List of changes to Thesis

Analysis of Systems' Performance in Competitions Challenges

December 16, 2024

Dr. Hugo Jair Escalante

Comment 1.1. *Título "Competitions challenges" se lee raro.*

Response 1.1. Se propone modificar el título de la tesis para que sea *"Analysis of systems' performance in supervised learning challenges"*

Comment 1.2. *Se podrían agregar algunos detalles al caption de la Fig 1?, explicando la intuición o flujo de info en el diagrama*

Response 1.2. Se modificó el caption añadiendo la siguiente descripción *From left to right, the flow moves from the desired reality, obtaining a reality sample until competitors receive the training and validation data (with labels) and the test data (without labels). From right to left, the path follows the predictions for evaluation, where organizers compare them with the gold standard.*

Comment 1.3. *Después de los objetivos y antes del outline se podría agregar una sección resumiendo el trabajo realizado y resaltadndo los principales hallazgos de la tesis.*

Response 1.3. Se insertó una seccción titulada **Summary of the Work and Main Findings** con el siguiente contenido:

This thesis addresses the challenges present in algorithmic competitions, proposing novel methodologies and tools for accurately comparing algorithms under competitive conditions. Through an exhaustive review of current evaluation methods, this work identifies limitations in traditional approaches when applied to competition settings. The methods proposed in this research consider the unique dynamics of competitive scenarios, allowing for fair and robust algorithm comparisons.

The results obtained in this thesis highlight the advantages of the proposed methodologies, demonstrating significant improvements in the accuracy and adaptability of algorithm performance evaluation in competitions. These findings contribute to developing a more solid framework for competition evaluation, with positive implications for future research and the design of competitions in data science and machine learning.

Comment 1.4. *Pág. 12: por qué en las pruebas sugeridas por Dietterich se requiere tener acceso al algoritmo?, es así? (ya vi que se dice más adelante, quizá sea bueno comentar algo en la pág. 12)*

Response 1.4. Se sustituyó el segundo y cuarto párrafos de la sección 1.2, que originalmente eran

Dietterich [21] reviews various statistical tests to determine algorithm performance differences. The paper discusses five closely related statistical tests for assessing whether one learning algorithm outperforms another in specific learning tasks. These tests include the paired t-test, cross-validated t-test, and McNemar's test, among others. However, these tests require access to the underlying

algorithm. In a competition scenario, there is only access to the predictions, not the algorithms. This limitation makes it challenging to apply these tests directly in competitive environments where only prediction results are available.

...

Despite their strengths, these tests are limited by the need for access to the internal mechanics of the algorithms. In competitive scenarios, only the output predictions are typically available, rendering these tests less applicable. Therefore, while Dietterich's work provides a solid foundation for algorithm comparison, it underscores the necessity for developing methods that can operate effectively under the constraints of competition frameworks.

por

Dietterich [21] reviews various statistical tests to determine algorithm performance differences. The paper discusses five closely related statistical tests for assessing whether one learning algorithm outperforms another in specific learning tasks. These tests include the Resampled Paired t-Test, the K-Fold Cross-Validated t-Test, and the 5×2 *Cross-Validated Paired t-Test. However, these tests require access to the underlying algorithm because they repeatedly split the dataset for training and prediction (depending on the test) to assess algorithm variability. In a competition scenario, there is only access to the predictions, not the algorithms. This limitation makes it challenging to apply these tests directly in competitive environments where only prediction results are available.*

...

Despite their strengths, these tests are limited by the need for access to the internal mechanics of the algorithms. Access to the algorithm is necessary for conducting these tests because the algorithms need to be run on various training and test data subsets to assess performance variability accurately. This helps capture the variability in error rates across different data samples, which is essential for valid statistical comparisons. In competitive scenarios, where only prediction results are available, alternative approaches are needed, as the absence of algorithm access hinders the application of these statistical tests effectively. Therefore, while Dietterich's work provides a solid foundation for algorithm comparison, it underscores the necessity for developing methods that can operate effectively under the constraints of competition frameworks.

