International Journal of Psychological Research
ISSN: 2011-2084
ISSN: 2011-7922
Facultad de Psicología. Universidad de San Buenaventura, Medellín

Vélez, Jorge I.
Machine Learning based Psychology: Advocating for A Data-Driven Approach
International Journal of Psychological Research, vol. 14, no. 1, 2021, January-April, pp. 6-11
Facultad de Psicología. Universidad de San Buenaventura, Medellín

DOI: https://doi.org/10.21500/20112084.5365

Available in: https://www.redalyc.org/articulo.oa?id=299067861001

International Journal of PSYCHOLOGICAL RESEARCH

EDITORIAL

Vol 14, N° 1
https://revistas.usb.edu.co/index.php/IJPR
ISSN 2011-2084
E-ISSN 2011-7922

Machine Learning based Psychology: Advocating for A Data-Driven Approach

Psicología basada en Aprendizaje Automático: Abogando por un Enfoque basado en Datos

Jorge I. Vélez[1,⋆]

[1] *Universidad del Norte, Barranquilla, Colombia.*

⋆**Corresponding author**:
Jorge I. Vélez
Email: jvelezv@uninorte.edu.co

Declaration of data availability: All relevant data are within the article, as well as the information support files.

Conflict of interests: The author has declared that there is no conflict of interest.

How to Cite:
Vélez, J. I. (2021). Machine Learning Psychology: Advocating for a Data-Driven Approach. *International Journal of Psychological Research*, *14*(1), 6–11.
https://doi.org/10.21500/20112084.5365

Since its beginnings, Psychology has been prone to both data generation and understanding of human behavior through data analysis. Back in 1879, Dr. Wilheim Wundt opened the first experimental psychology lab at the University of Leipzig to study reaction times. To many, this is considered the start of Psychology as a separate scientific discipline and the use of data analysis for data-driven decision making in the field (Flis, 2019; Tweney, 2003). In this Editorial, we briefly discuss how Psychology students, clinicians, and researchers may take part of the data revolution and help transforming Psychology, as we know it, into Machine Learning Psychology.

1. Data Explosion in Psychology: A Place for Data Science

Nowadays, there is an explosion of data in different areas, and Psychology is no exception (Mabry, 2011; Zhu et al., 2009). In fact, considering the different branches of modern Psychology today (King University, 2019; Ritchie & Grenier, 2003), it seems that the amount of data generated by psychologists is far away from decreasing. Hence, there is no doubt that psychologists would greatly benefit from combining theoretical models with the right *Data Science* tools to correctly analyze data from experiments and surveys (Loftus, 1996). Thus, training psychologists in *Data Science* is essential for understanding and visualizing data, developing predictive models, and, as a consequence, fostering knowledge generation (Neth, 2021a, 2021b). In other words, we need, starting from undergraduate programs, to provide the necessary tools to Psychology students to take part of the data revolution and, in the near future, being able to make data-driven decisions (Jack et al., 2018; Mandinach, 2012; Tolle et al., 2011).

Data Science is an exciting multidisciplinary and broad discipline that allows you to turn raw data into understanding and insight, and involves principles, processes, and techniques for understanding phenomena through the analysis of data using a Galaxy of connected topics ranging from basic Statistics and Probability (i.e., descriptive and inferential statistics) to Machine Learning (ML) and Artificial Intelligence (AI; Provost & Fawcett, 2013). Broadly speaking, there are

five types of analytical approaches in Data Science: (1) descriptive analytics, which explains what happened; (2) diagnostic analytics, which explains why things happened; (3) predictive analytics, which, by using predictive models, forecasts what is likely to happen based on observed data; (4) prescriptive analytics, which recommends a course of action based on the results of a predictive model; and (5) cognitive analytics, which exploits the advances in ML and AI (i.e., intelligent systems) through High Performance Computing to develop analytic models with a human-like intelligence (Dey, 2016; Gudivada et al., 2016; Lepenioti et al., 2020). Note that these approaches open new possibilities on analysing Psychology data that go beyond traditional summary statistics (i.e., mean, median, range, and standard deviations), correlation/regression analyses, and the assessment of psychometric properties of a clinical instrument (Cuartas Arias, 2017).

