PRO HOMINUM
REVISTA DE CIENCIAS SOCIALES Y HUMANAS

Artículo original

Using classification and prediction algorithms to process surveys in the Altair® RapidMiner data mining software

Uso de algoritmos de clasificación y predicción para procesar encuestas en el software de minería de datos Altair® RapidMiner

https://doi.org/10.47606/ACVEN/PH0319

Gabriela Pazmiño Moreira[1]
https://orcid.org/0000-0002-8443-9189
zetagabriela@gmail.com

Olga Lilian Mendoza Talledo[2]
https://orcid.org/0000-0001-6053-562X
Olga.mendoza@ut.edu.ec

Homero Mendoza Rodríguez[2]
https://orcid.org/0009-0004-9049-4207
Homero.mendoza@utm.edu.ec

María José Pazmiño Moreira[1]
https://orcid.org/0000-0001-6035-4909

Jonathan Josué Proaño Morales[2*]
https://orcid.org/0000-0002-7140-5318
Jonathan.proano@utm.edu.ec

Recibido: 08/09/2024 **Aceptado:** 20/12/2024

ABSTRACT

Currently, analyzing and processing surveys to look for patterns has become essential to solve and find new product distribution strategies. For this reason, surveys have been analyzed using classification and prediction algorithms to look for patterns. To do this, the free software Altair® RapidMiner Studio Version 2024 was applied to the information extracted from surveys carried out through Google Forms and distributed online throughout the country. The surveys consisted of 30 questions, most of which were multiple choice. AdaBoost, Naive Bayes and deep learning algorithms were used to classify, analyze and find patterns between the questions. Thus, the vaccines used varied according to the age groups and the media in which the advertisements were shown. In conclusion, this tool is considered easy to use due to its simplicity, as it offers algorithms that allow accurate classification and prediction of surveys, as well as the search and visualization of patterns.

Keywords: Data; productivity; big data; marketing; population behavior.

1. Department of data processing and survey Design, prodata & Desing, 130103- Ecuador
2. Universidad Técnica de Manabí- Ecuador
* Autor de correspondencia: Jonathan.proano@utm.edu.ec

RESUMEN

En la actualidad, analizar y procesar encuestas para buscar patrones se ha convertido en algo esencial para resolver y encontrar nuevas estrategias de distribución de productos. Por esta razón, se han analizado encuestas utilizando algoritmos de clasificación y predicción para buscar patrones. Para ello, se aplicó el software libre Altair® RapidMiner Studio Versión 2024 a la información extraída de las encuestas realizadas a través de Google Forms y distribuidas online por todo el territorio nacional. Las encuestas constaban de 30 preguntas, la mayoría de opción múltiple. Se utilizaron algoritmos AdaBoost, Naive Bayes y deep learning para clasificar, analizar y encontrar patrones entre las preguntas. Así, las vacunas utilizadas variaron en función de los grupos de edad y de los medios en los que se mostraron los anuncios. En conclusión, esta herramienta se considera fácil de usar debido a su sencillez, ya que ofrece algoritmos que permiten una clasificación y predicción precisa de las encuestas, así como la búsqueda y visualización de patrones.

Palabras clave: Datos; Productividad; big data; Marketing; Comportamiento demográfico.

INTRODUCTION

Information analysis has evolved and there are more and more different platforms or programs that provide different services and deliver analysis results in large amounts, including Altair® RapidMiner Studio. It is an open source software for data mining and predictive analysis. Its results are available in XML files and are distributed under the AGPL license. It provides the development of information analysis processes by chaining operators through a graphical environment and helps to create predictive models. It is a cross-platform tool that runs in different environments such as Windows, Linux or Mac. In addition, it allows the application of algorithms from the Weka software (Fernández-Morales & Bonilla-Carrión, 2020).

Depending on the type of mining analysis, it provides many models that are quite comprehensive, such as Bayesian models, modeling, tree induction, neural networks, among others, as well as various classification methods, clustering, associations, and more. It is a popular analysis platform that provides graphical interfaces for creating data analysis workflows (Vidiya & Testiana, 2023).

The University of Chile analyzed student dropout rates and applied decision trees to take corrective and timely actions to improve university education. Similar cases were also found in a Dutch university, an institute in New Zealand, the University of India, and in research conducted in Latin America (Ramírez & Grandón, 2018).

The company Vidiya & Testiana (2023) considers Altair® RapidMiner a very useful tool for processing and managing their data, due to the patterns that can be used to improve marketing strategies, optimize product inventory and increase

customer satisfaction, using the FP growth algorithm implemented through the platform.

Private higher education institutions in Mexico studied the situation of their teaching methods and low enrollment through data mining (Estrada-Danell et al., 2016). In addition, a fraud detection process was developed that includes tools with algorithms that focus on both local and global levels to improve the accuracy of the solution and allow for the investigation of unusual patterns. Anomalous data were found in domains such as credit cards, security systems, and electronic health information, which led to the demonstration that the proposed method in Altair® RapidMiner is highly capable of detecting all introduced outliers (Orellana & Cedillo, 2020).

Understanding a large amount of data to determine how complex the user information process is on the Internet, in order to identify the natural activities of users on social media platforms, can be applied through various data mining research approaches such as clustering, classification, regression, among others. Social media platforms cannot function without users, as they allow the evaluation of critical data based on favorable or unfavorable interpretations of the collected information. Part of the process has been carried out to discover the nature of consumers through simple algorithmic experiments (Sandeep, V., & Vindhya, A. S. 2023). On the other hand, questionnaire design often proves to be imperfect, so that the structuring of the questions is insufficient for the respondent's understanding, leading to doubts. In order to contribute to the quality of the analysis, paradata were used (Fernández-Fontelo et al., 2023).

