Received: 28 April 2023 | Revised: 18 May 2023 | Accepted: 6 June 2023 | Published online: 7 June 2023

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2023, Vol. 1(2) 105–127
DOI: 10.47852/bonviewJDSIS32021027

BON VIEW PUBLISHING

Correlation Filters in Machine Learning Algorithms to Select Demographic and Individual Features for Autism Spectrum Disorder Diagnosis

Raquel S. Dornelas[1] and Danielli A. Lima[1,*]

[1]*Laboratory of Intelligent Computing and Robotics, Federal Institute of Triangulo Mineiro Campus Patrocínio, Brazil*

Abstract: Autism spectrum disorder is currently considered one of the main neurodevelopmental disorders with predominant characteristics of difficulty in social communication and cognitive skills, and limited and repetitive patterns. This disorder has no cure and has different levels of severity that vary according to the appearance of symptoms in each patient. Generally, the waiting time for the diagnosis of autism spectrum disorder is slow, having as one of the reasons for this situation the lack of development of simple screening procedures to be implemented and which have efficient results. The objective of this work is to analyze a public database in order to find patterns of the autism spectrum, that is, to isolate the attributes that together with behavioral characteristics can bring greater reliability to the precursor model. The preliminary results showed that the probabilistic neural network algorithm performed well in this classification. In addition, the application of correlation filters demonstrated greater efficiency in accuracy. By applying eight data mining algorithms and aggregating the demographic, individual, and behavioral attributes, and excluding some attributes, we obtained an accuracy of 100% through the support vector machine. Finally, the results with machine learning have shown that the patient's ethnicity, continent, and the presence of jaundice tend to reveal more likely that the patient will be diagnosed with autism spectrum disorder.

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

1. Introduction

Autism spectrum disorder (ASD) is not a disease but a neurological condition characterized by remarkable alterations in the development of language and social interaction (Vieira & Baldin, 2017). Other authors, such as Siegel (2021) and Stone-Heaberlin et al. (2022), propose that autism is classified as one of the pervasive developmental disorders (PDD). Among its behavioral symptoms are difficulties in communication and social skills, varying in degrees. Additionally, individuals with autism may exhibit repetitive and restricted patterns of behavior, including a strong preference for routines, resistance to changes, and intense attachment to specific objects, among other characteristics. Because autism spectrum disorder (ASD) encompasses a range of presentations, it can be categorized into three distinct levels based on the level of support required by the individual, as outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[1]. According to Chagas (2016), these levels are: Level 1, which indicates mild ASD and necessitates support; Level 2, indicating moderate ASD and requiring substantial support; and Level 3, indicating severe ASD and necessitating extensive support.

In recent years, there has been a significant increase in scientific research on autism in several areas, aiming to increase knowledge, both about the nature of the disorder and possible treatment strategies (Zeidan et al., 2022). This increase has contributed to the development of increasingly quicker and more effective methods for identification and diagnosis.

It is estimated that 1.5% of the world population have some ASD; however, many cases have not yet been detected. The term autism was included in the current meaning, thanks to the scientist and psychiatrist Leo Kanner, who in 1943 used the expression to detail the clinical situation of 11 children who had similarities in their behavior as a language development disorder, deep insistence on preserving the unchanged environment, and certain rejections when relating to other people (Kanner, 1965). Since then, attempts have been made to gather the symptoms and behaviors of children diagnosed with autism, with the aim of standardizing them, based on the universalizing of the language used (Goulart & de Assis, 2002). Currently, the principles used to diagnose the autism have advanced and are represented in the DSM, published by the American Psychiatric Association (APA).

Although autism has no cure, after being diagnosed there are treatments to be performed that may or may not be complex depending on the degree of severity of the patient with the syndrome; these treatments must be performed by a group of professionals formed by physicians, speech therapists, psychologists,

[1]Diagnostic and Statistical Manual of Mental Disorders DSM-5. (S.l.): proposed by the American Psychological Association (APA) of ARTMED Editora LTDA, 2014. v. 5.

***Corresponding author:** Danielli A. Lima, Laboratory of Intelligent Computing and Robotics, Federal Institute of Triangulo Mineiro Campus Patrocínio, Brazil. Email: danielli@iftm.edu.br

and physical educators, among others. It is worth remembering, however, that precisely because they are chronic patients, this therapeutic vision will extend for long periods, requiring constant monitoring from the professionals involved, so that they have an exact dimension of the problem (Assumpção & Pimentel, 2000). After going through several modifications over time, currently the term "autism" is known as autism spectrum disorder (ASD), coming from the Diagnostic and Statistical Manual of Mental Disorders (DSM-V). ASD is used to describe different neurodevelopmental disorders that have limitations in mutual social communication and social interaction, in addition to repetitive and limited behavior patterns. The term came to be used to represent all the subdivisions that existed for these types of disorders (Klin, 2006), since then it includes Asperger's syndrome, unspecified pervasive developmental disorder, and childhood disintegrative disorder.

