, images, and time series of varying length, explainable ML methods have also permeated into those domains. For example, Assaf and Schumann [4] proposed a deep neural network to explain time series predictions. Liu et al. [55] developed a framework for generating explanations for natural language processing tasks; specifically, text classification. Furthermore, in recent years explainable AI methods outside the supervised learning domain have been developed, for example, for unsupervised clustering [23, 32] and reinforcement learning [68] (see also Bhatti [8]).

One aspect that is closely related to explainable ML and that is not covered in BKZ's book (and is also left aside in many other explainable ML materials) is a model's *uncertainty quantification*. By design, many ML models always produce a prediction regardless of their quality and without providing guarantees of their uncertainty; for example, *p*-values for classification and confidence intervals for regression. In medical applications and other domains it is of critical importance to know if a model's prediction can be trusted; alas, such information is not usually available. Many models like neural networks, decision trees, K-NN, and so on can produce prediction scores or probabilities; however, those are relative to the given data point and class (in the case of classification) but do not necessarily represent the overall probability distribution. When analysing predictions, it is important to consider both their explanations and their trustworthiness. The later can be assessed with *conformal prediction*, which is a framework proposed by Vovk et al. [90] to estimate the predictions' uncertainty. One of the advantages of this framework is that it is model agnostic. A recent method for uncertainty estimation was also proposed by Sensoy et al. [76]; however, it is specific to deep learning models.

There are several implementations of many explainable ML methods in the form of R packages. An extensive list of 27 packages was compiled and analysed by Maksymiuk et al. [58], with `DALEX` [9], `lime` [66], and `iml` [61] being some of the most popular (based on GitHub stars). Some R packages for general ML are implemented in `EnsembleML`, `cvms` (cross-validation for model selection), and `MachineShop` (see also the CRAN site on ML at https://cran.r-project.org/web/views/MachineLearning.html). Finally, there is another ML-related technique that the reader of BKZ's book should be aware of that is known as 'targeted learning'. This approach relies on ML algorithms to assess uncertainty and provide reliable

estimations of the true target parameters of the probability distribution of the data [86–89] (an online free book can be found at https://tlverse.org/tlverse-handbook/ and the key R packages are `SuperLearner` and `tmle`). In our view, BKZ's book invites the reader to conceive ML analyses and models that are interpretable so that their utility is optimised.

6 Final Thoughts

ML is a technique that automates data analysis by resorting to the power of statistical tools [21, 31] and has become a favoured framework to cope with big data by producing predictive models [24] across several fields [1, 43, 51, 56]. However, those models are known for lacking interpretability and explainability [17] and this, in turn, reduces their accountability because issues relating to risk assessment and safe adoption are overlooked [39]. BKZ's book aims to alleviate that problem by providing a concise and engaging tutorial on how to carry out careful and responsible ML analyses. We thus recommend their book as complementary reading for those undertaking ML-related courses. Different from current introductory textbooks on ML (e.g. Ghatak [35]), BKZ's book shows that an ML-based analysis is not about fiddling with black-box algorithms and praying for the best. Instead, the authors show that ML-based analyses require carefully selecting and tuning algorithms that, while giving accurate predictions, retain a good level of interpretability and explainability. That is, the ML analysis and analyst become responsible. We believe this message applies not only to ML modelling but to all forms of data analysis.

Acknowledgements The authors thank Kim Wilson (insightediting.com.au) for copyediting this manuscript. The flipbook version of BKZ's book can be found at (https://betaandbit.github.io/RML/#p=1) and a .pdf version can purchased from (https://leanpub.com/RML). The data set and R code used by BKZ are available at (https://github.com/MI2DataLab/ResponsibleML-UseR2021) and (https://htmlpreview.github.io/?https://raw.githubusercontent.com/MI2DataLab/ResponsibleML-UseR2021/main/modelsXAI.html), respectively. More work by the lead author of BKZ's book can be found at (https://www.mi2.ai/).