2. ML Psychology: Predictive Models, Clustering, and Intelligent Systems

In Psychology, data can be generated from different and diverse sources: ranging from surveys and clinical instruments/batteries, which asses important aspects of human behavior, to EEG, reaction times, and genetic and omics data that quantify changes in the brain and the frequency distribution of traits or gene/protein expression function and evolution (Bell & Cuevas, 2012; Bragazzi, 2013; Jiménez-Figueroa et al., 2017; Suarez et al., 2020). ML has called the research communitys attention for disclosing patterns, detecting objects, and developing predictive frameworks in several diseases (Dey, 2016; Dhall et al., 2020) as well as in several areas of Psychology, including Psychometrics, Experimental Psychology, Diagnosis, Treatment, follow-up, and Personalized and Predictive Care (Dwyer et al., 2018; Jacobucci & Grimm, 2020; Koul et al., 2018; Lin et al., 2020; Orrù et al., 2020; Rosenfeld et al., 2012; Shatte et al., 2019), demonstrating its usefulness for elucidating important aspects of disease.

When using ML, the data can be of any nature (i.e., binary, multinomial, ordinal or continuous), and the underlying assumptions are minimal. Whether or not we have an outcome variable for each individual in our sample, it defines the type of ML techniques to be applied (i.e., Supervised ML vs. Unsupervised ML). Broadly speaking, supervised ML refers to developing predictive models for an outcome of interest Y based on a set of predictors $X = (X_1, X_2, \ldots, X_P)^T$; the selection of the predictive model fitting the data best is performed based on an error-related measure (i.e., the root mean squared error [RMSE] and the mean absolute error [MAE] for continuous outcome variables, and the sensitivity, specificity, accuracy, and lift for dichotomous variables; Kuhn, 2008, 2020). Some of the most common supervised ML algorithms include Classification and Regression Trees (CART; Breiman et al., 1984), Random Forrest (RF; Breiman, 2001), Support Vector Machines (SVMs; Cortes & Vapnik, 1995) and eXtreme Gradient Boosting (XGBoost; Chen & Guestrin, 2016).

When the data lacks an outcome variable (i.e., case/control status or 'labels') while having different measures available (i.e., responses for a clinical battery), unsupervised ML techniques can be used to identify hidden complex structures in the data. Three of the main methods used in unsupervised ML are principal component analysis (PCA), multidimensional scaling (MDS), and clustering. PCA is a dimensionality reduction exploratory technique, based on the eigenvalue decomposition of the variance-covariance matrix, that allows visualizing high-dimensional data (i.e., $k \geq 3$ variables are measured) while preserving as much statistical information as possible (Joliffe & Morgan, 1992; Ringnér, 2008; Ritchie & Grenier, 2003). MDS allows the visualization of the similarity level of individuals in a data set by calculating a dissimilarity or distance function $D(\boldsymbol{X})$ such that individuals closely related to each other have low dissimilarity (Mead, 1992). In this sense, the choice of an appropriate dissimilarity function is crucial (Harmouch, 2021). Clustering methods, on the other hand, help to identify, based on a set of features or variables, groups of individuals that would be impossible to spot otherwise. Multiple clustering techniques available in the literature could be applied (i.e., K-means clustering, Hierarchical clustering, and distribution-, model- and density-based clustering techniques; Roman, 2019). However, the choice of which of these methods should be used depends heavily on the data and involves assessing the stability and compactness of the derived clusters using different performance measures (Pedregosa et al., 2011; Scikit-learn Project, 2021).

For high-dimensional data, combining PCA+clustering or MDS+clustering is a go-to recipe to graphically represent individuals relationships and subgroups according to some features. Subsequent work may include to develop ML predictive models that can classify new individuals to such derived groups (Roman, 2019). Interestingly, the combination unsupervised ML techniques may lead to the identification of individuals exhibiting differential clinical profiles (i.e., extreme phenotypes; Acosta et al., 2011; Arcos-Burgos et al., 2019; Elia et al., 2009; Pérez-Gracia et al., 2010; Vidal et al., 2020; Yu et al., 2017; Yu et al., 2018), hence contributing to the development of personalized interventions, treatments, and follow-up strategies. The combination of supervised and unsupervised ML techniques as well as the automation of the data analysis process could allow the development of data-driven Intelligent Systems supporting psychologists to make more accurate and timely decisions (de Mello & de Souza, 2019; Luxton, 2016).