For some time, researchers (Thabtah et al., 2022) have been adhering to efficient machine learning (ML) techniques (Carvalho et al., 2019; Garai et al., 2023; Garai & Paul, 2023; Lopes & Lima, 2021) with the aim of improving the ASD diagnosis method by improving symptom analysis time and accuracy to provide patients with faster access to treatment. It is currently perceived that there is a large amount of accessible data capable of overcoming the effectiveness of manual analysis. Another existing factor is the emergence of increasingly evolved algorithms that, connected to datasets, make them capable of carrying out studies and searches in a broader and more complex way. Due to these factors, there was a growth of applications conveyed to data science, which is linked to other definitions, such as big data and decision making based on data, which are also of great importance. It is also known that the number of cases of ASD is growing rapidly all over the world and one of the difficulties in the diagnosis process is to create an automated model to carry out the classification that defines whether a case is ASD or not.

Different works have already been developed in order to help children in different age groups with ASD in the school environment (Farias & Cunha, 2013; Sousa et al., 2012); other works focus on creating software for the diagnosis of people with ASD (Crane et al., 2016; Davis et al., 2007). However, there are few works that seek, through data mining, to create associations with relevant characteristics for the prototyping of an application. Having then known that the autism diagnosis process is a work carried out through classification, it is expected to be able to carry out the identification of ASD cases from analysis and evaluation of instances obtained through an input dataset containing behavioral characteristics. Data science techniques will be utilized to classify ASD behavior by leveraging a dataset obtained from the UCI Machine Learning Repository. The integration of this dataset with the KNIME Analytics Platform, an open-source platform, will enable the extraction of valuable insights through the creation of analytics and reports.

The general objective of this paper is to perform a data treatment on a consolidated basis and through different ML techniques extracting the most relevant characteristics for the diagnosis of ASD more accurately. This study encompasses the analysis of three comprehensive databases comprising a total of 1100 instances, encompassing responses obtained from ASD screenings conducted among children, adolescents, and adults. Each dataset consists of 20 attributes, where 10 pertain to behavioral characteristics and the remaining 10 to individual traits, which have proven to be effective in successfully identifying cases of ASD. The data were structured in ARFF files and later applied in the KNIME tool used to perform the data mining. An extensive analysis was undertaken using eight supervised learning techniques to compare and evaluate the obtained results. It is important to note that the primary objective of this study was not to develop software but rather to uncover and establish a model capable of accurately classifying patterns that can aid in the identification of ASD. By doing so, this research aims to significantly reduce the waiting time for patients seeking diagnosis.

This work is organized into five sections. The first section is the Introduction and also brings the objectives of the work. In the second section, we have the Background with some important definitions about the work, as well as related works. In the third section, we have the proposal of the work where we can see the steps of implementation of models, schemes and supervised learning algorithms in the data mining tool adopted here. In the fourth section, we have the Analysis of Results and the discussion of the present work where we analyze and present the results. In the fifth section, we have the Conclusion of the work with future work.

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

[2]With NODA, doctors observe the behaviors captured by the family at home through a secure platform. NODA was developed by Behavior Imaging Solutions Inc, https://behaviorimaging.com/noda/ and https://floreotech.com/team/.

[3]ASDTests: mobile app for ASD screening. www.asdtests.com, was accessed on December 20, 2020).

[4]Autism Speaks is the largest autism advocacy organization in the United States. She sponsors autism research and conducts awareness and outreach activities aimed at families, governments and the public https://www.autismspeaks.org/.

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

[5]M-learning application with version for Android 4.2 –Jelly Bean https://apps.mgov.gov.in/descp.do?appid=1045&param=app.

[6]Test for the signs of autism: https://play.google.com/store/apps/details?id=ru.atec&hl=pt_BR&gl=RU.

[7]Inquiry Health LLC is an Android developer that has been active since 2014. The current app portfolio contains 31 apps https://play.google.com/store/apps/details?id=com.autismadult.test.

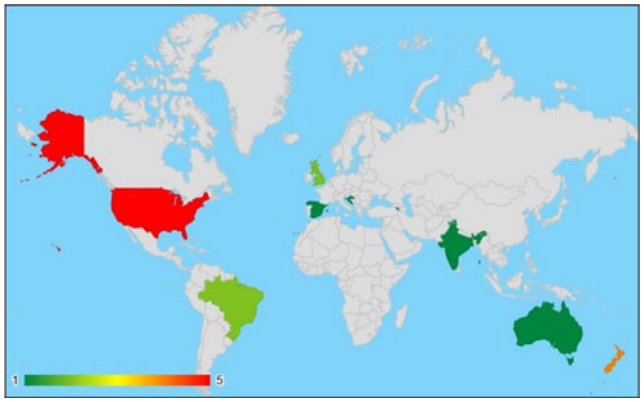
[8]Autism Quiz: AQ-10 Quotient Test: a tool to support self-referrals https://play.google.com/store/apps/developer?id=Theodore+Tollet.

[9]Developed by Android in London, United Kingdom and last updated in August 2016 https://play.google.com/store/apps/details?id=com.androidinlondon.autismtests.