The most common difficulties in processing surveys arise from off-topic responses or unanswered questions. Over time, efforts have been made to advance surveys to save time and increase response rates. Colloquial response alternatives have been used to reduce database errors, with significant results, but they did not fully cover the margin of error (Ongena & Unger, 2020). To address this issue, fuzzy noise removal models applied to vision were employed, incorporating both theoretical and practical contributions to the field. Three generic diffusion modeling frameworks based on probabilistic noise elimination diffusion models, noise-conditioned scoring networks, and stochastic differential equations are identified and presented (Shamshad, F. et al., 2023). In light of the aforementioned considerations, Altair® RapidMiner was deemed an appropriate tool for survey analysis employing AdaBoost, Naive Bayes, and Deep Learning algorithms to gain deeper insights and rectify analysis errors. These algorithms facilitate problem-solving and the execution of complex activities through a series of instructions. They simulate human intelligence processes through learning, reasoning, self-correction, and pattern recognition, thereby contributing to the technological development of this new era.

METHODOLOGY

Survey design and objectives

This section presents the survey processing and its algorithms, as well as their fundamentals, which allow us to describe the development of the study. Altair® RapidMiner Studio Version 2024 was employed for the analysis, and surveys were conducted using the Google Forms platform.

The survey link was distributed in WhatsApp and Instagram groups. The responses were in the form of multiple-choice questions, and the questions were divided into different topics, yet simultaneously related to social media, health, lifestyle, and interests.

Distribution and Participants for the Surveys

The surveys comprised 30 multiple-choice questions. The surveys were conducted on 166 individuals from diverse geographical regions and age groups, spanning from 18 to over 60 years of age (snowball sampling). The participants were selected from a diverse range of backgrounds, including men and women, and represented various professions, non-professional roles, and unemployed individuals. However, the objective of the survey was to obtain data for use in Rapid Miner. Consequently, the responses were randomly replicated multiple times to achieve the target of 2000 surveys. Ultimately, 2058 surveys were replicated.

Operating System

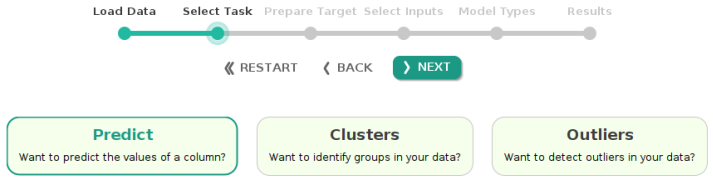
The work was conducted on a Sony Vaio laptop with 4 GB of RAM, a third-generation Intel Core i3 processor, and a 480 GB SSD, utilizing Linux Mint 21.3 x86_64 as the operating system.

Algorithms used

According to the automodel

In Altair® RapidMiner, the Auto Model can be utilized to identify and select the optimal model for analyzing the survey data. This enables the simulation of potential scenarios based on the chosen analysis methodology, whether it be predictive, classification, or the identification of missing values. In the aforementioned case, the selected vaccine was used to predict outcomes, as illustrated in Figure 1.

Figure 1.



Selection of the predictor variable.

In the performance analysis of the models, the model with the lowest classification error was the Decision Tree and the Support Vector Machine, both with a classification error of 0.00 and a standard deviation of 0.00, indicating highly accurate and consistent performance. The Decision Tree also demonstrated the greatest gains, with a value of 434.00, while the Support Vector Machine exhibited gains of 442.00.

In terms of total processing time, the Decision Tree exhibited notable efficiency, requiring only 2.31 seconds, whereas the Support Vector Machine required 15.31 seconds. In contrast, the Generalized Linear Model exhibited the poorest performance, with a classification error of 0.27 and a standard deviation of 0.03, along with gains of only 134.00. The model also exhibited relatively slow processing times, with a total runtime of 2.44 seconds. Other models, such as Gradient Boosted Trees and Random Forest, demonstrated favorable outcomes with classification errors of 0.02 and 0.03, respectively, and gains of 446.00 and 420.00. However, they exhibited slower processing times compared to the Decision Tree and the Support Vector Machine (Table 1).

Table 1.
Visualization of the results obtained by each model

Model	Classification Error	Standard Deviation	Gains	Total Time (min)	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Naive Bayes	0.22	0.01	194	2.5	0.48	2
Generalized Linear Model	0.27	0.03	134	2.44	8	2
Logistic Regression	0.19	0.02	204	7.37	4	6
Fast Large Margin	0.11	0.02	334	22.22	6	16
Deep Learning	0.12	0.03	298	3.17	3	2
Decision Tree	0	0	434	2.31	0.3	2
Random Forest	0.03	0.01	420	28.5	0.53	17
Gradient Boosted Trees	0.02	0.01	446	11.28	2	5
Support Vector Machine	0	0	442	15.31	2	9

The rationale behind the deployment of Bayes and deep learning techniques

Upon examination of the models proposed by Auto Model, it became evident that Naive Bayes is capable of accurate predictions even in the absence of complete data classification. In contrast, the classification trees were excluded due to the ambiguous interpretation of their structure.

Similarly, the Deep Learning model demonstrated remarkable proficiency in classifying the five classes, exhibiting minimal error metrics and a perfect confusion matrix.

This indicates that the model has effectively discerned the data characteristics and is adept at making precise predictions. Moreover, these models were seamlessly integrated with the AdaBoost algorithm.

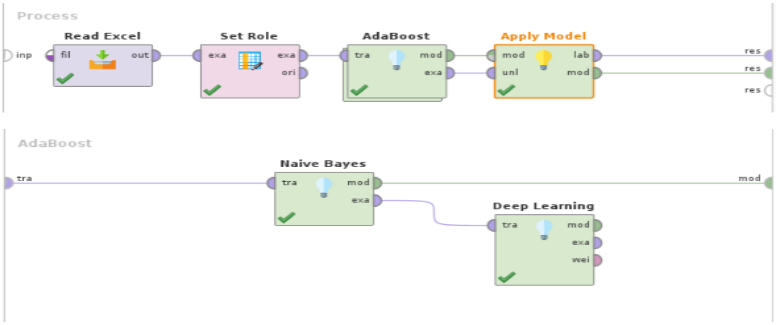
The algorithms utilized in this study was as follows:

AdaBoost, developed by Freund and Schapire in 1996, provides minimal error rates and is known for its simplicity, speed, and ease of implementation, resulting in improved accuracy (Wang et al., 2019). It takes a positive approach to decision making and provides generic techniques that can be applied to any basic classification method. AdaBoost works by selectively resampling training data to generate inferred subsets (Dong et al., 2015). In Altair® RapidMiner it works as a box algorithm. Naive Bayes learns from training data and predicts the class of the test instance, providing better categorization accuracy on real-time datasets than other classifiers and requiring a small fraction of training data (Mosquera, Castrillón & Parra, 2018). It achieves an accuracy of about 82% with increasing dataset size (Liu et al., 2013). Geoffrey Hinton deepened artificial intelligence in 2006 by explaining deep neural networks, which greatly enhanced the capabilities of the model. Over time, this algorithm has been applied to machines to simulate human learning. Deep Learning provides a comprehensive foundation for deep learning, covering aspects from network design to training, evaluation, and adaptation. Its explanations make the mathematical representation accessible to researchers in other subfields of artificial intelligence (Heaton, 2018).