Author Contributions RO wrote the section 'inductive inference'. JCC wrote the section 'causality'. EG-C wrote the section 'interpretability (explainable ML and AI)'. FM-R conceptualised the overall idea and wrote the remaining sections.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability No datasets were generated or analysed during the current study

Declarations

Conflict of Interest There are no relevant financial or non-financial competing interests to report.

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. Ahmed, S., Alshater, M., El Ammari, A., Hammami, H.: Artificial intelligence and machine learning in Finance: a bibliometric review. Res. Int. Bus. Finance **61**, 101646 (2022). https://doi.org/10.1016/j.ribaf.2022.101646
2. Ambainis, A.: Probabilistic inductive inference: a survey. Theoret. Comput. Sci. **264**(1), 155–167 (2001)
3. Angluin, D., Smith, C.H.: Inductive inference: Theory and methods. ACM Comput. Surv. (CSUR) **15**(3), 237–269 (1983)
4. Assaf, R., Schumann, A.: Explainable deep neural networks for multivariate time series predictions. In: Proceedings of the Twenty-eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Macao, pp. 6488–6490 (2019)
5. Beck, J.: Can bootstrapping explain concept learning? Cognition **158**, 110–121 (2017)
6. Belle, V., Papantonis, I.: Principles and practice of explainable machine learning. Front. Big Data (2021). https://doi.org/10.3389/fdata.2021.688969
7. Bergadano, F.: Machine learning and the foundations of inductive inference. Mind. Mach. **3**(1), 31–51 (1993)
8. Bhatti, M.I.: Cluster Effects in Mining Complex Data. Nova Science Publishers, New York (2012)
9. Biecek, P.: DALEX: explainers for complex predictive models in R. J. Mach. Learn. Res., **19**(84), 1–5 (2018). Retrieved from http://jmlr.org/papers/v19/18-416.html. Accessed 1 Sept 2022
10. Biecek, P., Burzykowski, T.: Explanatory Model Analysis. Explore, Explain, and Examine Predictive Models. CRC Press, New York (2021)
11. Biecek, P., Kozak, A., Zawada, A.: The Hitchhiker's Guide to Responsible Machine Learning. The R Version. Warsaw University of Technology, Warsaw (2022)
12. Blum, L., Blum, M.: Toward a mathematical theory of inductive inference. Inf. Control **28**(2), 125–155 (1975)
13. Bontempi, G., Flauder, M.: From dependency to causality: a machine learning approach. J. Mach. Learn. Res. **16**(1), 2437–2457 (2015)
14. Breiman, L.: Statistical modelling. The two cultures. Stat. Sci. **16**(3), 199–231 (2001)
15. Butzer, T.: Bootstrapping and dogmatism. Philos. Stud. **174**(8), 2083–2103 (2017)
16. Cardona, J., Grisales-Cardenas, J.S., Trujillo-Llano, C., Diazgranados, J.A., Urquina, H.F., Cardona, S., Marmolejo-Ramos, F.: Semantic memory and lexical availability in Parkinson's disease: a statistical learning study. Front. Aging Neurosci. (2021). https://doi.org/10.3389/fnagi.2021.697065
17. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics (2019). https://doi.org/10.3390/electronics8080832
18. Case, J., Smith, C.: Comparison of identification criteria for machine inductive inference. Theoret. Comput. Sci. **25**(2), 193–220 (1983)
19. Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., Van Calster, B.: A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. Clin. Epidemiol. **110**, 12–22 (2019). https://doi.org/10.1016/j.jclinepi.2019.02.004
20. Clarke, B.S., Clarke, J.L.: Predictive Statistics: Analysis and Inference Beyond Models. Cambridge University Press, Cambridge (2018)
21. Cunningham, S.J.: Machine learning and statistics. A matter of perspective. Working Paper 95/11. Department of Computer Science, the University of Waikato. Hamilton, NZ (1995)

