International Journal of AI & Data Science Review
Volume 1, Issue 2, 2026

Correlation Filters in Machine Learning Algorithms to Choose Personal and Demographic Characteristics for the Diagnosis of Autism Spectrum Disorder

Ciro Rucco[1]

Laboratory of Intelligent Computing and Robotics, Federal Institute of Triangulo Mineiro Campus Patrocínio, Brazil

Article Info

Received: 28-12-2025 Revised:10-02-2026 Accepted:29-03-2026 Published:04-04-2026

Abstract: Currently regarded as one of the primary neurodevelopmental diseases, autism spectrum disorder is characterized by restricted and repetitive patterns, as well as difficulties with social communication and cognitive abilities. There is no known treatment for this illness, which varies in severity according on how each patient's symptoms manifest. The absence of effective screening methods that are easy to use is one of the factors contributing to the often long wait times for autism spectrum disorder diagnoses. The goal of this effort is to identify patterns of the autism spectrum by analyzing a public database; in other words, to identify the traits that, when combined with behavioral traits, might increase the precursor model's reliability. The probabilistic neural network approach did well in this categorization, according to the first findings. Furthermore, the use of correlation filters showed increased accuracy and efficiency. We achieved 100% accuracy using the support vector machine by using eight data mining techniques, combining the individual, behavioral, and demographic factors, and eliminating certain features. Lastly, machine learning studies have shown that a patient's ethnicity, continent, and presence of jaundice seem to indicate a higher likelihood of receiving an autistic spectrum disorder diagnosis.

Keywords: artificial intelligence, machine learning, autism spectrum disorder, data mining, diagnosis in neurodevelopment

1. Introduction

According to Vieira and Baldin (2017), autism spectrum disorder (ASD) is a neurological condition marked by notable changes in language and social interaction development rather than an illness. According to Siegel (2021) and Stone-Heaberlin et al. (2022), among other writers, autism is categorized as a pervasive developmental disorder (PDD). Communication and social skills issues are among its behavioral indicators, with varied degrees of difficulty. A strong propensity for routines, intolerance to change, and an extreme connection to certain things are just a few of the repetitive and constrained patterns of behavior that people with autism may display. According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)1, autism spectrum disorder (ASD) may be divided into three different categories according to the degree of assistance needed by the person since it includes a variety of presentations. These levels are as follows, according Chagas (2016): Level 1 denotes mild ASD and calls for assistance; Level 2 denotes moderate ASD and calls for significant assistance; and Level 3 denotes severe ASD and calls for considerable assistance. Scientific research on autism has significantly increased in a number of areas in recent years with the goal of learning more about the disorder's nature as well as potential treatment options (Zeidan et al., 2022). This rise has aided in the creation of more rapid and efficient techniques for diagnosis and identification. Although many instances have not yet been identified, it is believed that 1.5% of the global population has some kind of ASD. Leo Kanner, a scientist and psychiatrist, coined the term "autism" in 1943 to describe the clinical circumstances of eleven children who shared characteristics such as a language development disorder, a strong desire to maintain the same environment, and some rejections when interacting with others (Kanner, 1965). Since then, efforts have been undertaken to compile the behaviors and symptoms of kids with autism diagnoses in an effort to standardize them using universal language (Goulart & de Assis, 2002). The American Psychiatric Association (APA) publishes the DSM, which reflects the evolving diagnostic standards for autism. Even though there is no cure for autism, there are treatments that must be administered after a diagnosis. Depending on the severity of the patient's syndrome, these treatments may or may not be complex, and they must be carried out by a team of professionals that includes doctors, speech therapists, psychologists, and physical educators, among others. However, it is important to keep in mind that since they are chronic patients, this therapeutic vision will last for extended periods of time, necessitating ongoing observation from the experts engaged in order to have a thorough understanding of the problem's dimensions (Assumpção & Pimentel, 2000). After undergoing a number of changes over time, the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) now refers to "autism" as autism spectrum disorder (ASD). ASD is used to characterize a variety of neurodevelopmental diseases characterized by repetitive and restricted behavior patterns, as well as limits in social engagement and reciprocal social communication. Asperger's syndrome, childhood disintegrative disorder, and nonspecific pervasive developmental disorder are among the subtypes of these diseases that the word came to refer to (Klin, 2006). Researchers (Thabtah et al., 2022) have been using effective machine learning (ML) techniques for a while (Carvalho et al., 2019; Garai et al., 2023; Garai & Paul, 2023; Lopes & Lima, 2021) in an effort to improve the ASD diagnosis method by increasing the speed and accuracy of symptom analysis so that

patients can receive treatment more quickly. Nowadays, it is thought that there is a lot of available data that can outperform manual analysis. The development of more sophisticated algorithms that, when linked to databases, enable them to conduct research and searches in a more comprehensive and intricate manner is another current factor. These considerations led to an increase in the applications of data science, which is related to other concepts like big data and data-driven decision making, both of which are very significant. It is also known that the number of ASD cases is increasing quickly worldwide, and developing an automated model to do the classification that determines whether a case is ASD or not is one of the challenges in the diagnostic process. A variety of programs have already been created to assist children with ASD in various age groups in the classroom (Farias & Cunha, 2013; Sousa et al., 2012); other programs concentrate on developing software for the diagnosis of individuals with ASD (Crane et al., 2016; Davis et al., 2007). Nevertheless, there aren't many studies that use data mining to establish connections with pertinent traits for application development. After learning that the process of diagnosing autism involves categorization, it is anticipated to be able to identify cases of ASD by analyzing and evaluating examples received via an input dataset comprising behavioral features. Using a dataset from the UCI Machine Learning Repository, data science methods will be used to categorize ASD behavior. By integrating this information with the open-source KNIME Analytics Platform, useful insights may be extracted via the production of analytics and reports. This paper's main goal is to undertake a consolidated data treatment and extract the most relevant attributes for a more accurate ASD diagnosis using various machine learning approaches. In this research, answers from ASD tests of children, adolescents, and adults are analyzed from three extensive databases totaling 1100 cases. There are 20 qualities in each dataset, 10 of which are related to behavioral traits and the other 10 to individual features that have been shown to be beneficial in diagnosing ASD patients. The data mining was done using the KNIME program once the data was organized into ARFF files. Eight supervised learning approaches were used in a thorough investigation to compare and assess the findings. It is crucial to remember that the main goal of this research was not to create software, but rather to identify and create a model that can correctly categorize patterns that may help identify ASD. The goal of this study is to drastically cut down on the amount of time patients must wait for a diagnosis. There are five portions in this work. The introduction and the work's goals are presented in the first part. The Background, which includes some key definitions of the work and related works, is included in the second part. The third portion contains the work proposal, which shows how models, schemes, and supervised learning algorithms are implemented in the data mining tool used here. The findings are analyzed and presented in the fourth part, which also includes a commentary on the current work. The work's conclusion and future directions are covered in the fifth part.

2. Background

In this section, we will provide a comprehensive theoretical framework that will serve as a guide for readers to grasp the main concepts essential to comprehend the data mining model. Initially, we will delve into the fundamental concepts of ASD, including an exploration of the varying levels of autism and its core characteristics. Following that, we will introduce the database, which houses the data utilized in this study. Subsequently, we will shed light on ML, highlighting some of the most widely employed tools used for data analysis. Lastly, we will present a review of relevant previous works centered around applications designed for the early detection of ASD.

2.1. Autism spectrum disorder

ASD is a developmental brain disorder that includes difficulties that limit certain social and communication behaviors of natural growth (Thabtah et al., 2022). Although the patient's symptoms portray the beginning of the disorder, it is important to pay attention to the fact that the severity varies according to the manifestation of these symptoms and that even if they are reduced and have a change in their prognosis, from procedures performed early, there is no cure for ASD. Although ASD originated in the first years of life, its initial trajectory is not the same. In some children, for example, symptoms are noticed soon after birth, but in most cases the symptoms are only identified between 12 and 24 months (Thabtah et al., 2022). However, the diagnosis of ASD is performed at an average of 4–5 years of age. This is a deplorable condition considering that the measures taken to carry out early treatment have enormous gains in terms of the child's cognitive and adaptive functioning and may even prevent the complete manifestation of ASD, as it occurs in a period of development in which the brain is quite malleable (Thabtah et al., 2022).

In recent years, estimates of the prevalence of autism have increased dramatically. One of the causes of this increase is the result of the expansion of diagnostic criteria and the development of screening and diagnosis instruments with adequate psychometric properties (Thabtah et al., 2022). There are currently clinical and non-clinical diagnostic procedures for ASD (Thabtah et al., 2022); the Autism Diagnostic Interview (ADI) and the Revised Autism Diagnostic Observation Schedule (ADOS-R) are examples of clinical procedures. Non-clinical procedures are also used, such as the Social Communication Questionnaire and the Autism Quotient Trait (AQ).

Patients with ASD have specific needs according to their actions and behaviors; therefore, it requires a specific therapeutic analysis that allows the indication of an individualized treatment strategy (Thabtah et al., 2022). The patient's therapeutic process is usually initiated by the pediatrician, who is also responsible for referring other professionals who will resort to

interdisciplinary therapeutic procedures (Thabtah, 2017). Usually, patients with autism also need psychopharmacological treatment to control symptoms related to the medical condition, when these interfere negatively with their quality of life (Biswas et al., 2021).