Process used in Altair® RapidMiner studio

The following operators were used to load and process the data, as shown in Figure 2: To load the data, the Read Excel operator was used because the file downloaded from Google Forms was in XLSX format. Once loaded into the operator, it was connected through extensions to Set Role, where the target variable was selected (this variable can be selected depending on what you want to visualize or analyze). After selecting the variable, the next operator selected was the AdaBoost algorithm (A). Since this is a box algorithm, meaning that it can be combined with one or more algorithms, the Naive Bayes and Deep Learning algorithms (B) were added within this operator; this was done to increase the efficiency in classifying the responses. Finally, the algorithm was applied to the data using Apply Model. By combining these operators, the results were standardized between zero and one.

Figure 2.
Evolution of the data process using the algorithms: A. AdaBoost (box algorithm) B. Naive Bayes and Deep Learning.



.

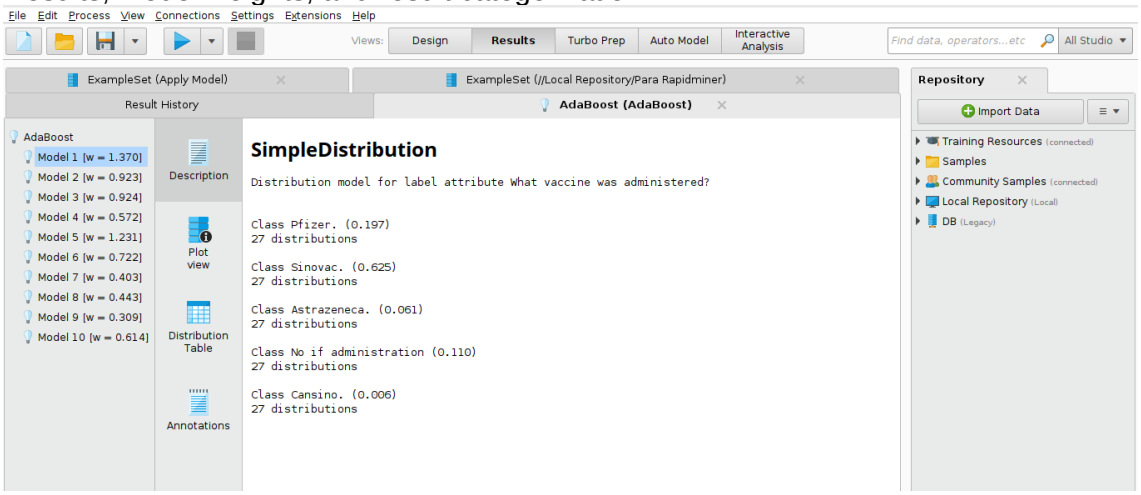
Methodology for Optimizing Writing and Translation

The ChatGPT tool was employed to refine the Spanish text, ensuring clarity and coherence. Subsequently, DeepL was utilized to perform an accurate and fluent translation of the content from Spanish to English, ensuring that the context and original style of the message were preserved.

RESULTS

Applying the aforementioned algorithms, we obtained the following results.

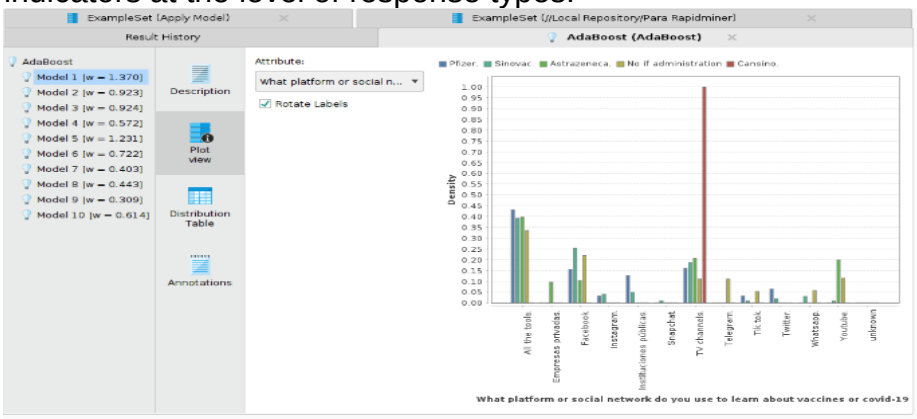
Figure 3.
Results, model weights, and result categorization



The collected data was fed into AdaBoost, resulting in Figure 3, which generated 10 prediction models. Each of these models has a different weight, with model 1 being the most accurate in processing the data according to the previous questions. Moving to Simple Charts, we see the graphs of the questions as well as