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



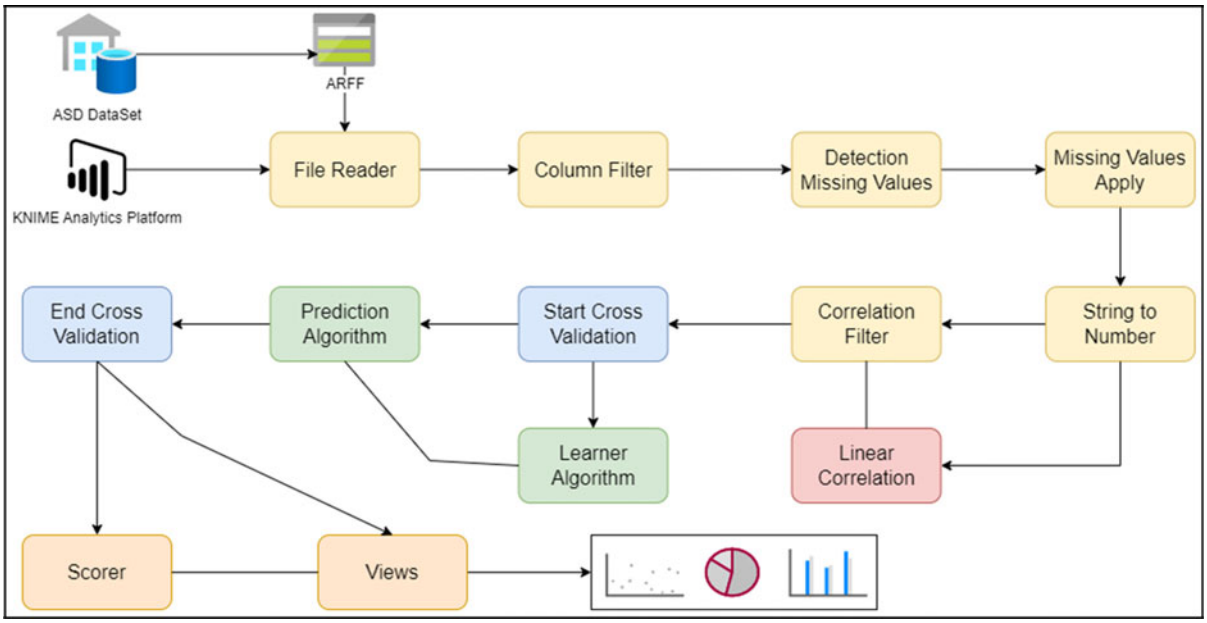
2.4. Workflow methodological process

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.

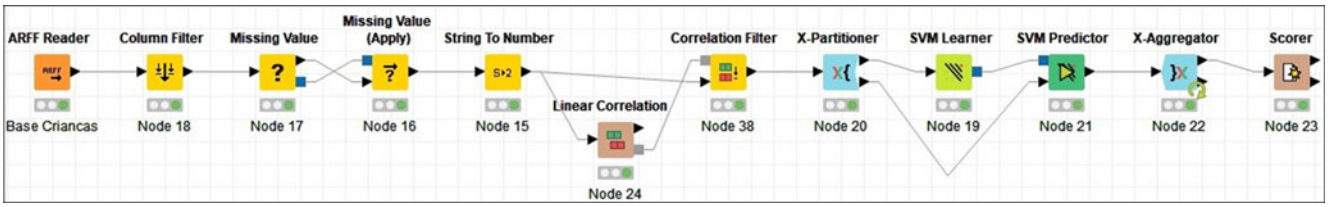
At the conclusion of the entire process, the analysis and assimilation of results areconducted. This involves performing a validation (as explained in the next section) for each of the models. The goal is to rank the most effective algorithms for extracting behavioral characteristics of individuals with 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 **String to Number** is responsible for converting string values (*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.
12. 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.
    (a) 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 is