22. Dalkey, N.C.: Inductive inference and the representation of uncertainty. Mach. Intell. Pattern Recogn. **4**, 393–397 (1986)
23. Dasgupta, S., Frost, N., Moshkovitz, M., Rashtchian, C.: Explainable k-means and k-medians clustering. In: Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, pp. 12–18 (2020)
24. Donoho, D.: 50 years of data science. J. Comput. Graph. Stat. **26**(4), 745–766 (2017)
25. Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**(1), 1–26 (1979)
26. Efron, B.: Resampling plans and the estimation of prediction error. Stats **4**(4), 1091–1115 (2021)
27. Ellis, P.D.: The Essential Guide to Effect Sizes: Statistical Power, Metaanalysis, and the Interpretation of Research Results. Cambridge University Press, Cambridge (2010)
28. Falk, R., Greenbaum, C.W.: Significance tests die hard: the amazing persistence of a probabilistic misconception. Theory Psychol. **5**(1), 75–98 (1995)
29. Flener, P., Schmid, U.: An introduction to inductive programming. Artif. Intell. Rev. **29**(1), 45–62 (2008)
30. Freivalds, R., Kinber, E.B., Wiehagen, R.: On the power of inductive inference from good examples. Theoret. Comput. Sci. **110**(1), 131–144 (1993)
31. Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Friede, T.: Is there a role for statistics in artificial intelligence? Adv. Data Anal. Classif. (2021). https://doi.org/10.1007/s11634-021-00455-6
32. Frost, N., Moshkovitz, M., Rashtchian, C.: Exkmc: expanding explainable k-means clustering. arXiv preprint arXiv:2006.02399 (2020)
33. Garcia-Ceja, E.: Behavior Analysis with Machine Learning Using R. CRC Press, New York (2021)
34. Gelman, A., Vehtari, A.: What are the most important statistical ideas of the past 50 years? J. Am. Stat. Assoc. **116**(536), 2087–2097 (2021). https://doi.org/10.1080/01621459.2021.1938081
35. Ghatak, A.: Machine Learning with R. Springer, New York (2017)
36. Gigerenzer, G.: Adaptive Thinking: Rationality in the Real World. Oxford University Press, USA (2000)
37. Gold, E.M.: Language identifition in the limit. Inf. Control **10**(5), 447–474 (1967)
38. Guo, R., Cheng, L., Li, J., Hahn, P., Liu, H.: A survey of learning causality with data: poblems and methods. ACM Comput. Surv. (2020). https://doi.org/10.1145/3397269
39. Hall, P., Gill, N., Cox, B.: Responsible Machine Learning. Actionable Strategies for Mitigating Risks and Driving Adoption. O'Reilly, Boston (2021)
40. Hayes, B.K., Heit, E.: Inductive reasoning 2.0. Wiley Interdiscipl. Revi. Cogn. Sci. **9**(3), e1459 (2018)
41. Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**(396), 945–960 (1986)
42. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Mach. Learn. **11**(1), 63–90 (1993). https://doi.org/10.1023/A:10226311189321
43. Hopkins, E.: Machine learning tools, algorithms, and techniques in retail business operations: consumer perceptions, expectations, and habits. J. Self-Gov. Manag. Econ. **10**(1), 43–55 (2022). https://doi.org/10.22381/jsme10120223
44. Hothorn, T., Zeileis, A.: Predictive distribution modeling using transformation forests. J. Comput. Graph. Stat. **30**(4), 1181–1196 (2021). https://doi.org/10.1080/10618600.2021.1872581
45. Hubbard, R., Haig, B.D., Parsa, R.A.: The limited role of formal statistical inference in scientific inference. Am. Stat. **73**(sup1), 91–98 (2019)
46. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning with Applications in R. Springer, New York (2021)
47. Johnson, G.M.: Algorithmic bias: on the implicit biases of social technology. Synthese **198**(10), 9941–9961 (2021)
48. Kahneman, D., Slovic, S.P., Slovic, P., Tversky, A.: Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge (1982)
49. Kamath, U., Liu, J.: Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning. Springer, Cham (2021)
50. Kim, B., Xu, C., Barber, R.: Predictive inference is free with the jackknife+-after-bootstrap. Adv. Neural. Inf. Process. Syst. **33**, 4138–4149 (2020)
51. Kumar, R., Saha, P.: A review on artificial intelligence and machine learning to improve cancer management and drug discovery. Int. J. Res. Appl. Sci. Biotechnol. **9**(3), 149–156 (2022)