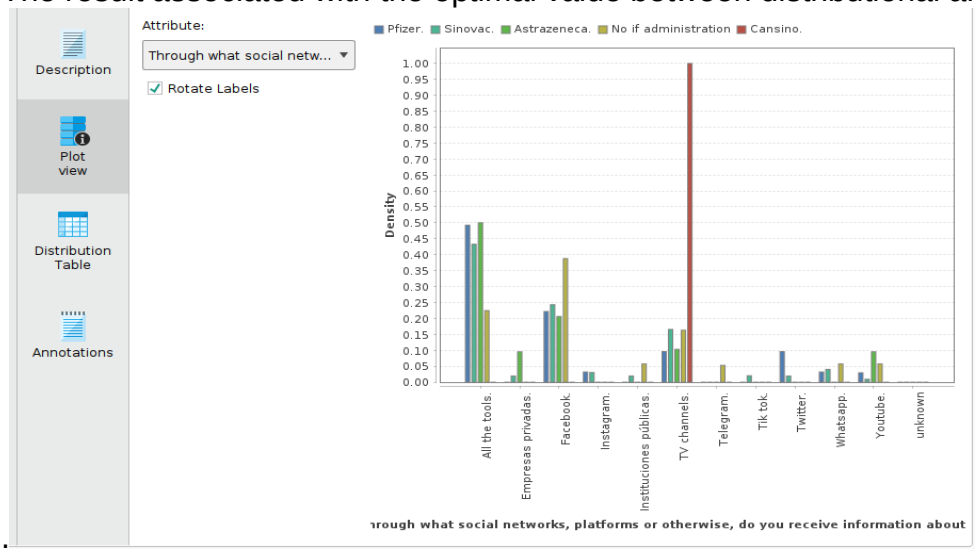
the target variable, which in this case used vaccine was administration. This resulted in a score of 0.987 and a weight of 1.370 for the first model (Figure 4).

Figure 4.
A bar graph showing the parameters obtained in the sampling process and density indicators at the level of response types.



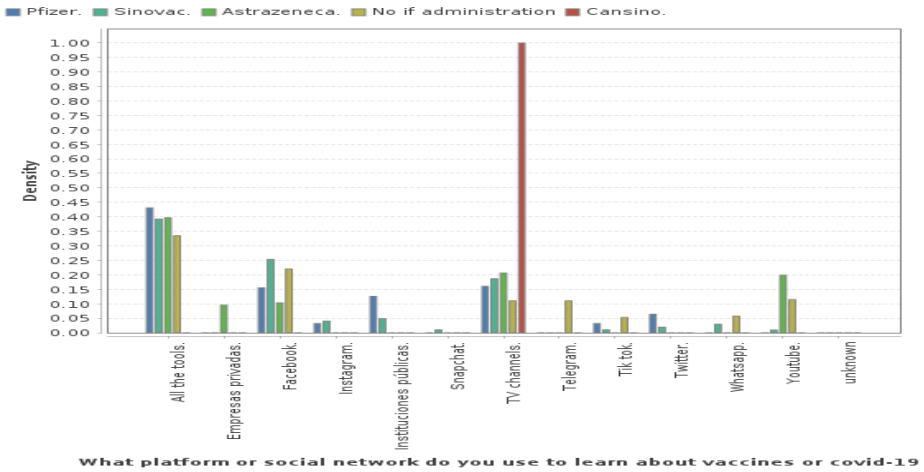
Moving to the Naive Bayes algorithm, we obtain a distribution and visualize how the algorithm classified the questions. This allowed us to process a large number of questions simultaneously without complicating the analysis, since we were working with some questions with yes, no answers and others with multiple-choice options (Figure 5).

Figure 5.
The result associated with the optimal value between distributional analyses



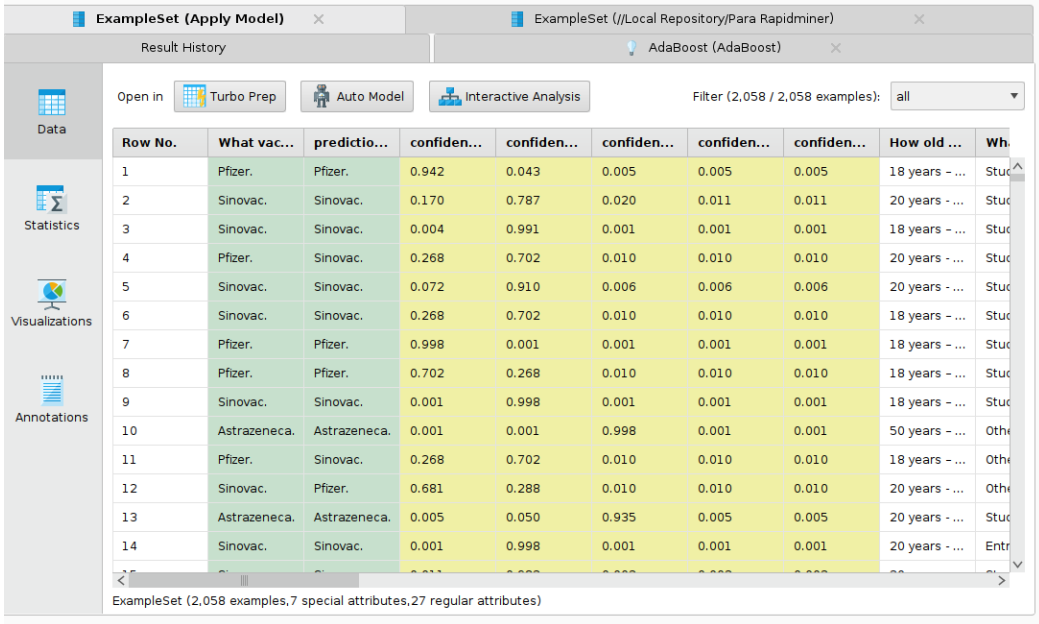
The results of each question can be exported to PNG, JPG, SVG, EPS and PDF files. They can then be visualized as follows (Figure 5).

Figure 6.
The range of analyzed values and the platforms or social network.



In 'Example Set Apply model', questions are selected and then they can be visualized in a bar graph. Here we see the selected vaccines on the x-axis and the density color-coded on the y-axis. It also shows the age ranges. On the other hand, we can see that none of the survey results were left unclassified, since they were correctly classified (Figure 6).