Figure 4
Box plot for children's age and test result binning by ASD class

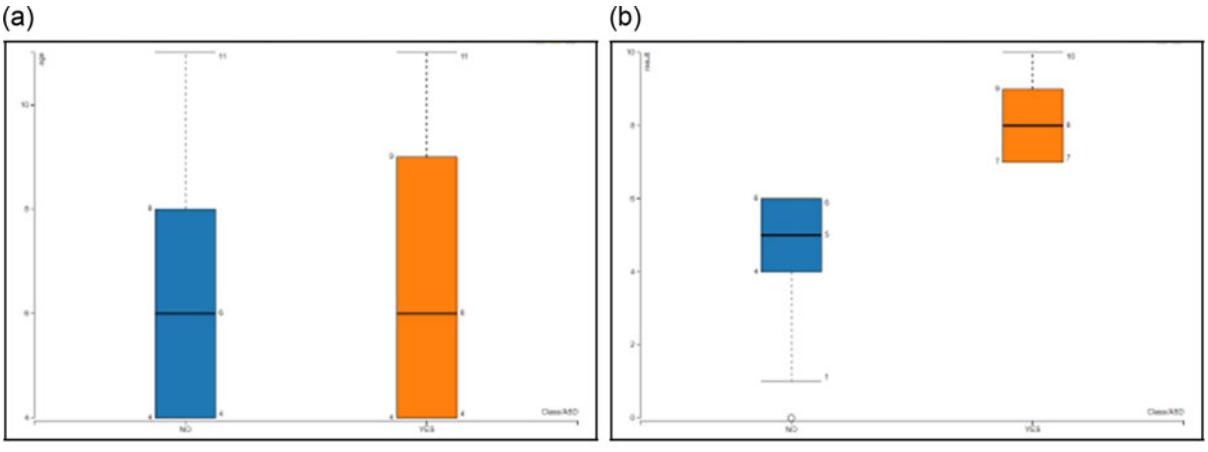
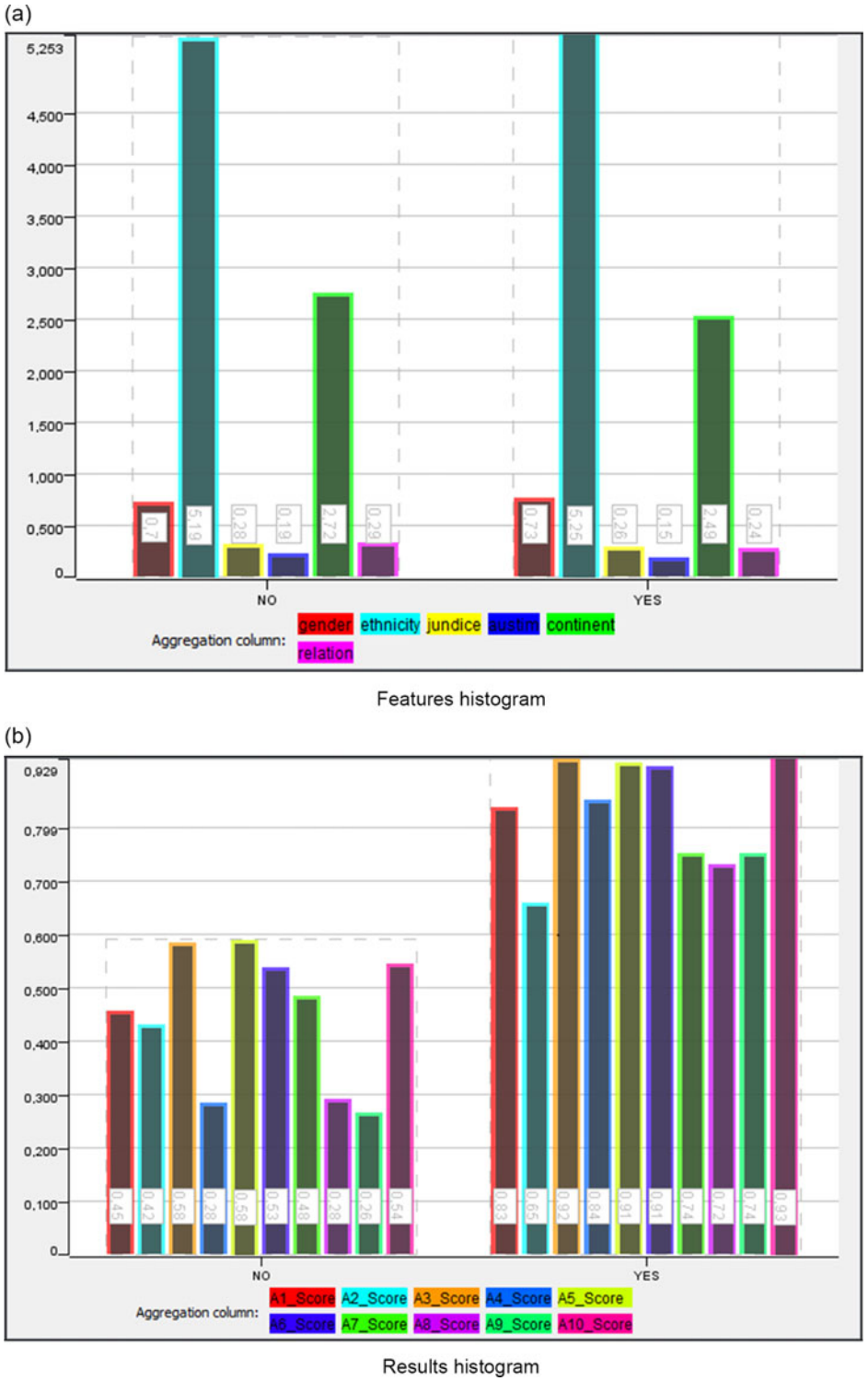