Comment 1.5. *Segunda oración Sec. 1.3: They emphasize*

Response 1.5. Se modificó "They emphasizes" a "They emphasize"

Comment 1.6. *Missing reference, justo antes de inciar Sec. 2.4.2, y justo después de iniciar la sección*

Response 1.6. Ya se corrigió la referencia

Dr. Eric Téllez Ávila

Comment 2.1. *La introducción puede confundir al lector, ya que se enfoca en crowdsourcing y cualquiera puede pensar que es el tema principal*

Response 2.1. Se modificó el texto introductorio de la sección "Context and Background".

Decía:

"The growing popularity of crowdsourcing, along with its ability to leverage collective knowledge and skills, has opened new avenues for solving complex problems. As we explore this concept further, it is crucial to understand the underlying mechanisms and challenges involved. In the following sections, we will dive deeper into specific examples, methodologies, and crowdsourcing applications, highlighting their relevance across various domains."

y se cambió a:

"Building upon the concepts of crowdsourcing and academic competitions, the next chapter delves into the methodologies and structures that define modern challenges. Exploring the critical aspects of problem definition, dataset selection, and evaluation metrics provides a comprehensive framework for organizing and analyzing competitions. This chapter highlights the significance of fair and robust evaluations, essential for fostering innovation and advancing research in competitive environments."

Además se cambió la subsección "Crowdsourcing and Challenges". El nuevo texto de esta subsección es:

"Crowdsourcing, popularized by Jeff Howe in 2006, involves leveraging a geographically dispersed group to achieve common goals such as innovation, problem-solving, or efficiency. While historically rooted, it has gained prominence in academia, business, and humanitarian efforts.

Key examples include Wikipedia for collaborative content creation, Amazon Mechanical Turk for micro-tasks, and the Galaxy Zoo project, which accelerates scientific research through public participation [14, 39, 46].

Benefits include cost efficiency, speed, and access to diverse expertise, making it invaluable for tasks requiring scalability and creativity. However, ensuring quality, coordinating contributors, and addressing ethical concerns like fair compensation and privacy remain critical [62, 32]. "

Comment 2.2. *¿Es adecuado el nombre "methodology of organizing challenges" en el primer párrafo de la subsección "Methodology" de la Introducción?*

Response 2.2. Se sustituyó el texto del primer párrafo de "Methodology" que orginalmente es :

"The methodology of organizing challenges involves several critical steps. Figure **??** illustrates the challenge scheme."

y se sustituyó por :

"A classical challenge scheme involves several critical steps. Figure **??** illustrates the challenge scheme, showcasing the flow of processes in a typical competition framework. From left to right, the flow moves from the reality or problem to solve, obtaining a reality sample until competitors receive the training and validation data (with labels) and the test data (without labels). From right to left, the path follows the predictions submitted by the competitors for evaluation, where organizers compare them with the gold standard. This systematic flow ensures fairness and rigor

in assessing algorithmic performance."

Comment 2.3. *En la subsubsección "estructura" debería citarse el apéndice de métricas*

Response 2.3. Se incluyó la cita al apéndice

Comment 2.4. *En el primer párrafo de la página 5 es necesario incluir una referencia de la afirmación*

Response 2.4. Se incluyó la referencia solicitada

Comment 2.5. *En la página 9 al final se indica "falta una hoja o faltan capítulos listados"*

Response 2.5. Ya se corrigió asegurándose que estén listados una breve descripción del contenido de los capítulos de la tessis.

Comment 2.6. *Corregir en la sección "Evaluation Metrics and Framework" dice Transparency and reproducibility are critical themes throughout the paper, corregir la palabra "paper"*

Response 2.6. Se corrigió, ahora dice *Transparency and reproducibility are critical themes throughout the thesis, underpinning the methods, analyses, and conclusions presented.*

Comment 2.7. *En la sección "Approaches for Competition Scenarios" ¿a qué se refiere "a Bayessian approach"?*

Response 2.7. Se sustituyó el párrafo que originalmente dice

Lacoste et al. [44] introduce a Bayesian approach for comparing machine learning algorithms on single and multiple datasets. This method provides a probabilistic framework for evaluating algorithm performance differences, accounting for uncertainty and variability across datasets. The authors emphasize that their framework requires access to the algorithms, not merely their predictions. This is crucial because their approach involves modeling the algorithms' internal behavior and decision-making process, which predictions cannot capture alone. The Bayesian approach involves modeling the performance of algorithms as random variables and using Bayesian inference to estimate performance differences. This method offers a robust way to incorporate prior knowledge and quantify uncertainty in performance comparisons.