Figure 7.
Parameters analyzed by the optimization process of the Apply Model analysis

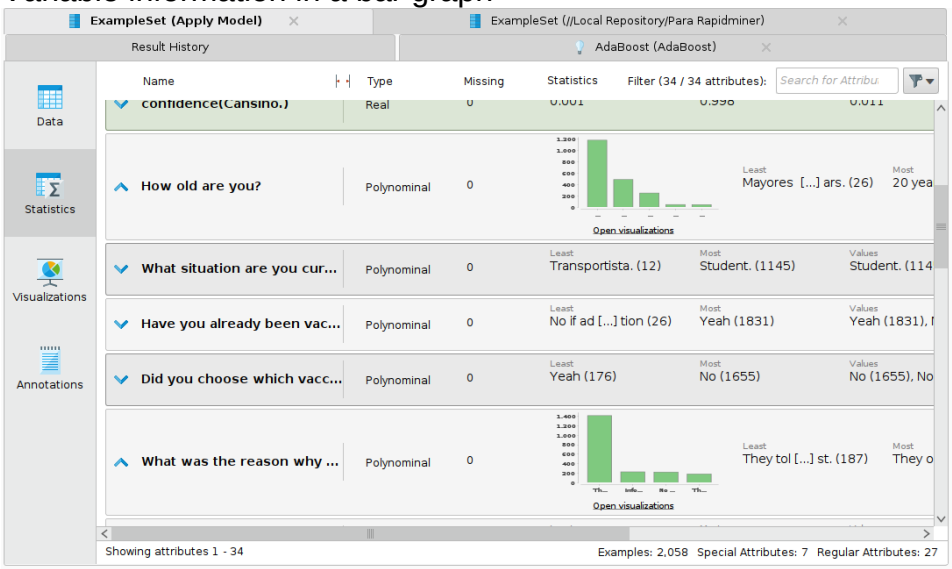


.

Another way to visualize the questions is in the example set (Apply Model), where the prediction made for each question depending on the user selected was

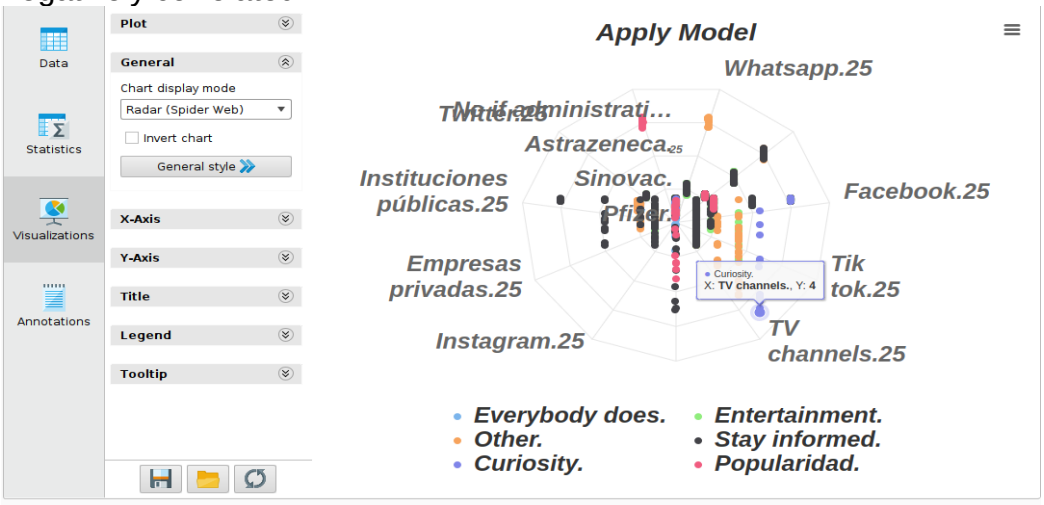
observed (Figure 7). In addition, we can observe the bar graph of the answers in tabular format and in a small histogram (Figure 8).

Figure 8.
Variable information in a bar graph



In addition, visualizations can select questions for both the x-axis and the y-axis, in this case against the target variable of age ranges. In addition, the graph type and visualization mode can be selected (Figure 9).

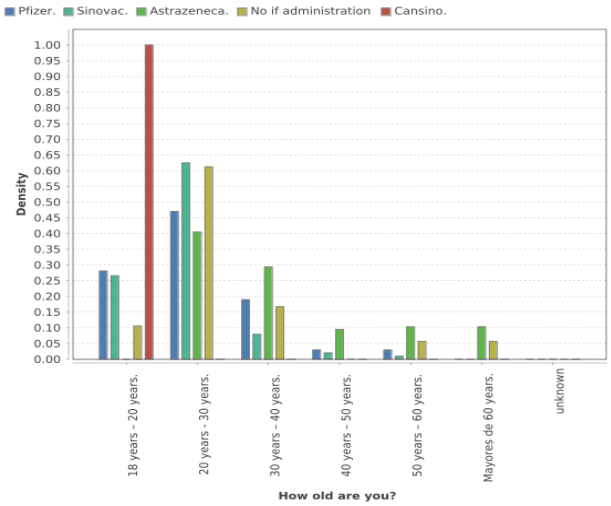
Figure 9.
Shows the procedure for selecting the variables to observe and their distribution in the radar chart, considering more than one question that are positively or negatively correlated.



Visualization of modeling results

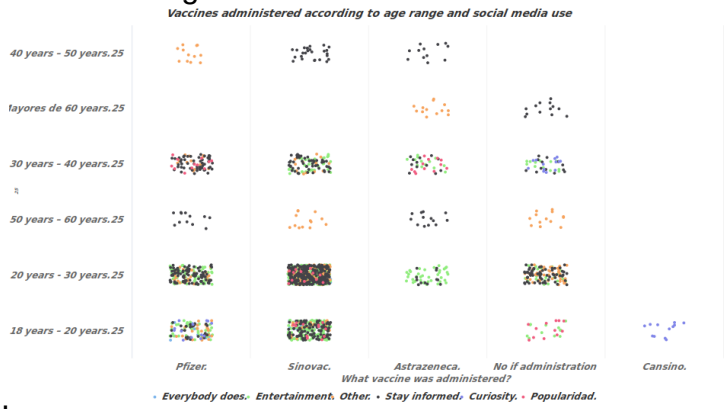
To obtain the variable graph, we export the image and select the format. In this way, we obtain the complete visualization based on the question, the age ranges and the density, with reference to the variables that were not classified, if applicable. However, they were perfectly classified, so this value is zero (figure 10).

Figure 10.
Visualization of vaccine use density by age group .