Figure 5
Features and results histogram for children database

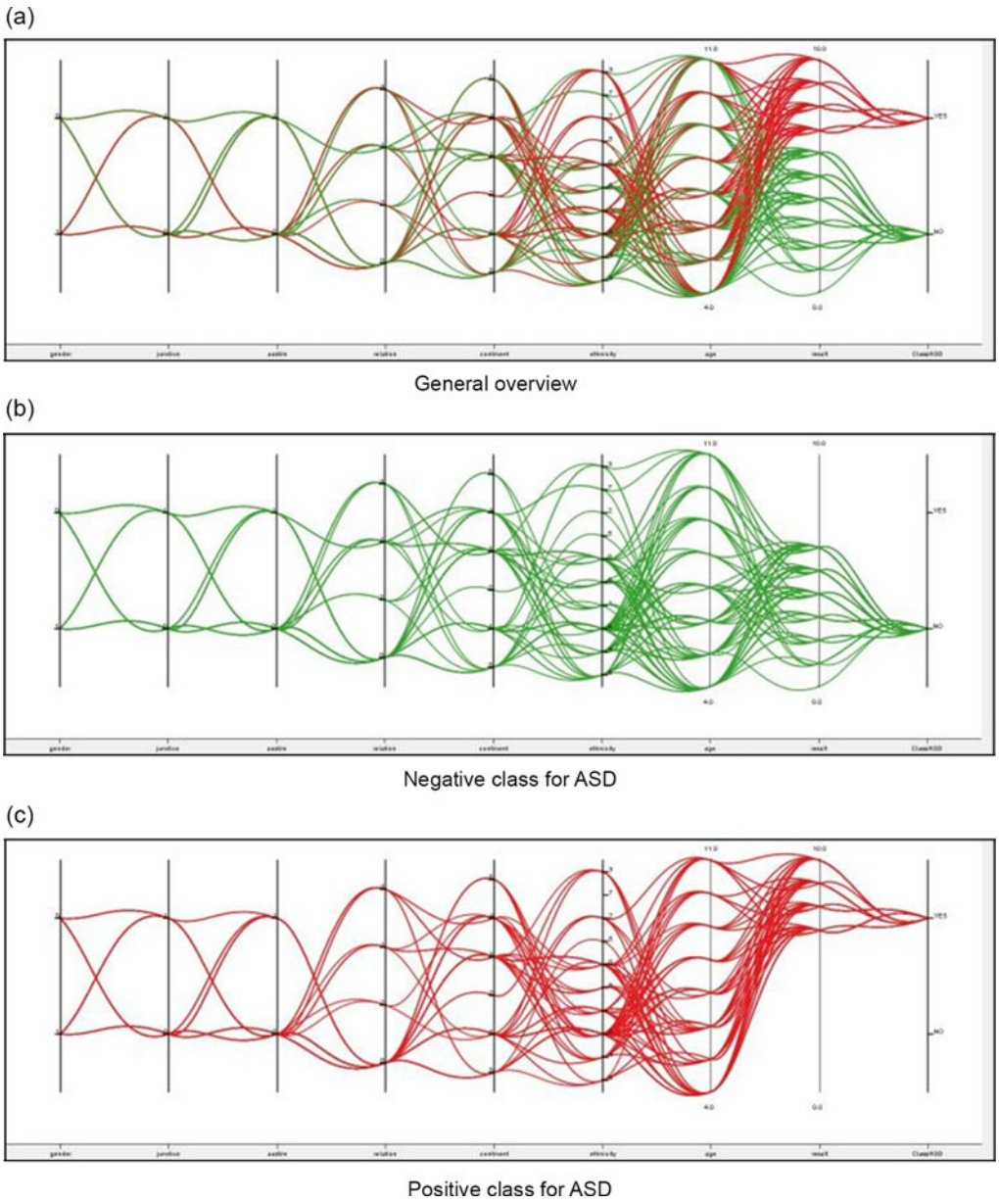


represented by a line segment that intersects these parallel lines at specific positions.

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

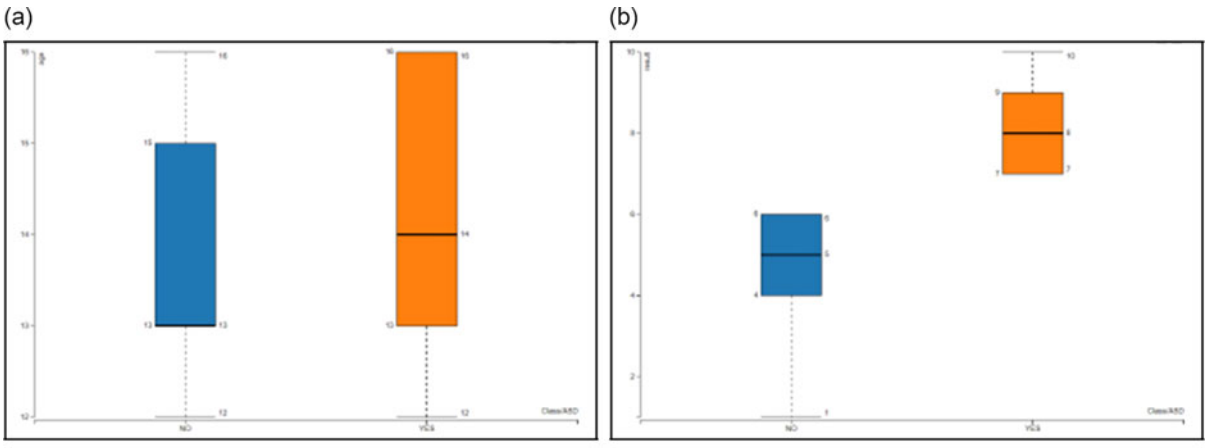


Table 2
Results with the different supervised learning techniques for the children's database

Algorithm	Results					
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Naive Bayes	48.973%	51.027%	−0.013	48.63%	51.37%	−0.028
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	49.315%	50.685%	−0.002	53.425%	46.575%	0.057
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Decision Tree	46.233%	53.767%	−0.078	58.219%	41.781%	0.163
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	58.562%	41.438%	0.173	55.137%	44.863%	0.099
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
SVM	50%	50%	−0.01	46.575%	53.425%	−0.078
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	47.945%	52.055%	−0.043	52.397%	47.603%	0.032
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
MLP	45.548%	54.452%	−0.094	56.849%	43.151%	0.136
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	55.479%	44.521%	0.109	55.822%	44.178%	0.11
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Random Forest	53.767%	46.233%	0.068	58.219%	41.781%	0.162
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	51.37%	48.63%	0.024	55.822%	44.178%	0.11
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Tree Ensemble	52.397%	47.603%	0.039	55.479%	44.521%	0.109
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	51.027%	48.973%	0.015	57.534%	42.466%	0.145