2.2. Data science in classification

Data are facts collected and normally stored; however, only when some analysis technique is applied, and they are able to produce information (de Carvalho et al., 2017; de Morais et al., 2020; Lima et al., 2021). Data Science is the term used to define the extraction of this information and is also composed of specialized skill sets, such as programming, techniques such as predictive analysis, data mining, and visualization, among others. The purpose is to extract information efficiently and transform them into knowledge that will later help formulate actions that generate results, even helping in decision making (de Carvalho et al., 2017). Machine Learning (ML) is a subfield of artificial intelligence (AI) that enables computers to recognize patterns and uncover hidden information without being explicitly programmed for it. ML algorithms have the ability to learn from data, meaning they can identify and extract meaningful insights without being explicitly instructed on what to look for (Guandaline & de Campos Merschmann, 2017). Among the existing types of ML, two are considered as the main ones: the supervised and the unsupervised. The first ML's method involves training the computer using a labeled dataset, which is the most commonly used approach. In contrast, the second method involves detecting patterns in unlabeled data, where the computer must make sense of the data without historical labels (classes).

Data Mining (DM) is a branch of computing that began in the 1980s, when professionals in companies and organizations began to worry about the large volumes of computer data stored and unused within the company (Lima et al., 2021). Currently, data mining consists mainly in the practice of analyzing data after being collected, aiming to produce new extremely useful information that is more significant than the original dataset. Knowing how to differentiate a task from a mining technique is essential. The task is equivalent to identifying what one intends to look for in the data, such as the type of predictability or which sets of patterns are useful to see. On the other hand, the mining technique is equivalent to identifying procedures that certify the exploration of the patterns that really matter. Among the most important techniques used for data mining is the machine learning technique (Lima et al., 2021).

2.3. Related apps for ASD diagnosis

This section highlights and critically reviews available autism screening tests developed for the mobile environment. As autism is a complex disorder that compromises social and communication skills, affected children (Eder et al., 2016) have different behaviors and mental states, which encourages the development of software with adaptive interfaces (Fletcher-Watson et al., 2016) for them (Sousa et al., 2012). Furthermore, this section has been enhanced by the inclusion of educational applications (Aziz et al., 2014), as well as awareness and non-screening apps. The functionalities of current autism apps are also briefly explained, and their strengths and weaknesses identified. AaB app: AaB was developed in the USA, which is a screening app created in the USA by researchers based on video to assist parents in identifying and the mental health challenges of their children who may be related to autism (Thabtah et al., 2022).

Naturalistic Observation Diagnostic Assessment app: Naturalistic Observation Diagnostic Assessment (NODA) is not a screening application; it is grounded in medical research and was created to give parents appropriate guidance on the characteristics of autism, which was developed by Floreo Tech in the USA and account with a 2017 update (Thabtah, 2017). Technology company Floreo Tech is led by CEO and co-founder Vijay Ravindran.

ASDTests: ASDTests is a screening application that includes four quizzes containing 10 questions based on the age groups of users developed in New Zealand. In addition, the application collects the patient's demographic, geographic, and health data. If the user presents a value (7) as the final result of the questions, it means that the classification of this user is as having autism. This was the application we chose to collect the data, that is, the database we collected refers to the data collected by Thabtah et al. (2022) over these years. The application is available to download for mobile phones with Apple's iOS operating system. More specifically, it is a questionnaire for babies, children, teenagers, and adults (Thabtah et al., 2022).

Asperger test app: The app was developed at Cambridge University, UK, which was developed for adults to identify Asperger traits or autism[2] traits (Allison et al., 2012). Although the AT app doe[3]s not provide a formal diagnosis, it does provide guidance for users to understand their socially related behaviors. Currently, two versions of the app with 50 objective questions are available: one for adults/teenagers and one for children. Some users found the questions too strategic in the sense of inducing the user.

Autism and developmental disorder screening app: or shortly ANDDS. The ANDDS was developed in the United States to assess the risk of ASD in children aged 6 months to children under 3 years. The app progressively assesses the child (6, 12, 15, 18, 24 to 36 months) through a series of Boolean questions (yes/no). Therefore, the application developed by behavioral scientists excludes most of society. Interestingly, the app displays the result by color bands and ANDDS is not a

popular screening test among users, however, because it has no rating, the app was created in 2005 by Suzanne and Bob Wright who founded Autism Speaks[4] because her grandson is diagnosed with autism.

Autism test app: The software developed in Croatia called Autism Test Application (ATA) is a self-assessment application on autism and other psychological issues related to the challenges that adults and, in some cases, parents may face when performing the test on behalf of a child. Its educational foundation through 20 questions (3 possible alternatives) to identify symptoms of autism in adults and does not serve for diagnosis. The application is developed by Consurgo, rated 2.2 in the app download stores, however the methodology used to design the questions was unclear. According to Thabtah (2017), it is unclear what methodology is used to design the questions and whether it is based on known and published screening methods.

ASDetect app: the ASDetect application developed in Australia by La Trobe University in partnership with Olga Tennison Autism Center (Thabtah, 2017). The authors use a series of videos combined with quizzes for children. The questionnaire assesses social and behavioral characteristics and can also be taken by parents and caregivers. The tests are not so short, as they take about 2030 min to complete, and at the end the possibility of the child between (12, 18 and 24) months old is shown to have autism or not, with an accuracy of 81%. The application managed to obtain 4.5 in the ratings ranking with 33 reviews, with very positive reviews.

Indian scale for assessing autism: One of the methods used to assess an individual's level of autism is the Indian Scale for Assessing Autism (ISAA) (Chakraborty et al., 2015). An application for Android 4.2 developed in India by CDAC Hyderabad, based on ISAA[5] with English and Hindi versions was developed to help parents, particularly in India, assess their child's level of autism by answering 40 questions with alternatives covering different areas of autism. After each user navigates through a large number of screens they will receive their final score and corresponding autism rate. According to Thabtah (2017), any score below 76 shows no autism traits, while any score above 153 exemplifies severe traits of autism. Intermediate scores exhibit different levels of autism.

Autism AI: The first work by Thabtah (2017), called ADSTests, presents an application developed for the Apple Store and has 10 questions about characteristics of autism and more questions about some geographic, demographic or even data. However, in order to further improve the application, efforts by the authors Thabtah et al. (2022) and Thabtah (2017) were presented in an attempt to use machine learning to improve the accuracy of the previous application. In this case, the authors from New Zealand developed a first improvement (Thabtah et al., 2022) and a second improvement (Shahamiri & Thabtah, 2020) in the United States. These improvements were based on artificial intelligence (AI) to enhance the initial ADSTests assessment. The ADSTests test was previously limited to considering only answers with values greater than an integer. Another application proposed by Allison et al. (2012) originated from the United States.

Worldtour: the Brazilian application called Worldtour published Sousa et al. (2012) proposes a software to support the cognitive development of autistic children through playful activities that involve planning. According to the authors, the software is in the prototype stage and considers the Human-Computer Interface (HCI) recommendations recommended for children. Aprende con Zapo: the application was developed in Spain by the authors Lozano-Martínez et al. (2011) to demonstrate a situation in which the application can support the teaching-learning process in primary and secondary students with ASD. Participants had difficulties in recognizing emotional states; thus, according to the authors, there were nine students aged between 8 and 18 years. The teaching process was developed in two weekly sessions of 45 min each, with an interval of two academic years, and then they were retested. According to Lozano-Martínez et al. (2011), the results obtained confirm that the use of educational software in the teaching of emotional and social skills helps students to improve their skills and overcome tasks aimed at understanding emotional and social skills, being observed by teachers and family members.

TEACCH prototype: the application prototype developed in Brazil based on the TEACCH methodology. The authors Farias and Cunha (2013) present the prototype of a software that helps in the literacy of autistic children, seeking to streamline the use of the TEACCH methodology, currently applied manually, and already internalized by children who receive some kind of monitoring.

ATEC autism signals: is an app Autism Treatment Evaluation Checklist (ATEC) proposto por Zakhar Lobanov[6] to check for signs of autism and the dynamics of autism rehabilitation and contains 77 questions that take between 5 and 10 min to complete the test. It is divided into 4 parts: (a) speech/language/ communication, (b) sociality, (c) tact/cognitive skills, and (d) health/physical development/behavior. It has an option to accompany several children simultaneously and has a comment section. At the end, a graph and tables are generated to monitor the evolution of individuals.

Autism screening test adult and child: The application was developed by the company Inquiry Health LLC[7],

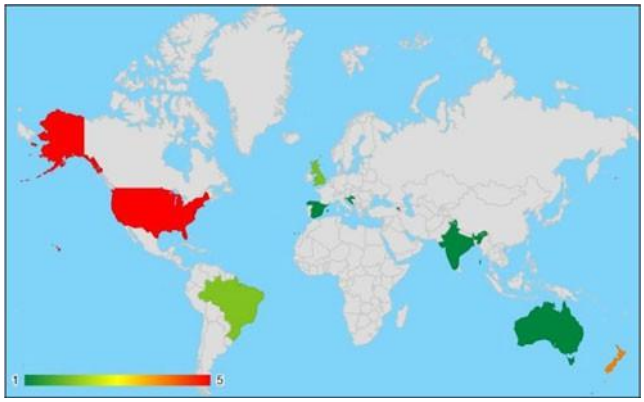
located in the United States of America (USA). It was last updated in August 2022 and has garnered over 500 downloads on Android. According to the developers, the data is not shared with third parties, and no data is collected.