por

Lacoste et al. [44] introduce a Bayesian approach for comparing machine learning algorithms on single and multiple datasets. This method evaluates performance differences probabilistically, explicitly modeling uncertainty and variability across datasets. Unlike frequentist methods, which

rely on fixed hypothesis tests and binary outcomes, Bayesian methods provide a more flexible and interpretable framework. The authors emphasize that their framework requires access to the algorithms, not merely their predictions, as it involves modeling the algorithms' internal behavior and decision-making processes. By treating algorithm performance as random variables and using Bayesian inference, this approach offers a robust way to incorporate prior knowledge and quantify uncertainty in performance comparisons.

Comment 2.8. *En la página 33 hay que corregir la referencia de Breiman (2001) pues aparece como [?]*

Response 2.8. Se corrigieron las referencias

Comment 2.9. *En la página 41 falta indicar que es "OffendMEX"*

Response 2.9. Se añadió la siguiente oración al final del párrafo: *OffendMEX is a dataset comprising samples of offensive, aggressive, and vulgar text in Mexican Spanish, primarily collected from Twitter.*

Comment 2.10. *En el capítulo 3, corregir el espacio entre caracteres de los nombres de los competidores de MeOffendEs*

Response 2.10. Se corrigieron los nombres que estaban como matemáticas y se pusieron en formato texto itálico

Comment 2.11. *El capítulo cuatro inicia con "Chapter 4 centers on ... "*

Response 2.11. Se cambió el texto y ahora inicia "This chapter centers on ..."

Comment 2.12. *Cambiar el nombre de la sección 4.1 de "Competitions Comparison" a "Comparison of competitions"*

Response 2.12. Se realizó el cambio pues es más formal y adecuado para contextos académicos. De la misma forma se cambió en nombre del capítulo.

Comment 2.13. *En el capítulo cuatro se abusa del uso de los términos competitive, challenges, competitive, foster*

Response 2.13. Se reparafrasearon los dos primeros párrafos de este capítulo. El primer párrafo decía :

"This chapter centers on the comparative analysis of competitive challenges across various

fields. Competitions, whether in academic, professional, or community-based contexts, serve as vital platforms for fostering innovation and identifying top performers. This chapter aims to explore the impact and structure of these challenges, with a focus on how they drive excellence and creativity among participants. By examining different competitive frameworks, this chapter provides insights into the role of competition in advancing methodologies, enhancing participant skills, and promoting the development of novel solutions. The comparison of competitions is essential for understanding their significance and for improving future competitive frameworks."

y see sustituyó por:

"This chapter centers on the comparative analysis of competitive challenges in various fields. Whether in academic, professional, or community contexts, they serve as vital platforms and are crucial to innovation and identifying top performers. The objective is to explore their structure and impact, focusing on how they drive excellence and creativity among participants. By examining different competitive frameworks, this chapter provides insights into the role of competition in advancing methodologies, enhancing participant skills, and promoting the development of novel solutions. Comparison of competitions is essential for understanding their significance and improving future competitive frameworks."

De igual manera la primera sección de este capítulo se reescribió. Originalmente era :

"Competitive challenges are prevalent across various fields, from academic competitions to professional and community-based contests. These challenges are designed to identify the best performers and foster excellence and innovation among participants. According to Escalante (2023), competitive challenges push participants to enhance their abilities and achieve higher standards [26]. Similarly, Egele (2024) notes that such challenges can lead to significant personal and professional growth by motivating individuals to push their boundaries and engage in continuous improvement [25]."

ahora es:

"Competitions are widespread across diverse domains, ranging from academia to professional and community settings. These events aim to recognize excellence and inspire participants to achieve their best. As noted by Escalante (2023), they encourage individuals to enhance their abilities and pursue higher standards [26]. Similarly, Egele (2024) highlights their role in fostering personal and professional growth, pushing participants to expand their boundaries and continually improve [25]."