(*Continued*)

Table 2
(*Continued*)

Algorithm	Results					
	No filter			**Filter = 0.1**		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
PNN	46.918%	53.082%	−0.067	60.274%	39.726%	0.205
	Filter = 0.15			**Filter = 0.05**		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	58.219%	41.781%	0.163	57.534%	42.466%	0.145
	No filter			**Filter = 0.1**		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Gradient Boosted	51.37%	48.63%	0.027	58.904%	41.096%	0.177
	Filter = 0.15			**Filter = 0.05**		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	57.534%	42.466%	0.15	58.219%	41.781%	0.162

adolescents who answered the questionnaire (60.58%) belong to the YES class. In Figure 7(a), we have that the median for the class NO is $\tilde{x} = 13$ years old, and for the class YES, it is $\tilde{x} = 14$ years old. In Figure 7(b), we have the median for the class NO as $\tilde{x} = 5$ and the median for the class YES equal to $\tilde{x} = 8$. In Figure 8, we have the box plots for the analysis of some of the attributes and also for the questions answered in the AQ-10. In Figure 8(a), we have the variables gender, ethnicity, jaundice, autism, continent, and relation. In this case, ethnicity was the attribute that most varied between the classes YES and NO. For the same reason, the variable continent, although not having such a significant difference, also presented a variation between the positive and negative classes for the ASD. In Figure 8(b), we have the histogram that demonstrates the difference between the responses for the classes YES and NO for adolescents with ASD. In this sense, for adolescents who do not have (NO class) ASD, the answer with the highest average was the answer to the question A_1 that most people hear small sounds while other people do not, and the lowest average was A_7 question that represents the question: "When I'm reading a story I find it difficult to work out the characters' intentions." For the positive class YES, we have the question A_2 which, on average, has the lowest number of positive responses for those among adolescents with ASD, as well as in the base of children. On the other hand, Question A5 pertains to the ability to read between the lines when engaging in conversation. Figure 9 presents a parallel coordinates plot that illustrates the overall representation of attributes for the positive and negative classes of ASD in teenagers. In Figure 9(a), the parallel lines represent the attributes specific to teenagers, including gender, jaundice, autism in the family, relation, continent, ethnicity, age, result (ranging from 0 to 10), and the classes YES and NO. Figure 9(b) focuses on the data points belonging to teenagers in the NO class. This plot comprises 41 data points, showcasing how the attributes vary for individuals who do not have ASD. On the other hand, Figure 9(c) displays the parallel coordinates plot specifically for positive cases of ASD, represented by the YES class. Here, we observe how the attributes align for teenagers diagnosed with ASD.

Similar to the observations made with children, the primary differentiating factor between the positive and negative classes for teenagers is the "result" attribute. The "result" attribute plays a crucial role in distinguishing individuals with ASD (YES class) from those without (NO class). By examining these PC plots, we gain insights into the patterns and relationships between attributes in determining the classification of teenagers into the (YES, NO) classes for ASD.

Table 3 contains data pertaining to the individual characteristics of adolescents as assessed by the questionnaire. It includes 8 attributes, and the results have been analyzed using correlation filters. The adolescent database has a total of only 104 data. The results refer to data from the adolescent database, according to the settings in Table 1 From the results of the second table, the algorithm that presented the best performance was the SVM with the correlation filter at 0.15 (considering ethnicity, incidence of jaundice, and continent) with a value of 66,346% of accuracy in the results, being also the algorithm and filter with the best result in relation to the Cohen kappa coefficient ($k = 0.298$).