Autism Quiz: AQ-10 quotient: A simple app to automate the taking of the AQ-10 test, which is used by the National Healthcare Service in the UK as a quick reference guide for general practitioners to use. According to author Theodore Tollet[8], the test should be used to determine whether someone should be recommended for a specialist autism assessment and is not apt for use as a diagnostic tool on its own.

Awesomely autistic test: The Autism Quotient (AQ) Test is a simple app developed in London[9] for Android. It has gained popularity with over 10,000 downloads and 157 reviews, earning an average rating of 4.5 stars. The test was developed by Simon Baron-Cohen and colleagues at the Cambridge Autism Research Center. It consists of a short multiple-choice quiz designed to assess autism traits.

Figure 1 shows the places where most applications or works on software with the theme of Autism were developed. It can be seen that USA and New Zealand were the countries that invested the most efforts in works for the diagnosis of autism. When we started the development of this work, in 2019, there were still not so many works on this topic, especially regarding the use of machine learning. However, the authors Thabtah et al. (2022) and Thabtah (2017) have invested constant efforts to improve his first application called ASDTests. The other countries that constantly invest in research in the field of autism diagnosis are the United States. Brazil, in turn, has invested more significantly in software for the educational support of children and adolescents diagnosed with ASD; however, many Brazilian applications were not found in the Apple Store for iOS and Google Play for Android. Thus, it is necessary for the Brazilian government to invest more in research promoting research that unites informatics in health.

Figure 1
Countries that have developed work since 2010 on the use of software for the diagnosis or treatment of autism



2.4. Workflow methodological process

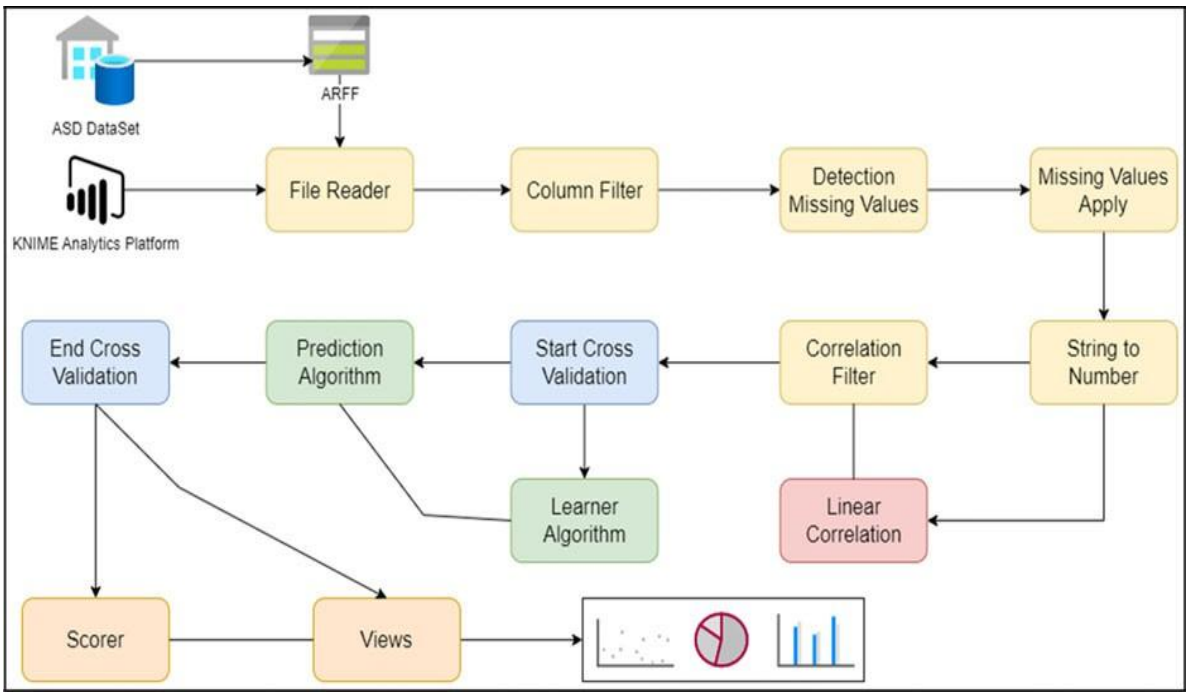
In the final model, we will also use the most usual steps in data mining and machine learning. The flowchart shown in Figure 2 demonstrates the model that was developed using the KNIME Analytics Platform tool. First, the Data Selection will be done through the file type.ARFF. A workflow will be performed for each supervised learning algorithm, which will have several nodes responsible for performing the entire data mining task. Some values changed from a string (string) to a number, so that the algorithms could handle the data properly, that is, the Pre-processing step of some data so that they were ready to be mined (Paul & Garai, 2022). The Transform step has the correlation node which is responsible for defining the redundant columns so that filtering is carried out, which is a part of the step of the data. For each column in the correlation model, the count of correlated columns is determined based on a threshold value (*threshold*) for the correlation coefficient. There are some data visualizations that help us understand the data better to create the best models.

3. Methodology

In the first step, the decision tree (DT) algorithm will be used. DT are models in the field of computational statistics that use supervised training for data classification and prediction. In other words, a training set is used in its construction, consisting of inputs, which in this case are patient data, and outputs, which represent the classes of autism spectrum disorder. Then, other models that were presented in the Materials Section were used in order to refine the best model for extracting the characteristics of ASD behaviors in children,

adolescents, and adults. In order to identify a more satisfactory model for the relationship between the attribute group and the class, 8 different types of algorithms for machine learning were used: the DT, Naive Bayes (NB), support

Figure 2
Flowchart for methodology used for learning machines algorithms



vector machine (SVM), multi-layer perceptron (MLP), probabilistic neural networks (PNN), random forest (RF), tree ensemble (TES), and gradient boosted (GB).

We know that DM comprises 5 essential steps: (1) Data selection; (2) Pre-processing; (3) Transformation; (4) Mining; and Analysis and assimilation of results, as shown by some researchers (Lima & Isotani, 2022; Baruh & Popescu, 2017). Thus, our analysis was based on two main steps, an initial test and refinement model, and a second final model that was used for our final quantitative precision task of data mining methods.

This KNIME correlation node determines which columns are redundant (that is, correlated) and filters them out. The output table will contain the reduced set of columns. The filtering step works roughly as follows: For each column in the correlation model, the count of correlated columns is determined based on a threshold value for the correlation coefficient (specified in the dialog). The column with the most correlated columns is chosen to "survive," and all correlated columns are filtered out. This procedure is repeated until no columns can be identified. The problem of finding a minimum set of columns to satisfy the constraints is difficult to solve analytically, so it is an approximate (Fillbrunn et al., 2017) process.

The first column filtering used in the Mining step was without the use of a correlation filter (= 1), thus containing all the individual characteristics of the patients. The correlation filter (CF) is a filter that determines by an iterative (approximate) process which columns are redundant; therefore, they are excluded (Garai et al., 2023). The higher the value, the fewer columns are filtered. Included and excluded column counts are shown in the label. The second was carried out containing as a limit value (= 0.1) which considered as redundant data the attributes of ethnicity, jaundice, and the result of the Classification/ASD. The third filter was carried out with the limit value (= 0.15), excluding the columns referring to age, gender, and genetic predisposition. Finally, in the filter with a threshold value of (= 0.05), gender, continent, and classification/ASD were considered as the most redundant characteristics.

At the conclusion of the entire process, the analysis and assimilation of results areconducted. This involves performing a validation (as explained in the next section) for each of the models. The goal is to rank the most effective algorithms for extracting behavioral characteristics of individuals with ASD based on the datasets utilized in this research project.