In this section, variables can be used for prediction as well as questions with their unprocessed responses, as seen in Figure 10, where we observe vaccine types on the x-axis and platforms and social networks used for information on the y-axis based on age groups. We can also see how these platforms influence vaccine choice in a scatter plot (Figure 11). In this category, there are several options for graphs such as radar charts.

Figure 11.
How data selection affects grouping. The process shows the search for data grouping across different ranges



.

Interpreting the graphs

The vaccines utilized vary contingent on the age cohort and the manner in which social media is employed. For instance, younger individuals exhibited a greater propensity to be influenced, in contrast to older demographics. The most prevalent vaccines were Sinovac and Astrazeneca, with the latter competing with Pfizer depending on the social network where the advertisement was published.

DISCUSSION

For survey analysis, Big Data Analytics and Altair® RapidMiner are excellent tools, both create an impact on decision-making and performance through techniques and good performance, ensuring the effectiveness in results. In addition, Big Data Analytics affects decision-making and generates a significant change in the research model, which is based on the vision of information processing and survey-based assessments, linking it to Big Data Analytics with the effectiveness of decision-making and performance, examining the relationship between the factor under investigation and decision-making, as well as the extent to which the factor under investigation influences decision-making, based on an adopted questionnaire and data collected from respondents, resulting in data analysis using statistics show that Big Data Analytics has a positive impact on an organization's decision-making ability and effectiveness. (Latif, A. et al., 2023).

When applying the survey processing technique through the deep learning model, which allows for better analysis based on a predefined algorithm or set of steps, in this case, deep learning is a type of machine learning. It has proven to be useful for analyzing, outlining, extracting, and detecting information modalities. DL techniques are broadly classified into segmentation and multi-step approaches, providing important insights into promising areas based on DL to effectively address new research variants and improve their development in emerging challenges (Khan et al., 2023).

On the other hand, it is important to note that Altair® RapidMiner influences the emerging area of research by improving the performance of survey studies through various techniques (Batool et al., 2023). These techniques explore hidden data relationships focusing on open source GUI with covered predictive analytics and data mining techniques: Exploratory analysis, visualization, decision trees, rule induction, k-nearest neighbors, naive Bayes, artificial neural networks, support vector machines, ensemble models, bagging, boosting, random forests, linear regression, logistic regression, association analysis using Apriori and FP-growth, K-means clustering, density-based clustering, self-organizing maps, text mining, time series prediction, anomaly detection, and feature selection, among others (Kotu & Deshpande, 2014).

In a separate study conducted by Demera H. et al. (2023), Altair® RapidMiner was utilized to analyze surveys conducted with 24 operators, 2 veterinarians, and 1 safety supervisor. In this study, the Random Forest algorithm was employed to ascertain the significance of each question, followed by the generation of a corresponding decision tree. The model achieved a score of 0.89,

indicating a high level of accuracy. Despite the survey comprising only 14 questions and therefore having limited data records, this methodology enabled the identification of significant patterns in the collected information.

With advances in data science technology, there are studies aimed at testing the effectiveness of algorithms that can help you make better decisions for any activity, using software such as Altair® RapidMiner. Using this program, the process of learning and testing data becomes faster and more efficient. Algorithms with the help of Altair® RapidMiner software can be used as an effective method of decision support, capable of providing decisions with quite high accuracy (Siregar, K. 2023).

Over time, Altair® RapidMiner has developed an intelligent learning enhancement based on the Adaboost algorithm, which was applied to lung cancer breath detection using electronic nose (ENOSE). This allowed to determine the average accuracy of the improved algorithm classifier to discriminate between people with lung cancer and healthy people, achieving an effectiveness of 98.47%, a sensitivity of 98.33%, and a specificity of 97%. In 100 independent and random tests, the coefficient of variation of the classifier's performance barely exceeded 4%. Compared to other integrated algorithms, the generalization and stability of the enhanced algorithm classifier are superior, indicating that the enhanced Adaboost algorithm can help select lung cancer more comprehensively. Furthermore, it will significantly advance the use of ENOSE in the early detection of lung cancer (Hao, L., & Huang, G. 2023).

In the same framework, Alzheimer's disease is a fatal disease that can cause dementia in its victims. Many studies have analyzed Alzheimer's disease using data mining techniques, using Altair® RapidMiner's Naive Bayes algorithm with Particle Swarm Optimization (PSO) feature selection and bagging to optimize unbalanced data. The results of the 10-fold cross-validation experiment showed that the first test using the Bayes algorithm achieved a precision value of 93.75% with an AUC value of 0.966. In addition, the test using PSO feature selection and bagging technique yielded a precision value of 98.21% with an AUC value of 0.989. From these results, it can be concluded that by using PSO feature selection and bagging techniques, the precision value increased significantly. This shows that optimizing algorithms with PSO feature selection and bagging techniques results in excellent classification (Saputra, R. et al., 2023).

By integrating big data management algorithms and object detection technologies based on automatic and deep learning algorithms, Altair® RapidMiner optimizes detection capabilities. Intelligent objects can process contextual data related to infrastructure and users through sensors and actuators to infer the environment within semi-IoRT systems and make seamless autonomous decisions. In addition, big data analytics along with data collection tools, data extraction, spatial cognition, and digital twin simulations contribute favorably together with intelligent production systems (Andronie, M. et al., 2023).

For survey analysis, there are a variety of tools for proper processing, whether open source or proprietary, such as Power BI, Tableau, SPSS, Qlikview, Python, SQL, R programming, Apache Spark, Flink, SAS, among others, which

also offer different developments. In the era of intelligent Internet, the management and analysis of massive space-time data is one of the important links for creating intelligent applications and building smart cities, where the interaction of data from multiple sources is the basis for managing and analyzing space-time data, serving as a key carrier to achieve interactive computation of massive data (Andronie, M. et al., 2023). Regarding the surveys conducted with 272 participants, which were analyzed from Qualtrics to SPSS to NVivo. Analysis of quantitative data investigating the associations between retrospective psychedelic experiences and the relationship with nature. NVivo software provides a useful way to break down large amounts of data, such as participant interviews, for subsequent code (theme) identification and requires early engagement with the data (Irvine et al., 2023).