3. R, Python and the Democratization of ML

Despite how promising transitioning to ML Psychology may seem, data-driven decision making requires not only a proficient understanding of Data Science, Data Analytics, and ML/AI techniques, as well as the Psychology component associated to the data at hand, but also a comprehensive computational set of tools that facilitates the implementation, validation, and deploying of ML models. Thus, ML Psychology imposes new challenges in terms of the level of training in computational tools and abstract thinking that Psychology students need to develop. In what follows, we present some open-source free-of-charge alternatives to get started.

For more than three decades, R (www.r-project.org) and Python (www.python.org) have taken the lead to democratize the use of ML algorithms to the general public by making them easily accessible at no cost. More recently, Julia (www.julialang.org) has emerged as a powerful, suitable, and efficient alternative.

Established as an open source project in 1995, R is a language and environment that provides a great variety of statistical and graphical techniques, including classical statistical tests, predictive modelling, clustering, and other ML algorithms, and it is highly extensible (R Core Team, 2021). For ML, R has multiple freely-available packages, which focused on ML, namely caret, dplyr, tensorflow, DataExplorer, ggplot2, kernLab, MICE, mlr3, plotly, randomForest, rpart, e1071, keras, and OneR. For more details, see the Comprehensive R Archive Network (CRAN) Task View (Hothorn, 2021).

On the other hand, Python, which was created by Guido van Rossum in 1991, is a widely-used, interpreted, object-oriented, and high-level programming language with dynamic semantics, used for general-purpose programming (Python Software Foundation, 2021, van Rossum 2009). For ML in Python, the scikit-learn (Pedregosa et al., 2011) is the go-to library. This library offers an open-source collection of simple, reusable and efficient classification, regression, clustering methods, as well as dimensionality reduction techniques, model selection algorithms, and pre-processing routines tools (Pedregosa et al., 2011) accessible to everybody for developing predictive models.

Finally, Julia is a general-purpose, dynamic, high-level, and high-performance programming language that started in 2012 by Jeff Bezanson, Stefan Karpinski, Viral B. Shah, and Alan Edelman. Julia was conceived to be as usable for general programming as Python, as easy for statistics as R, and as natural for string processing as Perl, as powerful for linear algebra as Matlab, as good at gluing programs together as the shell (Bezanson et al., 2012, para. 4). Similar to R and Python, Julia also offers tools for Data Visualization, Data Science, and ML.

4. Getting Closer and Closer to the Promised Land

With the explosion of data in Psychology, ML methods hold promise for personalized care by tailoring treatment decisions and clustering patients into taxonomies clinically meaningful. In other words, ML methods can be used to take us to a *Promised Land* where clinicians provide diagnosis and suggest treatment options based on data from an individual, instead of using a 'one-size-fits-all' approach (Cuartas Arias, 2019; Joyner & Paneth, 2019).

A recent review identified that depression, schizophrenia, and Alzheimer's disease were the most common mental health conditions studied via ML methods (Shatte et al., 2019). Other conditions included autism (Bone et al., 2015), frontotemporal dementia (Bachli et al., 2020), cognitive impairment (Na, 2019; Youn et al., 2018), and post-traumatic stress (Wani et al., 2020). Certainly, the challenge in the years to come is to expand the application of ML methods to other pathologies, especially in developing countries. Even more importantly, ML methods, properly applied, may lead to the discovery, for example, of relevant clinical aspects of understudied populations (Fröhlich et al., 2018). In this *Promised Land*, psychologists provide faster, timely, and more accurate diagnosis, and are able to dissect and identify individuals with subtle forms of the disease, and offer appropriately treatment options.

Despite getting us to this *Promised Land* where personalized psychological care is a reality for most people, ML can lead to misinformed conclusions in the absence of clinical domain expertise; focusing on Data Science and the application of ML methods only can produce misleading results and conclusions (Bone et al., 2015). Thus, it is not only important to deeply understand the clinical background of the field, but also to differentiate which ML methods can be used and how. In this regard, interdisciplinary collaboration between psychologists and researchers in areas related to Data Science and ML is crucial (Shatte et al., 2019). Because of this continuous interaction, communication is another relevant aspect. In ML Psychology, the practitioner must have excellent communication skills to be able to express his/her research questions to collaborators to synergically work and successfully address them as a team. It is also important for the ML Psychology practitioner to interpret and follow the results of applying ML methods, and be able to gain relevant insights into the psychology aspects of the condition under study (Bone et al., 2015).