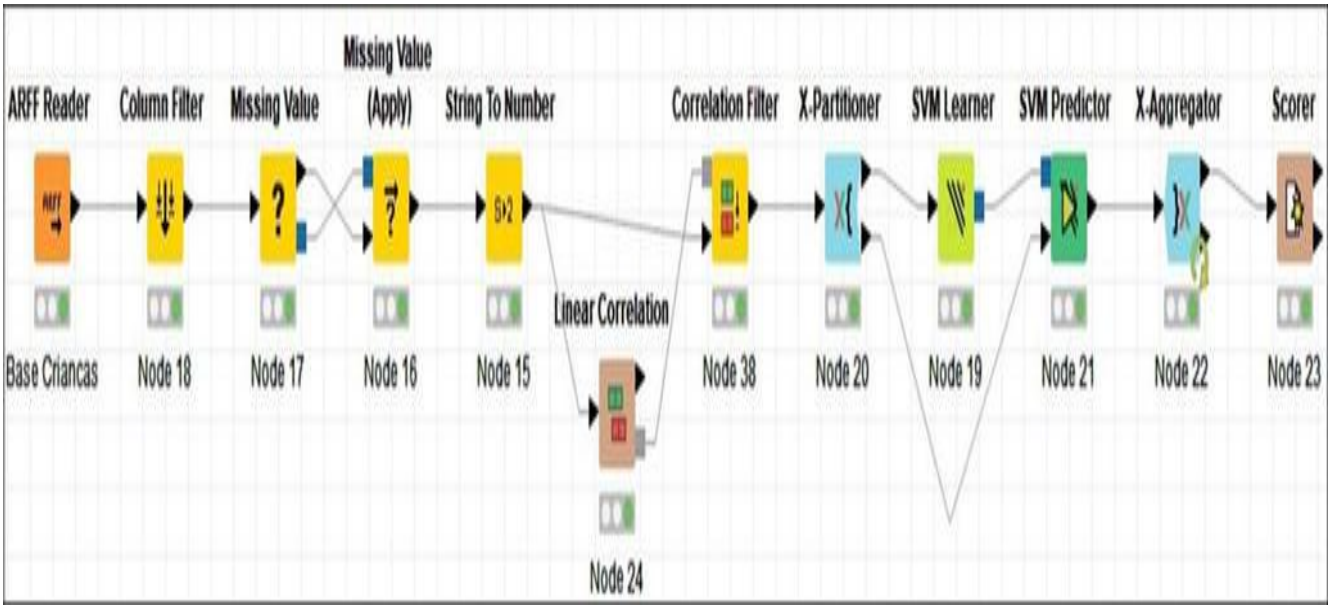
Our results are based in score rates, which compares two columns by their attribute value pairs and shows the

confusion matrix. The output of the node is the confusion matrix with the number of matches in each cell, that means accuracy. The Cohen kappa coefficient is also a statistic used to measure inter-rater reliability for qualitative items (McHugh, 2012). Thus, with these two measures, it is possible to rank the best algorithms considering the average of 10 executions performed in the KNIME Analytics Platform for each of the applied correlation filters.

Cross-validation (node cross-validation) is a technique to assess the generalizability of a model from a dataset (Kohavi, 1995). This technique is widely used in problems where the objective of modeling is prediction. In this case, an average of 10 iterations was used to attempt to generalize the model. The central concept of cross-validation techniques is the partitioning of the dataset into mutually exclusive subsets (Lima et al., 2021), and later, the use of some of these subsets to estimate the model parameters (training data), being the subsets remaining (validation or test data) used in the validation of the Lima et al. (2021) model. In the next section, it is expected that it will be possible to understand which attributes are most relevant for the survey of people with ASD, considering the present database used (Thabtah, 2017) for children, adults, and adolescents. In KNIME workflow presented in Figure 3, twelve (12) nodes were used, being the same in each of the 8 executed algorithms. This workflow is base in our last workflow presented in Figure 2. Furthermore, it should be considered that the set of attributes was filtered by the CF. Finally, 10 runs were performed and the final accuracy annotation is the average of the simulations carried out.

1. The first node will read Reader of the file of type ARFF, that is, the file with the database containing the responses of individuals of each age group who have gone through the screening process. ASD.
2. The Column Filter allows filtering the columns of the input table where the questionnaire of the 10 behavioral questions (A_1, A_2 A_{10}) already executed in the first step will be removed, while only the columns containing the 10 individual questions (AQ10) will be passed to the output table, and we substitute the country's name by continent. We also excluded Final Result, Age Description Relation remaining 8 attributes in database and also the result for ASD using AQ-10 test. In this case, the following attributes remain: age, gender, jaundice, genetic (autism) predisposition, prior use of app, continent, ethnicity, and Class/ASD final. In case of complete analysis, we considered 9 attributes, including age, gender, jaundice, genetic (autism) predisposition, prior use of app, continent, age description, ethnicity, and Class/ASD final.
3. The Missing Value allows filtering all missing values identified in the columns of the input table, so that the algorithms can flow normally. Some algorithms allow this setting to be set internally, others do not. To standardize, we use this approach.
4. Node Missing Value (Apply) reads the replacement settings from the previous node and applies them to the data, so that knowledge discovery by data mining techniques is not impaired.

Figure 3
KNIME Analytics Platform User Interface



Workflow model used on the individual features; in this example, the support vector machine (SVM) algorithm is shown

5. Node Stringto Numberisresponsibleforconvertingstringvalues (*string*) to number, so that algorithms can handle the data properly.
6. The node Linear Correlation is applied to analyze the attributes that most relate to each other; this node calculates for each pair of selected columns a correlation coefficient, that is, a measure of the correlation of the two variables. The value of this measure will range from -1 (strong negative correlation) to 1 (strong positive correlation). A value of 0 means no linear correlation.
7. The node Correlation Filter is responsible for determining the redundant columns, so that filtering is carried out. For each column in the correlation model, the count of correlated columns is determined based on a threshold value (*threshold*) for the correlation coefficient. The applied correlation filter is presented in Table 1, consisting of the values (1.00, 0.15, 0.10*, and* 0.05). A value of 1.00 indicates the absence of a correlation filter (all parameters were considered).
8. Node X-Partitioner is the first in a cross-validation loop. At the
end of the loop, there should be an X-Aggregator to collect the results of each iteration. All nodes between these two nodes will be executed $x = 10$ times, that is, the number of iterations defined to be performed, this loop is performed in order to provide an average of iterations.
9. Node will have the learning algorithm, Figure 2 shows as an example the SVM Learner, responsible for training the input dataset and searching for patterns for classification. However, the seven other algorithms can be changed in this step, being the SVM just one example.
10. Node SVM predictor contains the prediction algorithm, that is, it will predict the value of the class for new patterns; in this case, we use SVM as an example. However, any other Predictor algorithm could be used, depending on the Learner used.
11. Node X-Aggregator presents the end of the cross-validation loop, where $x = 10$ iterations. It stores the result of a predictor node, compares the predicted class and the actual class, and generates predictions for all rows and iteration statistics.

    (a) The workflow is ended by node Scorer which allows the selection of two columns so that the existing comparison between them can be performed; finally, it displays the confusion matrix with the number of matches in each cell. Accuracy: This node compares two columns by their attribute value pairs and shows the confusion matrix, that is, how many rows of which attribute and their classification match. The output of the node is the confusion matrix with the number of matches in each cell. Besides that it reports a number of True-Positives (TP), False-Positives (FP), True-Negatives (TN), and False-Negatives (FN).

    (b) Sensitivity: Sensitivity, also known as the true positive rate or recall, measures the proportion of actual positive instances that are correctly identified by a classification model. It is calculated as the ratio of true positives (TP) divided by the sum of true positives and false negatives (FN).

    (c) Specificity: Specificity measures the ability of a classification model to correctly identify negative instances. It is calculated as the ratio of true negatives (TN) divided by the sum of true negatives and false positives (FP).

    (d) Cohen's kappa: The Cohen's kappa coefficient (k) is a statistical measure commonly employed to assess the level of agreement between raters or within a single rater when dealing with categorical items, thus quantifying inter-rater or intra-rater reliability. It takes into account both the agreement that could occur by chance and the observed agreement beyond chance. The coefficient ranges from 1 to 1, with 1 indicating perfect agreement, 0 indicating agreement by chance, and 1 indicating complete disagreement.

These metrics provide valuable insights into the performance, accuracy, and reliability of the classification model applied in the workflow.

4. Results

In this section, we will present the results of the 8 machine learning algorithms (PNN, SVM, NB, MLP, RF, TES, GB, DT) used and the verification of which one has the best precision. The confusion matrix generated at the end of the workflow by Scorer

Table 1
Attributes for correlation filters applied in ASD dataset

Dataset	Filter	Included	Excluded	Remaining Attributes
Child	1.00	8	0	Age, gender, jaundice, autism predisposition, prior use of app, continent, ethnicity, Class/ASD
	0.15	6	2	Gender, ethnicity, jaundice, genetic predisposition, prior use of app, Class/ASD
	0.10	3	5	Ethnicity, jaundice, Class/ASD
	0.05	3	5	Gender, continent, Class/ASD
Teenager	1.00	8	0	Age, gender, jaundice, autism predisposition, prior use of app, continent, ethnicity, Class/ASD
	0.15	5	3	Jaundice, autism predisposition, prior use of app, continent, Class/ASD
	0.10	3	5	Jaundice, ethnicity, Class/ASD
	0.05	3	5	Age, prior use of app, Class/ASD
Adult	1.00	8	0	Age, gender, jaundice, autism predisposition, prior use of app, continent, ethnicity, Class/ASD
	0.15	6	2	Age, gender, autism predisposition, prior use of app, continent, Class/ASD
	0.10	5	3	Gender, ethnicity, jaundice, prior use of app, Class/ASD
	0.05	3	5	Age, prior use of app, Class/ASD
Complete	1.00	9	0	Age, gender, jaundice, autism predisposition, prior use of app, continent, ethnicity, age description, Class/ASD
0.15		6	3	Autism predisposition, prior use of app, continent, ethnicity, age description, Class/ASD
0.10		5	4	Age, jaundice, prior use of app, ethnicity, Class/ASD
0.05		3	6	Age, autism predisposition, Class/ASD

presents a table that allows the visualization of the performance of a classification algorithm. Generally, this special 2 2 contingency table is also called the error matrix. In this work, it is possible to observe all the results showing the percentage of accuracy, error, and Cohen's kappa of each of the tested algorithms. The Tables 2, 3, 4, and 5 were assembled based on these results, and on the left, we have the names of the algorithms used, and on the right, we have the results obtained after the application of each algorithm, changing the filters. First, through the CF node, a filter present in the KNIME Analytics Platform. Correlation filters of (0.05, 0.10, 0.15, 1.0) were utilized as stated in Table 1. It is important to emphasize that all these algorithms complement the previous model's 100% accuracy. In other words, with the 10 behavioral questions from the questionnaire, it was already possible to determine whether an individual had a diagnosis of ASD or not.