Comment 2.14. *a que se refiere "These metrics help understand how closely matched the participants are and how intense the competition is."*

Response 2.14. se reaparafraseó para dar claridad a este párrafo, ahora dice "These indicators provide insight into the level of parity among participants and how closely their performances align."

Comment 2.15. *En la bibliografía, a que se refiere la nota "Accedido" en la primer referencia*

Response 2.15. Se quitó la nota de la fecha en que se accedió el sitio web

Dra. Helena Gómez Adorno

Comment 3.1. *Falta 1 capítulo donde desarrolle sobre la implementación del esquema de evaluación propuesto ya que del capítulo 3 que es el marco teorico se salta al capítulo 4 que describe ya los resultados de la comparación de los resultados de competencias. Recomiendo agregar un capítulo que describa el desarrollo realizado.*

Response 3.1. Se incluyó un capítulo titulado *Implementation of the Proposed Evaluation Framework* para solventar este punto. Este capítulo está formado por elementos de los capítulos *Theory Framework* y *Performance Comparison in Challenge Scheme*. Además se adecuo el contenido de estos capítulos.

Comment 3.2. *Vale la pena mencionar en las conclusiones sobre como se podrías usar este mismo esquema para validar modelos de clasificación durante el proceso de desarrollo de dichos modelos*

Response 3.2. Se incluyó una sección en las conclusions titulada **Versatile Framework for Systematic Evaluation in Competitions and Beyond** que describe lo solicitado.

Dr. Aldo Marquez Grajales

Comment 4.1. *Sugiero incluir la hipótesis o preguntas de investigación en la introducción*

Response 4.1. Se incluyó en el capítulo *Introduction* una sección titulada *Research Hypotheses and Questions* para incluir las hipótesis y preguntas de investigación de la tesis.

Comment 4.2. *Sugiero, cambiar el nombre del capítulo 2, ya que por el nombre, esperaba que se describiera la metodología usada de tu propuesta.*

Response 4.2. Se corrigió el nombre en la portada del capítulo 2 que decía **Methodology** y debía decir **Theory Framework**

Comment 4.3. *Referencia rota justo antes de inciar Sec. 2.4.2, y justo después de iniciar la sección*

Response 4.3. Ya se corrigió la referencia

Comment 4.4. *Hace falta indicar cual es la propuesta de evaluación, ya que no está claro en el documento.*

Response 4.4. Se incluyó un capítulo titulado *Implementation of the Proposed Evaluation Framework* para solventar este punto

Comment 4.5. *En la sección* ***Measuring Competitiveness in Challenges****. Asumo que estas son las métricas propuestas. Sin embargo, no está claro en el documento. Faltan más métricas que están indicadas en la Tabla 4.2*

Response 4.5. Se modificó la sección para incluir solamente las métricas propuestas.

Sergio Nava

Comment 5.1. *Incluir en el apéndice el caso multimétrica*

Comment 5.2. *Revisar que cada capítulo tenga* ***Introduction*** *y* ***Summary***

Comment 5.3. *Mejorar el glosario*

Comment 5.4. *Mejorar los captions de figuras y tablas*

References

[1] Kaggle: Your machine learning and data science community, 2010.

[2] Agerri, R., Centeno, R., Espinosa, M., de Landa, J. F., and Rodrigo, Á. VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection. *Procesamiento del Lenguaje Natural 67*, 0 (sep 2021), 173–181.

[3] Álvarez-Carmona, M. Á., Aranda, R., Arce-Cardenas, S., Fajardo-Delgado, D., Guerrero-Rodríguez, R., López-Monroy, A. P., Martínez-Miranda, J., Pérez-Espinosa, H., and Rodríguez-González, A. Y. Overview of Rest-Mex at IberLEF 2021: Recommendation System for Text Mexican Tourism. *Procesamiento del Lenguaje Natural 67*, 0 (sep 2021), 163–172.

[4] Álvarez-Carmona, M. Á., Díaz-Pacheco, Á., Aranda, R., Rodríguez-González, A. Y., Fajardo-Delgado, D., Guerrero-Rodríguez, R., and Bustio-Martínez, L. Overview of Rest-Mex at IberLEF 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts. *Procesamiento del Lenguaje Natural 69*, 0 (sep 2022), 289–299.