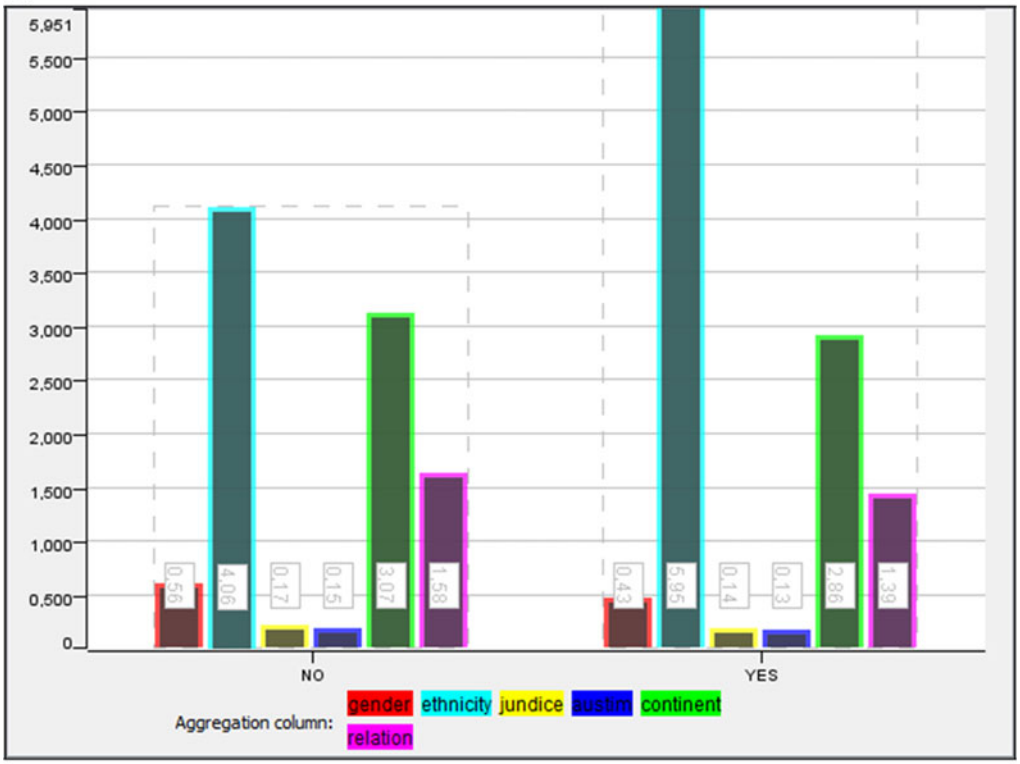
In this sense, the classification with the lowest performance was using the DT algorithm, with an accuracy of 52.885% and Cohen kappa $k = 0.008$, when using the correlation filter (= 0.05). On the other hand, the most relevant attributes for the classification were as follows: jaundice, autism predisposition, prior use of app, and continent. This means that, in a way, these attributes have a certain correlation, since ethnicity is often directly related to the continent where that person lives or was born.

4.3. Experiments performed on the adult database

In this section, the data visualization for the set of adults, as well as the data visualization, will be presented. The first data visualization is shown in Figure 10. Considering our adult database, we have a total of 704 individuals. Among them, 515 adults do not have ASD (class NO), while 189 adults are classified as positive for ASD (class YES). In Figure 10(a), the median age for those without ASD is $\tilde{x} = 26$ years old. In the NO class, the lower quartile of 21 years of age, the upper quartile of 33 years of age, and the upper whisker of 50 years of age were also observed, but with some outlier values, since the maximum observed value was 64 years of age, whereas for the positive class it is $\tilde{x} = 30$ years of age, with the lower quartile of 22 years of age and the upper quartile of 38 years of age, and upper whisker of 61 years of age. The minimum for both classes was 17 years of age. In Figure 10(b), we have the results for the class NO which has an average of 4.0 points in the AQ-10 test (smaller than the databases of children and adolescents), while for the class YES this value has an average of 8.0, the same value found for the results of the bases of children and adolescents.

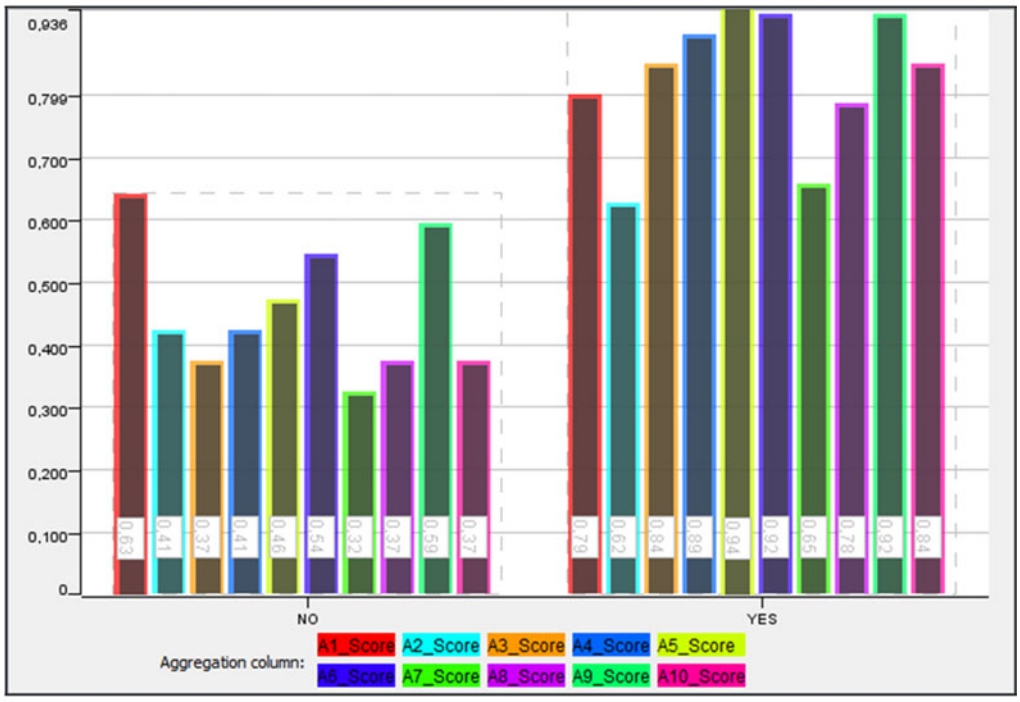
Figure 8
Features and results histogram for adolescents database