52. Kursa, M.B., Rudnicki, W.R.: Feature selection with the Boruta package. J. Stat. Softw., **36**(11), 1–13. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v036i11https://doi.org/10.18637/jss.v036.i11 (2010)
53. Kuusela, V.: Paradigms in statistical inference for finite populations: Up to the 1950s. Statistics Finland (2011)
54. Levy, J.J., O'Malley, A.: Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. BMC Medical Research Methodology (2020). https://doi.org/10.1186/s12874-020-01046-3
55. Liu, H., Yin, Q., Wang, W.Y.: Towards explainable NLP: a generative explanation framework for text classification. arXiv preprint arXiv:1811.00196 (2018)
56. Lowe, M., Qin, R., Mao, X.: A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring. Water (2022). https://doi.org/10.3390/w14091384
57. Lynam, A., Dennis, J., Owen, K., Oram, R.A., Jones, A.G., Shields, B.M., Ferrat, L.A.: Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. Diagn. Progn. Res. (2020). https://doi.org/10.1186/s41512-020-00075-2
58. Maksymiuk, S., Gosiewska, A., Biecek, P.: Landscape of R packages for eXplainable artificial intelligence. arXiv preprint arXiv:2009.13248 (2020)
59. Mayo, D.G., Spanos, A.: Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Article Title Science. Cambridge University Press, Cambridge (2010)
60. Molnar, C.: Interpretable machine learning. A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/ (2022). Accessed 1 Sept 2022
61. Molnar, C., Bischl, B., Casalicchio, G.: iml: An R package for interpretable machine learning. JOSS **3**(26), 786 (2018). https://doi.org/10.21105/joss.00786
62. Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: methods and benchmarks. J. Mach. Learn. Res. **17**(1), 1103–1204 (2016)
63. Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. U.S.A. **116**(44), 22071–22080 (2019)
64. Nusinovici, S., Tham, Y.C., Chak Yan, M.Y., Wei Ting, D.S., Li, J., Sabanayagam, C., Cheng, C.-Y.: Logistic regression was as good as machine learning for predicting major chronic diseases. J. Clin. Epidemiol. **122**, 56–69 (2020). https://doi.org/10.1016/j.jclinepi.2020.03.002
65. Onaindia, E., Aineto, D., Jiménez, S.: A common framework for learning causality. Progress Artif. Intell. **7**(4), 351–357 (2018). https://doi.org/10.1007/s13748-018-0151-y
66. Pedersen, T.L., Benesty, M.: lime: Local interpretable model-agnostic explanations [Computer software manual]. Retrieved from https://CRAN.R-project.org/package$=$lime (R package version 0.5.2) (2021). Accessed 1 Sept 2022
67. Pitt, L.: Inductive inference, dfas, and computational complexity. In: International Workshop on Analogical and Inductive Inference, pp. 18– 44 (1989)
68. Puiutta, E., Veith, E.M.: Explainable reinforcement learning: a survey. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE 2020), Virtual edition, pp. 77– 95 (2020)
69. Rodgers, J.L.: The epistemology of mathematical and statistical modeling: a quiet methodological revolution. Am. Psychol. **65**(1), 1–12 (2010)
70. Romeijn, J.-W.: Statistics as inductive inference. In: Bandyopadhyay, P., Forster, M. (eds.) Philosophy of Statistics, pp. 751–774. Elsevier, Amsterdam (2011)
71. Ross, L., Nisbett, R.: Human Inference: Strategies and Shortcomings of Social Judgment. Prentice-Hall, Englewood Cliffs (1980)
72. Royer, J.S.: Inductive inference of approximations. Inf. Control **70**(2–3), 156–178 (1986)
73. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
74. Schlosser, L., Hothorn, T., Stauffer, R., Zeileis, A.: Distributional regression forests for probabilistic precipitation fore-casting in complex terrain. Ann. Appl. Stat. **13**(3), 1564–1589 (2019). https://doi.org/10.1214/19-AOAS1247
75. Schölkopf, B.: Causality for machine learning. In: Geffner, H., Dechter, R., Halpern, J. (eds.) Probabilistic and Causal Inference: the Works of Judea Pearl, pp. 765–804 (2022)
76. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. arXiv preprint arXiv:1806.01768 (2018)