Flink provides an advanced operator union to facilitate user program development. In a Flink job with operations that join data from multiple sources, the selection of join sequences and data communication in the repair phase are key factors that affect job efficiency. The operator allows Flink to support operations that join data from multiple sources and reduces the amount of computation and data communication by introducing lightweight optimization strategies to improve job efficiency. The join sequence optimization strategy can reduce total execution time by 29% and data communication by 34% compared to traditional sequential execution. With the data repair optimization strategy, the job can achieve a performance improvement of 35%, and in the average case, data communication can be reduced by 43% (Ji, H. et al., 2023).

Other software, such as Qlikview, facilitates data processing with robust models, providing a wide variety of dashboard development and interactive models. However, due to its pricing, it experiences slower growth (Vanegas et al., 2020). In contrast, Power BI's interface is easy to master and precise for businesses of all types, allowing for the development of analytical models and reports. In contrast, Tableau software is suitable for organizations as it is suitable for storing information on a server (Batt et al., 2020). Meanwhile, SPSS can handle large amounts of data, create tables, charts, and simultaneously analyze other formats (Rahman & Muktadir, 2021).

In the current scenario, Python programming plays a vital role in the field of research, as it converts research work into a coding format that assists researchers worldwide. The Python program was instrumental in the development of the closure and interior of subsets of the given set. This was achieved by using the dictionary in Python, with each subset serving as the key and the respective closure as the value (Prabu, M. V., & Rahini, M. 2023).

Apache Spark, like Flink, enables users to implement queries on large distributed databases using functional APIs. In recent years, these APIs have gained popularity due to their functional interfaces that abstract much of the minutiae of distributed programming required by traditional query languages such as SQL. A novel column decomposition technique has been developed to split the synthesis task into smaller sub-control synthesis problems. This approach has been implemented as a new tool, RDD2SQL, which translates Spark RDD queries into SQL and empirically evaluates the effectiveness of RDD2SQL on a set of real-

world RDD queries. The study demonstrates that the majority of RDD queries can be translated into SQL, automating this translation and offering significant performance benefits (Zhang, G. et al,. 2023).

Measurements can produce consistent results and are crucial for any scientific research measurement. The intraclass correlation coefficient (ICC) is the most widely used method to determine the reproducibility of measurements across various statistical techniques. The ICC, along with its calculated confidence interval revealing the underlying sampling distribution, can help detect the ability of an experimental method to identify systematic differences among research participants in a test. The introduction of a new SAS macro, ICC6, provides advantages for calculating different forms of ICC and their confidence intervals. This macro employs the PROC GLM procedure in SAS to generate estimates of two-way random effects ANOVA. Validation analysis using commercial software packages STATA and SPSS yielded identical results, indicating the development of SAS methodology using publicly available statistical approaches in estimating six distinct forms of ICC and their confidence intervals (Senthil Kumar, V. S., & Shahraz, S. 2023).

R and Matlab are two high-level scientific programming languages frequently applied in natural sciences such as physics, mathematics, and computational biology. New methods and applications are often implemented solely in R or solely in Matlab. Additionally, the RCall interface provides users with access to a wide variety of methods programmed in R and Matlab.

RCall runs in Matlab and provides direct access to methods and software packages implemented in R, accessible for example from Bioconductor or CRAN. The straightforward combination of Matlab and R methods significantly enhances the functionality of the Matlab programming environment, rendering RCall an optimal choice for comparative evaluations of the two programming languages (Egert, J., & Kreutz, C. 2023).

CONCLUSION

By using this tool in survey analysis, companies achieve the desired effect of finding solutions in processing their extracted information, offering a variety of different algorithms with better results. The goal is to classify and predict new strategies through patterns, accurately classify questions and generate graphs with different distribution ranges used according to survey variants. For example, adolescents and adults made the decision to choose the vaccine they would receive based on the influence of information disseminated through social networks.

Conversely, the algorithms employed demonstrated effectiveness in classifying the questions and presenting the results. The utilization of this tool and the application of these algorithms represent a significant saving of time and financial resources in the processing process.

REFERENCES

Andronie, M., Lăzăroiu, G., Iatagan, M., Hurloiu, I., Ștefănescu, R., Dijmărescu, A., & Dijmărescu, I. (2023). Big data management algorithms, deep learning-based object detection technologies, and geospatial simulation and sensor fusion tools in the Internet of Robotic Things. ISPRS International Journal of Geo-Information, 12(2), 35. https://doi.org/10.3390/ijgi12020035

Batt, S., Grealis, T., Harmon, O., & Tomolonis, P. (2020). Learning Tableau: A data visualization tool. The Journal of Economic Education, 51(3–4), 317–328. https://doi.org/10.1080/00220485.2020.1804503

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies, 28(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y

Delgado-Demera, M. H., Proaño–Morales, J. J., Delgado-Demera, M. M., Burgos–Briones, G. A., & Cedeño–Palacios, C. A. (2023). Evaluación de riesgos sanitarios en el Centro de Faenamiento Municipal de Portoviejo – Manabí, Ecuador. Revista Científica De La Facultad De Ciencias Veterinarias De La Universidad Del Zulia, 33(2), 1-7. https://doi.org/10.52973/rcfcv-e33256

Dong, Z., Fang, Y., Wang, X., Zhao, Y., & Wang, Q. (2015). Hydrophobicity classification of polymeric insulators based on embedded methods. Materials Research, 18(1), 127–137. https://doi.org/10.1590/1516-1439.286414

Egert, J., & Kreutz, C. (2023). Rcall: An R interface for MATLAB. SoftwareX, 21(101276), 101276. https://doi.org/10.1016/j.softx.2022.101276

Estrada-Danell, R. I., Zamarripa-Franco, R. A., Zúñiga-Garay, P. G., & Martínez-Trejo, I. (2016). Aportaciones desde la minería de datos al proceso de captación de matrícula en Instituciones de Educación Superior particulares. Revista Electrónica Educare, 20(3), 1. https://doi.org/10.15359/ree.20-3.11

Fernández Morales, M. E., & Bonilla Carrión, R. (2020). Bibliominería, datos y el proceso de toma de decisiones. Revista interamericana de bibliotecologia, 43(2), eI8. https://doi.org/10.17533/udea.rib.v43n2ei8