In this second moment, we are further improving the inference power of the Thabtah (2017) application, since although demographic data were collected, it was not used for anything in the first work of Thabtah (2017). At this stage, a very important aspect that should be taken into account is that: 10 attributes were from the questions (A_1, A_2, A_3,, A_{10}), 1 question for the Result (values between 0 10, where 7 was considered autistic). Of the remaining 10 questions, one represented the class (binary attribute Class = YES, NO), which is obligatorily necessary for classification. The attributes of Relation and Age_desc were also excluded, for the first three results, presented in Tables 2, 3, and 4; these values were disregarded, for example, for all children the Age_desc will always be (4–11) years old, for teenagers it will always be (12–16 or 12–15) years old, and for adults (> 16) years of age. As in the first moment, we are doing for each base, these Age_desc values become redundant, as they repeat for all columns. Regarding the Relation, it was disregarded, because it is an attribute that shows the relationship of the person who took the test to the autism (or non-autism), for example, for children who do not yet know how to read, the father may have done. But for our analysis, this will not be considered.

4.1. Experiments performed on the children database

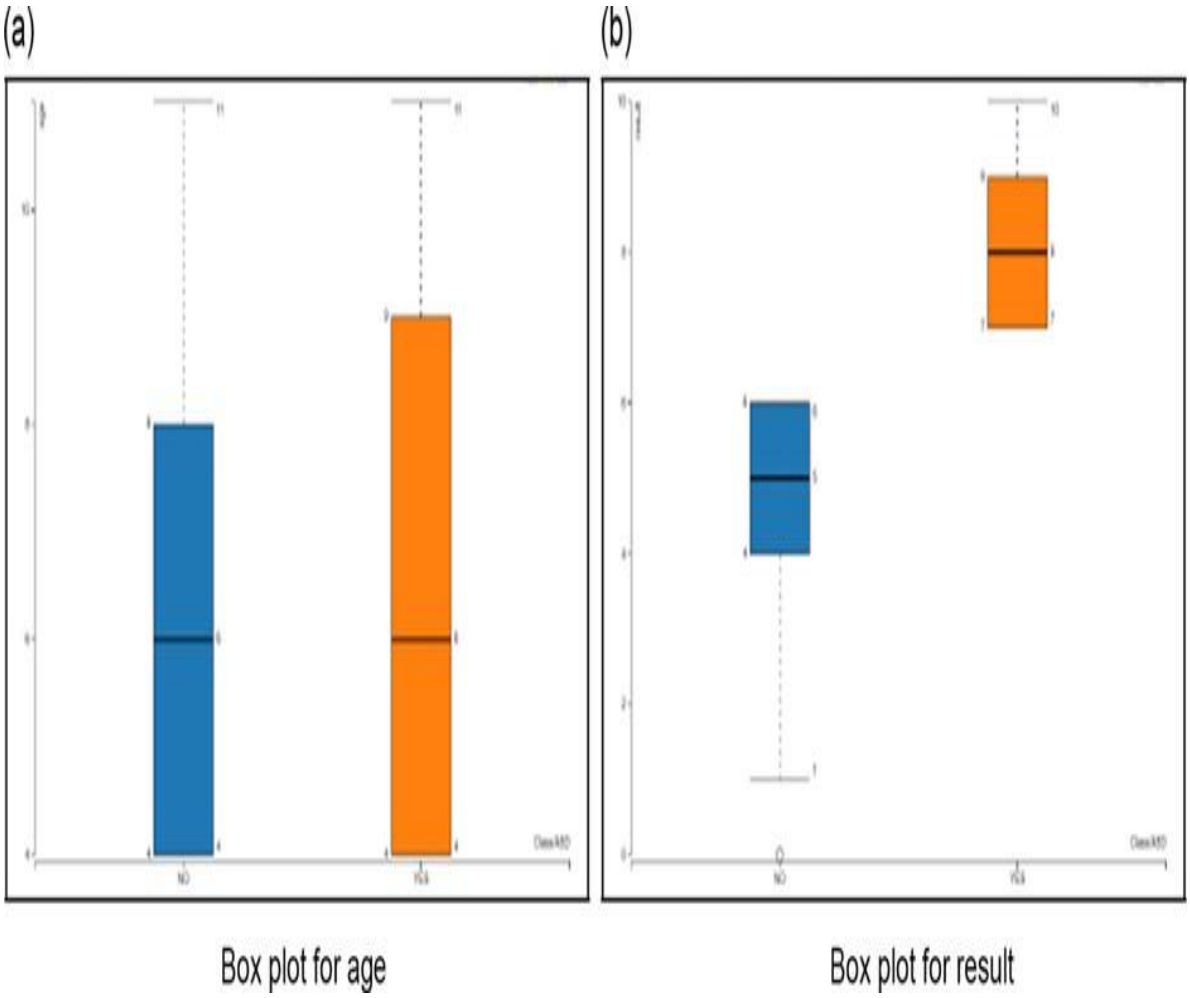
The children's database has 292 input data; in this sense, for a better understanding of the base, we used some data visualizations, see Figure 4. Considering, 48.29% (141 children) for positive (YES) class and 51.71% (151 children) for negative (NO) class. The first is a box plot in which two categories have been visualized. In descriptive statistics, a box plot is a graphical tool to represent the variation of observed data of a numerical variable through quartiles. The first (Figure 4(a)) is the average of ages that represents the positive (YES) and negative (NO) classes. In this way, we have the median of age for both classes as $\tilde{x} = 6$ years old. In Figure 4(b), we have the representation for result for the 10 questions (A_1, A_2,, A_{10}) median as $\tilde{x} = 5$ for negative (NO) class and here it is possible to observe one outlier when the child has no predisposition for the AQ-10 and scored 0 points. On the other hand, the median is $\tilde{x} = 8$ for positive (YES) class.

This means that, even if the child does not have ASD, the result has a high value for the median Result of the questionnaire, since from Result $\tilde{x} = 7$, according to Thabtah (2017), the person has ASD.

Figure 5 represents two histograms for viewing children's data. In this sense, in Figure 5(a) we have for the classes YES and NO the difference between the attributes: gender, ethnicity, jaundice, autism continent, relation. It can be seen in the image that the attributes do not suffer much variation between classes, since the average values did not present significant differences. The attribute that suffered the most difference between the classes was the continent, where for YES it presented an average of 2.49 and for NO it presented an average of 2.72. On the other hand, in Figure 5(b) we have that the differences for classes YES and NO are significant. In this sense, people in the YES class present answers to the lowest AQ-10 questions for A_2 in which children usually concentrate more on the whole picture rather than the small details. On the other hand, still for the YES classes, the question with the highest average of positive answers was the AQ-10 question A_{10} in which the person presents difficult to work out people's intentions. As for the NO class, the question that presented the lowest positive average was question A_4, in which the child can switch back to what he/she was doing. On the other hand, for the YES class, the question with the highest positive average was the A_3 questions that show the ease of doing more than one thing at the same time and the A_5 question that shows that the person is easily able to read between the lines while talking with someone else.

Parallel coordinates (PC) are a visualization technique used to analyze high-dimensional datasets. The visualization consists of a set of parallel lines, usually vertical and equally spaced, which represent different attributes or dimensions of the data. Each data point I