77. Sheng, J., Amankwah-Amoah, J., Khan, Z., Wang, X.: COVID-19 pandemic in the new era of big data analytics: methodological innovations and future research directions. Br. J. Manag. **32**, 1164–1183 (2021). https://doi.org/10.1111/1467-8551.12441
78. Solomonoff, R.J.: An inductive inference machine. In: I.R.E. Convention Record, Section on Information Theory, Vol. 2, pp. 56–62 (1957)
79. Solomonoff, R.J.: A formal theory of inductive inference. Part i. Inf. Control **7**(1), 1–22 (1964)
80. Solomonoff, R.J.: A formal theory of inductive inference. Part ii. Inf. Control **7**(2), 224–254 (1964)
81. Souza, F., Gottgtroy, M.: Considerations about the effectiveness of inductive learning process in data-mining context. Manag. Inf. Syst., 331–339 (2000)
82. Stone, M.: Cross-validatory choice and assessment of statistical predictions (with discussion). J. R. Stat. Soc. B **36**(2), 111–147 (1974)
83. Taniguchi, H., Sato, H., Shirakawa, T.: A machine learning model with human cognitive biases capable of learning from small and biased datasets. Sci. Rep. **8**(1), 1–13 (2018)
84. Tsamardinos, I., Greasidou, E., Borboudakis, G.: Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach. Learn. **107**, 1895–1922 (2018)
85. Tukey, J.: Analysing data: sanctification or detective work? Am. Psychol. **24**(2), 83–91 (1969)
86. van der Laan, M.: Targeted learning: the link from statistics to data science. STAtOR **18**(4), 12–16 (2017)
87. van der Laan, M., Rose, S.: Targeted learning. In: van der Laan, M., Rose, S. (eds.) Causal Inference for Observational and Experimental Data. Springer, New York (2011)
88. van der Laan, M., Rose, S.: Targeted learning in data science. In: van der Laan, M., Rose, S. (eds.) Causal Inference for Complex Longitudinal Studies. Springer, New York (2018)
89. van der Laan, M., Starmans, R.J.: Entering the era of data science: targeted learning and the integration of statistics and computational data analysis. Adv. Stat. (2014). https://doi.org/10.1155/2014/502678
90. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
91. Wiehagen, R.: From inductive inference to algorithmic learning theory. New Gener. Comput. **12**(4), 321–335 (1994)
92. Yu, B., Kumbier, K.: Veridical data science. Proc. Natl. Acad. Sci. U.S.A. **117**(8), 3920–3929 (2020)
93. Zhang, K., Schölkopf, B., Spirtes, P., Glymour, C.: (2017, 11) Learning causality and causality-related learning: some recent progress. Natl. Sci. Rev., **5**(1), 26–29. Retrieved from https://doi.org/10.1093/nsr/nwx137https://arxiv.org/abs/https:// academic.oup.com/nsr/article-pdf/5/1/26/31567604/nwx137.pdf 10.1093/nsr/nwx137
94. Zhu, H., Hall, P., May, J.: Inductive inference and software testing. Softw. Test. Verif. Reliab. **2**(2), 69–81 (1992)