Fernández-Fontelo, A., Kieslich, P. J., Henninger, F., Kreuter, F., & Greven, S. (2023). Predicting question difficulty in web surveys: A machine learning approach based on mouse movement features. Social Science Computer Review, 41(1), 141–162. https://doi.org/10.1177/08944393211032950

Hao, L., & Huang, G. (2023). An improved AdaBoost algorithm for identification of lung cancer based on electronic nose. Heliyon, 9(3), e13633. https://doi.org/10.1016/j.heliyon.2023.e13633

Heaton, J. (2018). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618. Genetic Programming and Evolvable Machines, 19(1–2), 305–307. https://doi.org/10.1007/s10710-017-9314-z

Irvine, A., Luke, D., Harrild, F., Gandy, S., & Watts, R. (2023). Transpersonal ecodelia: Surveying psychedelically induced biophilia. Psychoactives, 2(2), 174–193. https://doi.org/10.3390/psychoactives2020012

Ji, H., Wu, G., Zhao, Y., Wang, S., Wang, G., & Yuan, G. Y. (2023). joinTree: A novel join-oriented multivariate operator for spatio-temporal data management in Flink. GeoInformatica, 27(1), 107–132. https://doi.org/10.1007/s10707-022-00470-5

Khan, A., Khan, S. H., Saif, M., Batool, A., Sohail, A., & Waleed Khan, M. (2023). A Survey of Deep Learning Techniques for the Analysis of COVID-19 and their usability for Detecting Omicron. Journal of Experimental & Theoretical Artificial Intelligence: JETAI, 1–43. https://doi.org/10.1080/0952813x.2023.2165724

Kotu, V., & Deshpande, B. (2014). Predictive Analytics and Data Mining: Concepts and practice with Altair® RapidMiner. Morgan Kaufmann. ISBN: 9780128016503

Latif, A., Fairdous, R., Akhtar, R., & Ambreen, M. (2023). Exploring the impact of Big Data Analytics on organizational decision-making and performance: Insights from Pakistan's industrial sector. Pakistan Journal of Humanities and Social Sciences, 11(2). https://doi.org/10.52131/pjhss.2023.1102.0475

Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013). Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier. 2013 IEEE International Conference on Big Data. DOI: 10.1109/BigData.2013.6691740

Mosquera, R., Castrillón, O. D., & Parra, L. (2018). Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos. CIT Informacion Tecnologica, 29(6), 153–162. https://doi.org/10.4067/s0718-07642018000600153

Ongena, Y., & Unger, S. (2020). The effects of task difficulty and conversational cueing on answer formatting problems in surveys. In Advances in Questionnaire Design, Development, Evaluation and Testing (pp. 259–286). Wiley. https://doi.org/10.1002/9781119263685.ch11

Orellana Cordero, M. P., & Cedillo, P. (2020). Detección de valores atípicos con técnicas de minería de datos y métodos estadísticos. Enfoque UTE, 11(1), 56–67. https://doi.org/10.29019/enfoque.v11n1.584

Prabu, M. V., & Rahini, M. (2023). Application of Kuratowski's closure operator in Python program. 5th International Conference On Current Scenario In Pure And Applied Mathematics (ICCSPAM-2022). https://doi.org/10.1063/5.0137779

Rahman, A., & Muktadir, M. G. (2021). SPSS: An imperative quantitative data analysis tool for social science research. International Journal of Research and Innovation in Social Science, 05(10), 300–302. https://doi.org/10.47772/ijriss.2021.51012

Ramírez, P. E., & Grandón, E. E. (2018). Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en

Árboles de Decisión con Parámetros Optimizados. Formación Universitaria, 11(3), 3–10. https://doi.org/10.4067/s0718-50062018000300003

Sandeep, V., & Vindhya, A. S. (2023). Lack of accuracy in ascertaining nature of users based on Naive Bayes algorithm comparing K-means algorithm. The 6th International Conference On Energy, Environment, Epidemiology And Information System (ICENIS) 2021: Topic of Energy, Environment, Epidemiology, and Information System. https://doi.org/10.1063/5.0124446

Saputra, R. A., Puspitasari, D., Wahyudi, M., Ramdhani, L. S., & Ramanda, K. (2023). Optimization the Naive Bayes algorithm using particle swarm optimization feature selection and bagging techniques for detection of Alzheimer's disease. AIP Conference Proceedings. https://doi.org/10.1063/5.0128553

Senthil Kumar, V. S., & Shahraz, S. (2023). Intraclass correlation for reliability assessment: the introduction of a validated program in SAS (ICC6). Health Services & Outcomes Research Methodology. https://doi.org/10.1007/s10742-023-00299-x

Siregar, K. (2023). Testing the c4.5 algorithm with Rapid Miner to determine decisions for implementing sports activities. INFOKUM, 11(04), 40–47. https://doi.org/10.58471/infokum.v11i04.1790

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical Image Analysis, 88(102802), 102802. https://doi.org/10.1016/j.media.2023.102802

Vanegas, D. A., Tarazona Bermudez, G. M., & Rodriguez Rojas, L. A. (2020). Mejora de la toma de decisiones en ciclo de ventas del subsistema comercial de servicios en una empresa de IT. Revista Científica, 38(2), 174–183. https://doi.org/10.14483/23448350.15241

Vidiya, E. C., & Testiana, G. (2023). Analisis Pola Pembelian di Lathansa Cafe & Ramen dengan Menggunakan Algoritma FP-Growth Berbantuan Altair® RapidMiner. G-Tech: Jurnal Teknologi Terapan, 7(3), 1118–1126. https://doi.org/10.33379/gtech.v7i3.2739

Wang, J., He, Z., Ji, J., Zhao, K., & Zhang, H. (2019). IoT-based measurement system for classifying cow behavior from tri-axial accelerometer. Ciencia Rural, 49(6). https://doi.org/10.1590/0103-8478cr20180627

Zhang, G., Mariano, B., Shen, X., & Dillig, I. (2023). Automated translation of functional big data queries to SQL. Proceedings of the ACM on Programming Languages, 7(OOPSLA1), 580–608. https://doi.org/10.1145/3586047