Figure 4

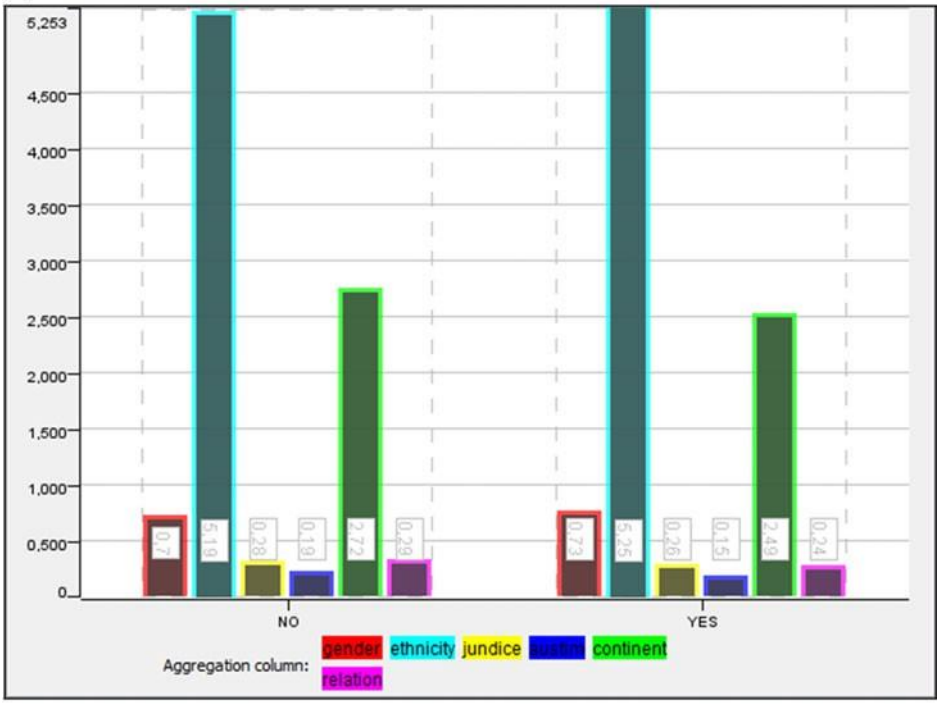


Box plot for children's age and test result binning by ASD class

Figure 5

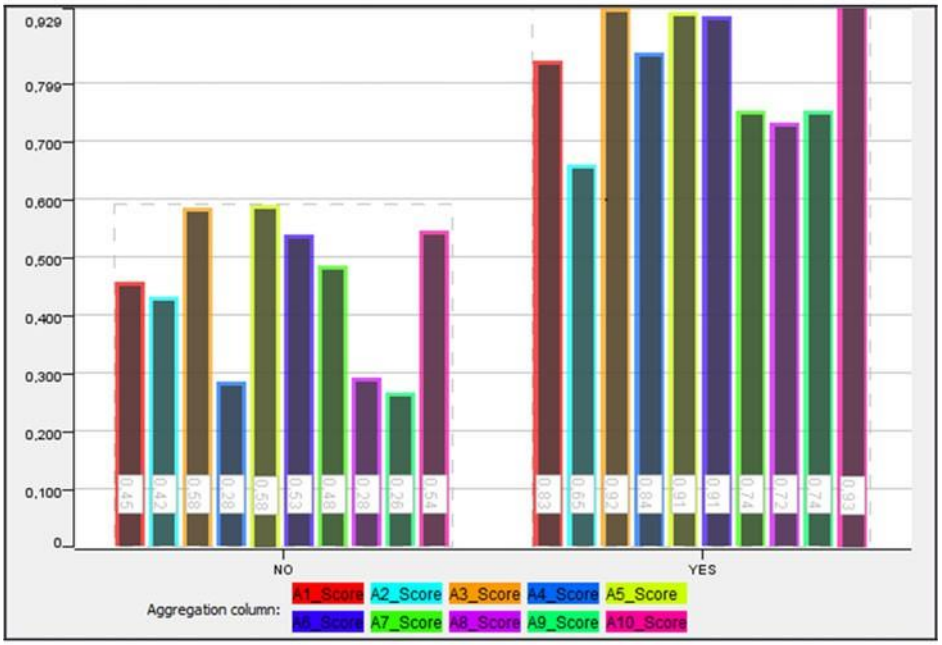
Features and results histogram for children database

(a)



Features histogram

(b)



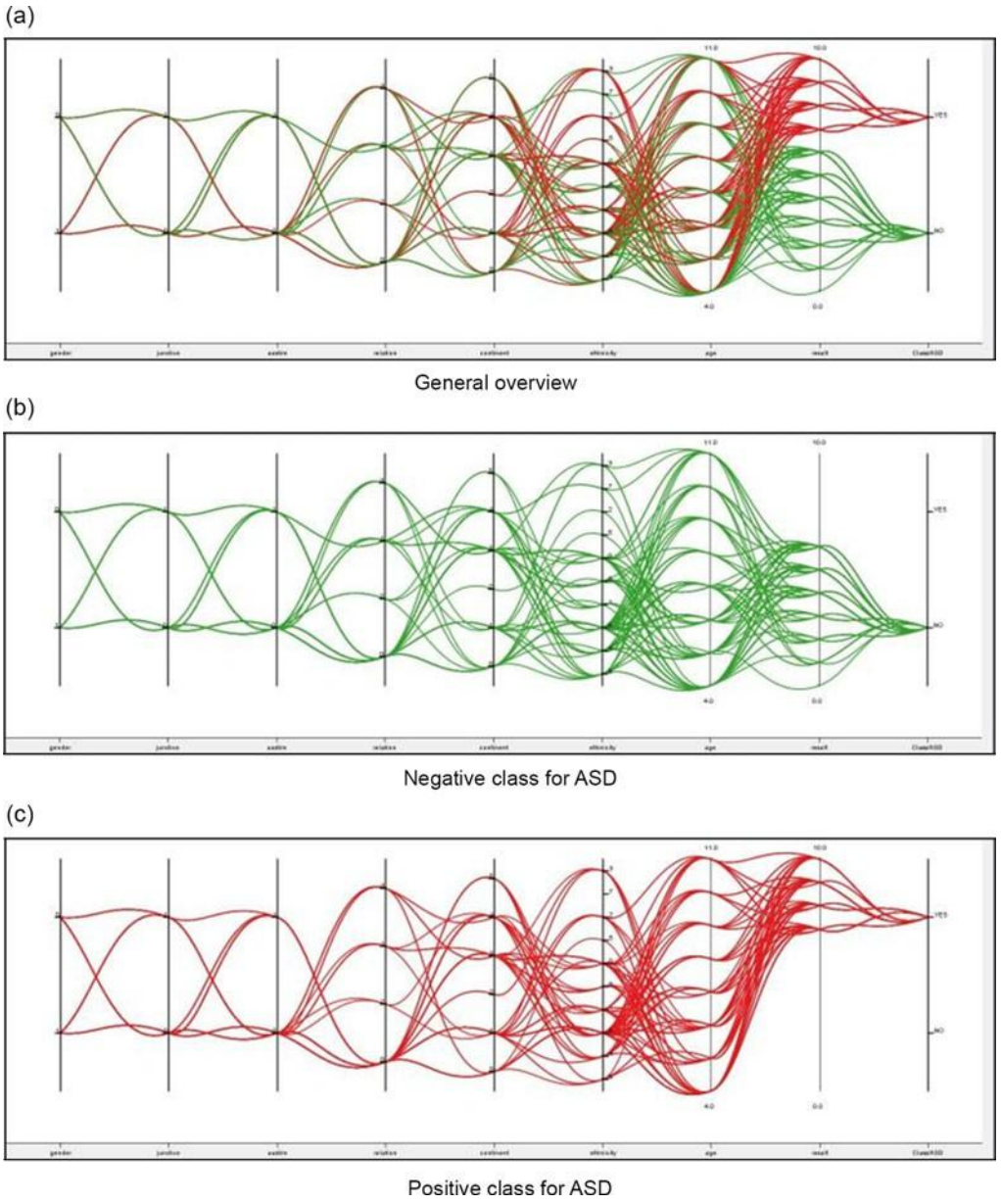
Results histogram

represented by a line segment that intersects these parallel lines at specific positions.

In Figure 6, we present the parallel coordinates plot, which provides insights into the relationship between different attributes in the dataset. The plot consists of three parts: Figure 6(a) shows the general parallel coordinates plot with all attributes, including gender, jaundice, autism in family, relation, continent, ethnicity, age, result (ranging from 0 to 10), and class (YES and NO).

Figure 6(b) focuses on the attributes for the NO class, where the lines intersect the parallel axes, indicating the values for each attribute. Similarly, Figure 6(c) displays the attributes for the YES class. By comparing these two plots, we can observe that the

Figure 6
Parallel coordinates for child database including positive and negative classes for ASD



attribute that varies the most between the YES and NO classes is the "result" coordinate. This attribute serves as a differentiating factor between the two classes, as it shows distinct values for each class.

Table 2 shows the average accuracy data for the 10 simulations for each of the algorithms, considering the values of the correlation filters. The correlation column automatically filters data for the user. The results refer to data from the children's database, which has 292 data, according to the following configurations, considering the remaining attributes filtered by correlation filter according to Table 1. From the reference data in Table 2, it was possible to observe that the PNN algorithm with the CF at 0.1 (considering ethnicity and the incidence of jaundice) showed better performance compared to the others, obtaining the value of 60,274% accuracy in results. Additionally, this algorithm was also the one that presented the best result in relation to the Cohen kappa coefficient (k = 0.205). It is generally considered to be a more robust measure than the simple percentage agreement calculation, as k takes into account the possibility of agreement occurring by chance. On the other hand, the worst classification we obtained was with the MLP algorithm, with an accuracy of only 45,548% (with the error rate surpassing the hit rate) and k = 0.094, when no CF is used. (= 1), that is, considering all attributes.