[5] Aragón, M. E., Álvarez-Carmona, M., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., and Moctezuma, D. Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. *CEUR Workshop Proceedings 2421* (2019), 478–494.

[6] Arlot, S., and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics Surveys 4* (2010), 40–79.

[7] Bel-Enguix, G., Sierra, G., Gómez-Adorno, H., Torres-Moreno, J.-M., Ortiz-Barajas, J.-G., and Vásquez, J. Overview of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task. *Procesamiento del Lenguaje Natural 69*, 0 (sep 2022), 255–263.

[8] Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57* (1 1995), 289–300.

[9] Berg-Kirkpatrick, T., Burkett, D., and Klein, D. An empirical investigation of statistical significance in NLP. In *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference* (2012).

[10] Bergmeir, C., and Benitez, J. M. A note on the validity of cross-validation for evaluating time series prediction. *Computational Statistics & Data Analysis 120* (2018), 70–83.

[11] Bisani, M., and Ney, H. Bootstrap estimates for confidence intervals in ASR performance evaluation. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing 1* (2004).

[12] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 ed. Springer, 2007.

[13] Bonferroni, C. E. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[14] Brabham, D. C. *Crowdsourcing*. MIT Press, 2013.

[15] Breiman, L. Random forests. 5–32.

[16] Chai, T., and Draxler, R. R. Root mean square error (rmse) or mean absolute error (mae)? -arguments against avoiding rmse in the literature. *Geoscientific Model Development 7* (6 2014), 1247–1250.

[17] Chernick, M. R., and LaBudde, R. A. *An introduction to bootstrap methods with applications to R*. Wiley, 2011.

[18] Davison, A. C., and Hinkley, D. V. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[19] De Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. Mean absolute percentage error for regression models. *Neurocomputing 192* (2016), 38–48.

[20] Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7*, 1 (2006), 1–30.

[21] Dietterich, T. G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation 10*, 7 (oct 1998), 1895–1923.

[22] Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association 56*, 293 (1961), 52–64.

[23] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *https://doi.org/10.1214/aos/1176344552 7*, 1 (jan 1979), 1–26.

[24] Efron, B., and Tibshirani, R. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, may 1994.

[25] Egele, R., Junior, J. C. S. J., van Rijn, J. N., Guyon, I., Baró, X., Clapés, A., Balaprakash, P., Escalera, S., Moeslund, T., and Wan, J. Ai competitions and benchmarks: Dataset development, 2024.

[26] Escalante, H. J., and Kruchinina, A. Academic competitions, 2023.

[27] Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis. 88*, 2 (2010), 303–338.

[28] Fawcett, T. An introduction to roc analysis. *Pattern recognition letters 27*, 8 (2006), 861–874.

[29] Friedman, J. H., Hastie, T., and Tibshirani, R. The elements of statistical learning. *Springer series in statistics 1*, 10 (2001), 10.

[30] García, S., and Herrera, F. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research 9* (2008), 2677–2694.

[31] García-Vega, M., Díaz-Galiano, M. C., García-Cumbreras, M., Del Arco, F. M. P., Montejo-Ráez, A., Jiménez-Zafra, S. M., Cámara, E. M., Aguilar, C. A., Cabezudo, M. A. S., Chiruzzo, L., and Moctezuma, D. Overview of TASS 2020: Introducing Emotion Detection. *CEUR Workshop Proceedings 2664* (2020), 163–170.

[32] Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. Managing the crowd: Towards a taxonomy of crowdsourcing processes. *Proceedings of the Seventeenth Americas Conference on Information Systems 2011* (2011), 1–11.

[33] Good, P. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Science & Business Media, 2013.

[34] Good, P. I. *Introduction to statistics through resampling methods and R/S-Plus*. John Wiley & Sons, 2005.

[35] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

[36] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. Result analysis of the nips 2003 feature selection challenge. *Advances in Neural Information Processing Systems 17* (2004), 545–552.

[37] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[38] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics 6* (1979), 65–70.

[39] Howe, J. The rise of crowdsourcing. *Wired magazine 14*, 6 (2006), 1–4.