References

Acosta, M. T., Vélez, J. I., Bustamante, M. L., Balog, J. Z., Arcos-Burgos, M., & Muenke, M. (2011). A two-locus genetic interaction between LPHN3 and 11q predicts ADHD severity and long-term

outcome. *Translational psychiatry*, *1*(7), e17. https://doi.org/10.1038/tp.2011.14.

Arcos-Burgos, M., Vélez, J., Martinez, A., Ribasés, M., Ramos-Quiroga, J., Sánchez-Mora, C., Richarte, V., Roncero, C., Cormand, B., Fernández-Castillo, N., Casas, M., Lopera, F., Pineda, D., Palacio, J., Acosta-López, J., Cervantes-Henriquez, M., Sánchez-Rojas, M., Puentes-Rozo, P., Molina, B., & Muenke, M. (2019). ADGRL3 (LPHN3) Variants Predict Substance Use Disorder. *Translational Psychiatry*, *9*(1), e42. https://doi.org/10.1038/s41398-019-0396-7.

Bachli, M. B., Sedeño, L., Ochab, J. K., Piguet, O., Kumfor, F., Reyes, P., Torralva, T., Roca, M., Cardona, J. F., Campo, C. G., Herrera, E., Slachevsky, A., Matallana, D., Manes, F., García, A. M., Ibáñez, A., & Chialvo, D. R. (2020). Evaluating the reliability of neurocognitive biomarkers of neurodegenerative diseases across countries: A machine learning approach. *NeuroImage*, *208*, 116456. https://doi.org/10.1016/j.neuroimage.2019.116456.

Bell, M. A., & Cuevas, K. (2012). Using EEG to Study Cognitive Development: Issues and Practices. *Journal of cognition and development: official journal of the Cognitive Development Society. 13*(3). 281–294. https://doi.org/10.1080/15248372.2012.691143.

Bezanson, J., Karpinski, S., Shah, V., & Edelman, A. (2012). Why We Created Julia. https://julialang.org/blog/2012/02/why-we-created-julia/.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of autism and developmental disorders*, *45*(5), 1121–1136. https://doi.org/10.1007/s10803-014-2268-6.

Bragazzi, N. L. (2013). Rethinking psychiatry with OMICS science in the age of personalized P5 medicine: ready for psychiatome? *Philosophy, ethics, and humanities in medicine: PEHM*, *8,* Article 4. https://doi.org/10.1186/1747-5341-8-4.

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

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. (1984). *Classification and Regression Trees.* Routledge.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411.

Cuartas Arias, J. M. (2017). Big data for use in psychological research. *International Journal of Psychological Research*, *10*(1), 6–7. https://doi.org/10.21500/20112084.2828.

Cuartas Arias, J. M. (2019). Homo digitalis and Contemporary Psychology. *International journal of psychological research*, *12*(2), 6–7. https://doi.org/10.21500/20112084.4260.

de Mello, F. L., & de Souza, S. A. (2019). Psychotherapy and Artificial Intelligence: A Proposal for Alignment. *Frontiers in psychology*, *10*, 263. https://doi.org/10.3389/fpsyg.2019.00263.

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179.

Dhall, D., Kaur, R., & Juneja, M. (2020). Machine Learning: A Review of the Algorithms and Its Applications, In P. Singh, A. Kar, Y. Singh, M. Kolekar, & S. Tanwar (eds), *Proceedings of ICRIC 2019. Lecture Notes in Electrical Engineering* (pp. 47–63). Springer. https://doi.org/10.1007/978-3-030-29407-6_5.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual review of clinical psychology*, *14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Elia, J., Arcos-Burgos, M., Bolton, K. L., Ambrosini, P. J., Berrettini, W., & Muenke, M. (2009). ADHD latent class clusters: DSM-IV subtypes and comorbidity. *Psychiatry research*, *170*(2–3), 192–198. https://doi.org/10.1016/j.psychres.2008.10.008.

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181. https://doi.org/10.1177/0959354319835322.

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., Maathuis, M. H., Moreau, Y., Murphy, S. A., Przytycka, T. M., Rebhan, M., Röst, H., Schuppert, A., Schwab, M., Spang, R., Stekhoven, D., Sun, J., Weber, A., Ziemek, D., & Zupan, B. (2018). From hype to reality: Data science enabling personalized medicine. *BMC medicine*, *16*(1), 150. https://doi.org/10.1186/s12916-018-1122-7.