4.2. Experiments performed on the adolescent/ teenager database

First, in Figure 7 we will present the results of the 104 adolescents who responded to the database survey. We have 41 of teenagers (39.42%) do not have autism NO. On the other hand, 63

Figure 7
Box plot for adolescent's age and test result binning by ASD class

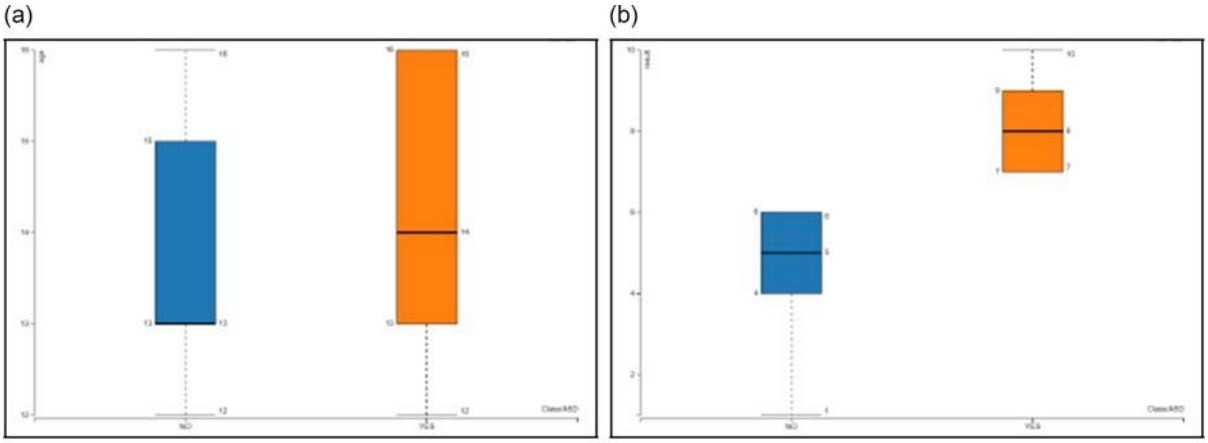


Table 5
Results with the different supervised learning techniques for all complete databases

Algorithm	Results					
	No filter			Filter = 0.1		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Naive Bayes	63.727%	36.273%	0.177	62.818%	37.182%	0.109
	Filter = 0.15			Filter = 0.05		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	64.091%	35.909%	0.158	63.273%	36.727%	0.072
	No filter			Filter = 0.1		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Decision Tree	64.182%	35.818%	0.205	64.636%	35.364%	0.206
	Filter = 0.15			Filter = 0.05		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	64.091%	35.909%	0.19	61.182%	38.818%	0.095

		No filter			Filter = 0.1	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
SVM	62.727%	37.273% Filter = 0.15	0.071	64.273%	35.727% Filter = 0.05	0.000
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	63.364%	36.636%	0.086	64.273%	35.727%	0.000
		No filter			Filter = 0.1	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
MLP	65.727%	34.273% Filter = 0.15	0.202	64.727%	35.273% Filter = 0.05	0.162
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	67.00%	33.00%	0.248	63.909%	36.091%	0.127
		No filter			Filter = 0.1	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Random Forest	65.727%	34.273% Filter = 0.15	0.198	65.727%	34.273% Filter = 0.05	0.186
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	65.818%	34.182%	0.189	65.727%	34.273%	0.138
		No filter			Filter = 0.1	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Tree Ensemble	65.545%	34.455% Filter = 0.15	0.196	65.182%	34.818% Filter = 0.05	0.177
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	66.00%	34.00%	0.189	65.636%	34.364%	0.16
		No filter			Filter = 0.1	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
PNN	66.788%	33.212% Filter = 0.15	0.2	64.422%	35.578% Filter = 0.05	0.151
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	64.545%	35.455%	0.197	62.147%	37.853%	0.115
		No filter			Filter = 0.1	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Gradient Boosted	65.273%	34.727% Filter = 0.15	0.213	64.727%	35.273% Filter = 0.05	0.192
Accuracy		Error	Kappa	Accuracy	Error	Kappa
65.727%		34.273%	0.224	63.273%	36.727%	0.159

4.3. Experiment for individual features and aggregation of AQ-10 behavior attributes

Based on the previous experiments, it was possible to discover a set of attributes from the databases that, linked to the AQ-10, we are able to know a tendency toward the respondents' autism, in this case, ethnicity, jaundice, age description, continent, and autism predisposition. However, to add more value to our experiments, we chose attributes that can influence our decision making, we used the AQ-10 in aggregation to the attributes previously filtered; thus, we obtained 15 attributes + Class/ASD = 16 attributes. Thus, considering the three databases: Child, Teenager, and Adult, the 1100 data (956 without missing rows) and the 16 attributes, we obtained the values from Table 6. This experiment was carried out to demonstrate that the accuracy can be increased when we take into account the attributes of the AQ-10, (A_1, A_2,, A_{10});

Table 6
Classification results considering behavioral attributes, both individual and demographic

Learner	Class	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Cohen's kappa
DTL	NO	542	34	334	46	0.921769	0.907609		
	YES	334	46	542	34	0.907609	0.921769		
.916318	0.824349								
GBL	NO	573	10	358	15	0.97449	0.972826	0.973849	0.944914
	YES	358	15	573	10	0.972826	0.97449		
MLP	NO	585	1	367	3	0.994898	0.997283	0.995816	0.991173
	YES	367	3	585	1	0.997283	0.994898		
NBL	NO	560	31	337	28	0.952381	0.915761	0.938285	0.869468
	YES	337	28	560	31	0.915761	0.952381		
PNN	NO	588	92	276	0	1.000000	0.750000	0.903766	0.786798
	YES	276	0	588	92	0.750000	1.000000		
RFL	NO	575	24	344	13	0.977891	0.934783	0.961297	0.917806
	YES	344	13	575	24	0.934783	0.977891		
SVM	NO	588	0	368	0	1.000000	1.000000	1.000000	1.000000
	YES	368	0	588	0	1.000000	1.000000		
TEL	NO	576	23	345	12	0.979592	0.937500	0.963389	0.922249
	YES	345	12	576	23	0.937500	0.979592		

however, we can have a relatively significant accuracy when only individual characteristics are observed.

When looking at Table 6, we observed that the SVM algorithm presented 100% of accuracy, being the best result; in addition, it presented sensitivity and specificity of 1.0, the highest value that could be achieved, plus k = 1.0 from Cohen's kappa is also the maximum value. For the worst accuracy result, we have the PNN algorithm, which reached only 90.37% of accuracy, with Cohen's kappa k = 0.78; for this same algorithm, we also had the worst performance of sensitivity (0.75) and specificity (1.00), this considering the YES class. Finally, we consider that, on average, all algorithms had accuracy results above 90%, which demonstrates stability of the measures used and attributes selected from dataset. The obtained results suggest that incorporating the AQ-10 attributes can enhance the accuracy of autism classification, but even without them, satisfactory accuracy levels can still be achieved by considering individual characteristics alone. The table provides a comprehensive overview of the performance of different learners in terms of true positive, false positive, true negative, and false negative counts. It also includes metrics such as sensitivity, specificity, accuracy, and Cohen's kappa value. These metrics allow us to evaluate the effectiveness of each learner in classifying the NO and YES classes. Overall, the learners showed varying levels of performance in classifying the NO and YES classes. MLP, SVM, and GBL stood out as top performers with excellent accuracy and balanced sensitivity and specificity values. DTL, NBL, RFL, and TEL also performed well, while PNN showed imbalanced sensitivity and specificity. Besides that, the results demonstrate the varying performance of different learners, with SVM achieving the highest accuracy of 100% for both classes, indicating its exceptional performance in this classification task.

5. Discussion of Results

In this section, we will carry out a balance and an analysis of the results obtained in both models: initial and final. The data obtained in the initial model were only an indication of how the data behaved in relation to the questions in the questionnaires. We used the correlation filter, and filtered 8 data mining algorithms were compared using 4 different filters. At this stage, the correlation filters with the highest accuracy rates by the algorithms was 0.15. Another relevant factor to be analyzed, this being the main contribution of this work, was the isolation of variables that tend to be more relevant for the diagnosis of autism are the ethnicity variables with 100% predominance for the filters, followed by the variables jaundice and continent, with 75% predominance. Even though no type of pre-analysis was performed, the algorithm ended up isolating the variables of ethnicity and continent, showing that there is a strong correlation between these variables in diagnosis of ASD.

In this case, health specialists and developers of medical applications aimed at the diagnosis of ASD are recommended to exclude filters with very low values, that is, filters close to 0, making the mining of data impaired due to the low diversity of attributes. In our case, specifically, the 0.05 filter was bad in 75% of the test cases. We did not have a prominent algorithm for the worst case, and the MLP, DTL, RFL, and PNN were observed. In this case, we will not consider this approach, as PNN also had 50% of the best cases. Regarding the variables that should be treated with less relevance was age, with 100% of occurrence for the performances with the worst accuracy. Furthermore, with 75% occurrence for the worst performance we have the variable (Used_app_before) previously used by the application. For the complete experiment carried out, considering the refinement of the attributes, even considering the worst accuracy value, the values were still higher than what was achieved for the highest accuracy with the adult database, the highest accuracy of 73,864% and error of 26,136%, where the PNN algorithm had the highest success rate. Therefore, some individual characteristics of the respondents can help us to predict the ASD; however, when we aggregate the AQ-10 questions, we start to have more confidence in our answers and in the prediction made by the algorithm, such as SVM that obtained 100% of accuracy considering the attributes: AQ-10 + jaundice, ethnicity, continent, autism predisposition, and age_desc. The model behaves better when we use behavioral attributes, both individual and demographic, but that only demographic and individual attributes can also suggest ASD predisposition.

6. Conclusion

7. The number of cases of ASD has increased dramatically in recent years, and as a result, many families struggle to receive a diagnosis at the right time for treatment to start. This is because early diagnosis has many benefits, including preventing the full emergence of ASD, while late diagnosis can cause greater harm to the child's development. Considering these factors, it has lately been feasible to see that one of the main focuses of research on ASD is the advancement of diagnostic techniques to shorten waiting times and provide patients rapid access to therapy. In light of these objectives, researchers have been using machine learning-based techniques; however, it should be noted that the use of these techniques for the identification and diagnosis of ASD is still in its early stages, despite the fact that they can show very positive results (Thabtah et al., 2022). We took on the task of proving the significance of data science for the classification of ASD in this paper because of the potential growth of diagnostic techniques and the creation of applications that can track using particular attributes that can improve prediction accuracy without compromising efficacy.