[40] Jafari, M., and Ansari-Pour, N. Why, When and How to Adjust Your P Values? *Cell Journal (Yakhteh) 20*, 4 (2019), 604.

[41] Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 - A meeting of SIGDAT, a Special Interest Group of the ACL held in conjunction with ACL 2004* (2004), pp. 388–395.

[42] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM 60*, 6 (2012), 84–90.

[43] Labatut, V., and Cherifi, H. Accuracy measures for the comparison of classifiers, 2012.

[44] Lacoste, A., Laviolette, F., and Marchand, M. Bayesian comparison of machine learning algorithms on single and multiple datasets, 2012.

[45] Lavesson, N., and Davidsson, P. Evaluating learning algorithms and classifiers. *Intelligent Information Systems 2*, 4 (2006), 37–52.

[46] Lintott, C. J., et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society 389*, 3 (2008), 1179–1189.

[47] Meinshausen, N., and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72* (9 2010), 417–473.

[48] Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[49] Nava-Muñoz, S., Graff Guerrero, M., and Escalante, H. J. Comparison of classifiers in challenge scheme. In *Pattern Recognition* (Cham, 2023), A. Y. Rodríguez-González, H. Pérez-Espinosa, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. A. Olvera-López, Eds., Springer Nature Switzerland, pp. 89–98.

[50] Nava-Muñoz, S., Graff, M., and Escalante, H. J. Analysis of systems' performance in natural language processing competitions. *Pattern Recognition Letters* (3 2024).

[51] Ojala, M., and Garriga, G. C. Permutation tests for studying classifier performance. *Journal of Machine Learning Research 11* (2010), 1833–1863.

[52] Olson, R. S., Bartley, N., Urbanowicz, R. J., and Moore, J. H. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining 10*, 1 (2017), 1–13.

[53] Pavao, A., Guyon, I., Letournel, A.-C., Tran, D.-T., Baro, X., Escalante, H. J., Escalera, S., Thomas, T., and Xu, Z. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research 24*, 198 (2023), 1–6.

[54] Plaza-del-Arco, F. M., Casavantes, M., Escalante, H. J., Martín-Valdivia, M. T., Montejo-Ráez, A., Montes-y Gómez, M., Jarquín-Vásquez, H., and Villaseñor-Pineda, L. Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants. *Procesamiento del Lenguaje Natural 67*, 0 (sep 2021), 183–194.

[55] Plevris, V., Solorzano, G., Bakas, N. P., and Seghier, M. E. A. B. Investigation of performance metrics in regression analysis and machine learning-based prediction models. *ECCOMAS Congress 2022 - 8th European Congress on Computational Methods in Applied Sciences and Engineering* (11 2022).

[56] Powers, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2011).

[57] Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning, 2020.

[58] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., and Donoso, T. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural 67*, 0 (sep 2021), 195–207.

[59] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis. 115*, 3 (2015), 211–252.

[60] Søgaard, A., Johannsen, A., Plank, B., Hovy, D., and Martinez, H. What's in a p-value in NLP? *CoNLL 2014 - 18th Conference on Computational Natural Language Learning, Proceedings* (2014), 1–10.

[61] Sokolova, M., and Lapalme, G. A survey of performance evaluation measures for classification systems. *Information Processing & Management 45*, 4 (2009), 427–437.

[62] Surowiecki, J. *The wisdom of crowds*. Anchor Books, 2005.

[63] Taulé, M., Ariza, A., Nofre, M., Amigó, E., and Rosso, P. Overview of DETOXIS at IberLEF 2021: DEtection of TOXicity in comments In Spanish. *Procesamiento del Lenguaje Natural 67*, 0 (sep 2021), 209–221.

[64] Vapnik, V., and Cortes, C. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[65] Wainer, J. A bayesian bradley-terry model to compare multiple ml algorithms on multiple data sets. *arXiv preprint arXiv:2208.04935* (2016).

[66] Willmott, C. J., and Matsuura, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research 30*, 1 (2005), 79–82.

[67] Zhang, Y., Vogel, S., and Waibel, A. Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System? In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004* (2004), pp. 2051–2054.