Gudivada, V., Irfan, M., Fathi, E., & Rao, D. (2016). Cognitive Analytics: Going Beyond Big Data Analytics and Machine Learning. *Handbook of Statistics*, *35*, 169–205. https://doi.org/10.1016/bs.host.2016.07.010.

Harmouch, M. (2021). 17 Types of Similarity and Dissimilarity Measures Used in Data Science. https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681.

Hothorn, T. (2021). CRAN Task View: Machine Learning & Statistical Learning. https://cran.r-project.org/web/views/MachineLearning.html.

Jack, R. E., Crivelli, C., & Wheatley, T. (2018). Data-driven methods to diversify knowledge of human psychology. *Trends in Cognitive Sciences*, *22*(1), 1–5. https://doi.org/10.1016/j.tics.2017.10.002.

Jacobucci, R., & Grimm, K. J. (2020). Machine Learning and Psychological Research: The Unexplored Effect of Measurement. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467.

Jiménez-Figueroa, G., Ardila-Duarte, C., Pineda, D. A., Acosta-López, J. E., Cervantes-Henríquez, M. L., Pineda-Alhucema, W., Cervantes-Gutiérrez, J., Quintero-Ibarra, M., Sánchez-Rojas, M., Vélez, J. I., & Puentes-Rozo, P. J. (2017). Prepotent response inhibition and reaction times in children with attention deficit/hyperactivity disorder from a Caribbean community. *Attention deficit and hyperactivity disorders*, *9*(4), 199–211. https://doi.org/10.1007/s12402-017-0223-z.

Joliffe, I. T., & Morgan, B. J. (1992). Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, *1*(1), 69–95. https://doi.org/10.1177/096228029200100105.

Joyner, M. J., & Paneth, N. (2019). Promises, promises, and precision medicine. *The Journal of clinical investigation*, *129*(3), 946–948. https://doi.org/10.1172/JCI126119.

King University. (2019). The Major Branches of Psychology. https://online.king.edu/news/major-branches-of-psychology-guide/.

Koul, A., Becchio, C., & Cavallo, A. (2018). PredPsych: A toolbox for predictive machine learning-based approach in experimental psychology research. *Behavior research methods*, *50*(4), 1657–1672. https://doi.org/10.3758/s13428-017-0987-2.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. http://dx.doi.org/10.18637/jss.v028.i05.

Kuhn, M. (2020). "Package 'caret' – Classification and Regression Training." In *R Package Version 6.0–86*. ftp://mirrors.ucr.ac.cr/cran/web/packages/caret/caret.pdf.

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, 57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003.

Lin, E., Lin, C. H., & Lane, H. Y. (2020). Precision Psychiatry Applications with Pharmacogenomics: Artificial Intelligence and Machine Learning Approaches. *International journal of molecular sciences*, *21*(3), 969. https://doi.org/10.3390/ijms21030969.

Loftus, G. R. (1996). Psychology Will Be a Much Better Science When We Change the Way We Analyze Data. *Current Directions in Psychological Science*, *5*(6), 161–171. https://doi.org/10.1111/1467-8721.ep11512376.

Luxton, D. D. (2016). An Introduction to Artificial Intelligence in Behavioral and Mental Health Care. In *Artificial Intelligence in Behavioral and Mental Health Care* (pp. 1–26). Academic Press.

Mabry, P. L. (2011). Making sense of the data explosion: The promise of systems science. *American journal of preventive medicine*, *40*(5 Suppl 2), S159–S161. https://doi.org/10.1016/j.amepre.2011.02.001.

Mandinach, E. (2012). A Perfect Time for Data Use: Using Data-Driven Decision Making to Inform Practice. *Educational Psychologist*, *47*, 71–85. https://doi.org/10.1080/00461520.2012.667064.

Mead, A. (1992). Review of the Development of Multidimensional Scaling Methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *41*(1), 27–39. https://doi.org/10.2307/2348634.

Na, K. (2019). Prediction of future cognitive impairment among the community elderly: A machine-learning based approach. *Scientific Reports*, *9*, Article 3335. https://doi.org/10.1038/s41598-019-39478-7.

Neth, H. (2021a). Data Science for Psychologists. https://bookdown.org/hneth/ds4psy/.

Neth, H. (2021b). Ds4psy: Data Science for Psychologists. https://cran.r-project.org/web/packages/ds4psy/index.html.