Several data mining algorithms were used in databases of up to 1100 data (children: 292 data, teenagers: 104 data, adults: 704 data, and complete: 1100 data) during the DM stage to show how highly relevant the behavioral and individual characteristics related to the performance of screenings that make up the bases used are. Therefore, it was shown that factors pertaining to race, continent, and jaundice at birth contribute to the accuracy of the diagnosis; both medical professionals and application developers should take these factors into account. The characteristics associated with the patient's age and prior usage of the application, which yielded the poorest outcomes, should be ignored by those working in the fields of neurology and health. There is a clear need for accessible screening methods given the previously indicated variables, such as the financial effect of ASD and the high expenditures of medical aid, as well as the shortcomings of current instruments that have not produced adequate results. These processes would help healthcare professionals in their evaluations and enable self-assessment (or assessments led by family or educators). In order to identify autism traits through more accurate screenings that can provide a timely prior diagnosis that can be used by a number of users, it was suggested in this work to isolate variables (demographic, geographical, and individual) and behavioral characteristics. As a result, it will provide those who want quicker advice access to specialized medical professionals. The ASD Tests application suggested by Thabtah (2017) might include the isolated factors (race, continent, autism genetic risk, and jaundice). These factors may increase the efficacy of diagnosing ASD in a variety of age groups by around 72% when paired with the application's ten behavioral questions. Additionally, the SVM learner attains a flawless accuracy rate of 100% when the AQ-10 is used with classification methods. Further investigation into the distinct effects of each variable (individual, regional, and demographic) on the dataset's behavioral aspects would be beneficial. The complex interaction between many causes and the presentation of ASD may be better understood by comprehending the separate contributions of these variables. Additionally, examining the possible relationships and interactions between these factors might improve our comprehension of the intricacy of diagnosing ASD. Furthermore, examining the dataset's longitudinal features and taking time into account may help clarify how behavioral patterns change in people with ASD. Future research may help us get a more thorough knowledge of the complex relationships between individual, regional, and demographic variables and how they affect the behavioral traits of people on the autism spectrum.

References

Baron-Cohen, S., Auyeung, B., and Allison, C. (2012). The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls are two quick "red flags" for autism screening. American Academy of Child & Adolescent Psychiatry Journal, 51(2), 202–212.Pimentel, A. C. M., and Assumpção Jr., F. B. (2000). Autismo infantil. Journal of Psychiatry in Brazil, 22, 37–39.Adnan, S. F., Aziz, M. Z., Abdullah, S. A., and Mazalan, L. (2014). Children with autism spectrum disorders (ASDS) may use this educational tool. Computer Science Procedia, 42, 70–77.Popescu, M., and Baruh, L. (2017). The boundaries of privacy self-management with big data analytics. 579–596 in New Media & Society, 19(4). Al Mamun, S., Biswas, M., Kaiser, M. S., Mahmud, M., Hossain, M. S., & Rahman, M. A. (2021). The context of a XAI-based autism detection system. Brain Informatics: 14th International Conference Proceedings, 14, 448–459.Francisco, R. D. P., Carvalho, T. P., Soares, F. A., Vita, R., Basto, J. P., & Alcalá, S. G. (2019). a thorough analysis of the literature on machine learning techniques used in predictive maintenance. Industrial Engineering & Computers, 137.F. Chagas (2016). Autismo: Caracterização e classifica-ção do grau de severidade dos alunos da associação maringaense dos autistas (AMA) con base no método cars. Journal of Surgery and Clinical Research in Brazil, 15(3), 37–41.Thomas, P., Bhatia, T., Nimgaonkar, V. L., Chakraborty, S., and Deshpande, S. N. (2015). The Indian scale for autism evaluation is used to determine the degree of autism. Journal of Psychological Medicine in India, 37(2).Goddard, L., Henry, L. A., Crane, L., Chester, J. W., and Hill, E. (2016). Over a thousand parents in the UK were surveyed about their experiences receiving an autism diagnosis. Autism, 20(2), 153-162. Powell, S. D., Davis, M., Otero, N., Dautenhahn, K., and Nehaniv, C. L. (2007). Developing software to help autistic youngsters comprehend narrative: thinking about how possibilities for reasoning and feedback are designed. 64–69 in IEEE 6th International Conference on Development and Learning. Sibaldo, M. A. A., Tsang, R., de Carvalho, T. B. A., and da Cunha Cavalcanti, G. D. (2017). Principal component analysis: A minimal classification error method for supervised learning. Information and Data Management Journal, 8(2), 131–131. Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., de Morais, J. I., & Barbon Jr., S. (2020). A multi-label categorization method to differentiate between news in Brazilian Portuguese that is fraudulent, satirical, objective, and authentic. Brazilian Journal of Information Systems, 13(4), 126–149 (iSys).Madela, J. R., Eder, M. S., Diaz, J. M. S., Magusara, M., & Sabellano, D. D. (2016). The Fill Me App is an interactive smartphone game designed for kids with autism. Interactive Mobile Technologies International Journal, 10(3), 59–63. ttps://doi.org/10.3991/ijim.v10i3.5553 Farias, E., and Cunha, M. (2013). prototype of a software tool

to support the treatment of children with autism. Anais do IX Simposio Brasileiro de Sistemas de Información, 332–342.Dietz, C., Fillbrunn, A., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). Knime for cross-domain analysis of life science data that is repeatable. Biotechnology Journal, 261, 149–156.McConachie, H., Fletcher-Watson, S., Pain, H., Hammond, S., and Humphry, A. (2016). An iPad app case study on designing for young children with autism spectrum disorder. Child-Computer Interaction International Journal, 7, 1–14. In 2023, Garai, S., and Paul, R. K. developed ensemble models based on MCS employing machine intelligence and Ceemdan decomposition. 18. Garai, S., Paul, R. K., Rakshit, D., Yeasin, M., Paul, A., Roy, H., & Manjunatha, B. Intelligent Systems with Applications (2023). An MLR model based on MRA that uses large-scale climatic variables to predict yearly rainfall in India. Environment and Climate Change International Journal, 13(5), 137–150.Goulart, P., and G. J. A. de Assis (2002). Research on autism in behavior analysis: Methodological aspects. Brazilian Journal of Comportamental and Cognitive Therapy, 4(2), 151–165.De Campos Merschmann, L. H., and Guandaline, V. H. (2017). Hcaim: A discretizer used on bioinformatics datasets for the hierarchical classification scenario. Kanner, L. (1965). Journal of Information and Data Management, 8(2), 146. Schizophrenia and infantile autism.Behavioral Science, 10(4), 412–420.A. Klin (2006). Autismo e síndrome de asperger: Uma visão geral.Journal of Psychiatry in Brazil, 28, s3–s11.R. Kohavi (1995). An analysis of bootstrap and cross-validation for model selection and accuracy estimates. 14(2), 1137–1145, International Joint Conference on Artificial Intelligence. Silva, A. F. F., Ferreira, M. E. A., and Lima, D. A. (2021). A robotics and programming project aimed at women is evaluated using machine learning and data visualization. Intelligent & Robotic Systems Journal, 103(1), 1–20.Isotani, S., and Lima, D. A. (2022). An investigation using a data clustering method to map and assess Google Classroom use during the COVID-19 outbreak. Brazilian Journal of Information in Education, 30, 20–49.

Lima, D. A., and H. J. Lopes (2021). Evolutionary tabu inverted ant cellular automata with elitist inertia for swarm robotics as a stand-in technique for e-puck architecture surveillance tasks. Autonomous Systems and Robotics, 144.Ballesta-Pagán, F. J., Lozano-Martínez, J., and Alcaraz-García, S. (2011). Software to detect emotions in students with autism spectrum disorders. Communicar: Journal of Communication and Education, 18(36), 139–148.M. L. McHugh (2012). The kappa statistic measures interrater dependability.

Biochemia Medica, 22(3), 276-282.Garai, S., and Paul, R. K. (2022). An artificial neural network method for predicting agricultural prices based on wavelets. Indian Society for Probability and Statistics Journal, 23(1), 47–61. Thabtah, F., and Shahamiri, S. R. (2020). Autism AI: A novel artificial intelligence-based autism screening method. 12(4), 766–777; Cognitive Computation.Siegel, B. (2021). Screening test for pervasive developmental disorders (PDDST). Autism Spectrum Disorders Encyclopedia, 3447–3451.Costa, E. A. B., de Castro, T. H. C., and Sousa, F. R. M. (2012). Worldtour: Software to assist in the process of creating original content. 23(1) Brazilian Symposium on Computers in Education.In 2022, Stone-Heaberlin, M., Rouse, M. L., Blake, H. S., Fodstad, J. C., Smith, J., Kerswill, S., & Bushnell, E. addressing eating issues in people with severe developmental disabilities and autism. The Handbook of Autism and Pervasive Developmental Disorder, edited by J. L. Matson and P. Sturmey. Springer.