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine Learning in Psychometrics and Psychological Research. *Frontiers in psychology*, *10*, 2970. https://doi.org/10.3389/fpsyg.2019.02970.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pérez-Gracia, J. L., Gúrpide, A., Ruiz-Ilundain, M. G., Alfaro Alegría, C., Colomer, R., García-Foncillas, J., & Melero Bermejo, I. (2010). Selection of extreme phenotypes: The role of clinical observation in translational research. *Clinical & translational oncology: official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, *12*(3), 174–180. https://doi.org/10.1007/s12094-010-0487-7.

Provost, F., & Fawcett, T. (2013). *Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking* (First edition). O'Reilly Media, Inc.

Python Software Foundation. (2021). What is python? executive summary. https://www.python.org/doc/essays/blurb/.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303.

Ritchie, P. L.-J., & Grenier, J. (2003). "Branches of Psychology." In *Encyclopedia of Life Support Systems*. EOLSS.

Roman, V. (2019). Unsupervised Machine Learning: Clustering Analysis. *Towards Data Science*. https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e.

Rosenfeld, A., Zuckerman, I., Azaria, A., & Kraus, S. (2012). Combining Psychological Models with Machine Learning to Better Predict People's Decisions. *Synthese*, *189*(1), 667–679.

Scikit-learn Project. (2021). Clustering Performance Evaluation. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation.

Shatte, A., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological medicine*, *49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151.

Suarez, I., De Los Reyes Aragón, C., Diaz, E., Iglesias, T., Barcelo, E., Velez, J. I., & Casini, L. (2020). How Is Temporal Processing Affected in Children with Attention-deficit/hyperactivity Disorder? *Developmental neuropsychology*, *45*(4), 246–261. https://doi.org/10.1080/87565641.2020.1764566.

Tolle, K. M., Tansley, D. S. W., & Hey, A. J. G. H. (2011). The fourth paradigm: Data-intensive scientific discovery. *Proceedings of the IEEE*, *99*(8), 1334–1337.

Tweney, R. D. (2003). Wilhelm Wundt in History: The Making of a Scientific Psychology. *Journal of the History of the Behavioral Sciences*, *39*(3), 299–300. https://doi.org/10.1002/jhbs.10127.

van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/gwydion-1/oldfiles/oldfiles/python/doc/ref.ps.

Vidal, J., O. M.and Acosta-Reyes, Padilla, J., Navarro-Lechuga, E., Bravo, E., Viasus, D., Arcos-Burgos, M., & Vélez, J. I. (2020). Chikungunya outbreak (2015) in the Colombian Caribbean: Latent classes and gender differences in virus infection. *PLoS neglected tropical diseases*, *14*(6), e0008281. https://doi.org/10.1371/journal.pntd.0008281.

Wani, A. H., Aiello, A., Kim, G., Xue, F., Martin, C., Ratanatharathorn, A., Qu, A., Koenen, K., Galea, S., Wildman, D., & Uddin, M. (2020). The Impact of Psychopathology, Social Adversity and Stress-relevant DNAm on Prospective Risk for Post-traumatic Stress: A Machine Learning Approach. *Journal of Affective Disorders*, *282*, 894–905. https://doi.org/10.1016/j.jad.2020.12.076.

Youn, Y. C., Choi, S. H., Shin, H. W., Kim, K. W., Jang, J. W., Jung, J. J., Hsiung, G. R., & Kim, S. (2018). Detection of cognitive impairment using a machine-learning algorithm. *Neuropsychiatric disease and treatment*, *14*, 2939–2945. https://doi.org/10.2147/NDT.S171950.

Yu, C., Arcos-Burgos, M., Licinio, J., & Wong, M. L. (2017). A latent genetic subtype of major depression identified by whole-exome genotyping data in a Mexican-American cohort. *Translational psychiatry*, *7*(5), e1134. https://doi.org/10.1038/tp.2017.102.

Yu, C., Baune, B. T., Fu, K. A., Wong, M. L., & Licinio, J. (2018). Genetic clustering of depressed patients and normal controls based on single-nucleotide variant proportion. *Journal of affective disorders*, *227*, 450–454. https://doi.org/10.1016/j.jad.2017.11.023.

Zhu, Y., Zhong, N., & Xiong, Y. (2009). Data Explosion, Data Nature and Dataology. In N. Zhong, K. Li, S. Lu, L. Chen (eds), *Brain Informatics. BI 2009. Lecture Notes in Computer Science* (pp. 147–158). Springer. https://doi.org/10.1007/978-3-642-04954-5_25.