(a)



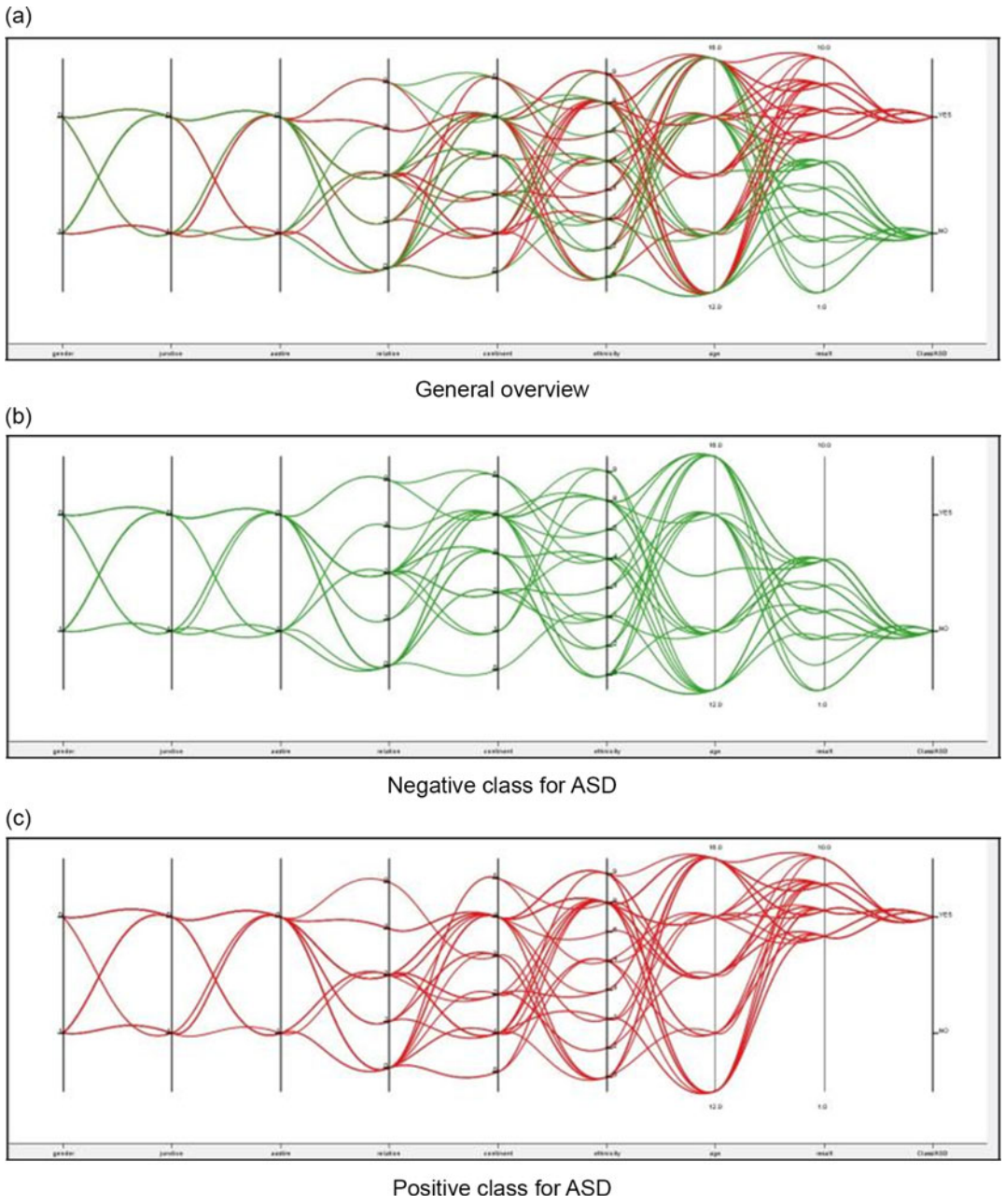
Features histogram

(b)



Result histogram

Figure 9
Parallel coordinates for adolescents database including positive and negative classes for ASD



In Figure 11, we have the data plot for 6 data attributes: gender, ethnicity, jaundice, autism predisposition, continent, and relation. As in the two previous databases, the data attributes continent and ethnicity are the attributes that most differ between the classes YES and NO, as can be seen in Figure 11(a). Furthermore, here it was also possible that the predisposition to autism could be considered a determining factor for autism as well as jaundice. These 4 attributes must be considered, as adults tend to respond better to the ASD test, so these attributes may be relevant when analyzing the complete database. In Figure 11(b), we have the results for the AQ-10 test, in which, as in the other databases, we have that the attribute A_2 is the one with the lowest weight for the class YES, while the attributes A_1: about hearing sounds while nobody else listens and the question A_5: about finding it difficult to read between the lines while someone is talking to the adult interviewed. For the NO class, the attribute A_1 is the one with the highest average of positive responses among the respondents, while the attribute A_6 is the one with the lowest average among the respondents, in this case, from the respondent know how to tell if someone listening to me is getting bored, which means that we have some results similar to what we found in the previous bases. However, when we join the bases, as will be explained in the next section, the attributes of this database will have more weight than the previous bases, since this base is 64% of the total of the lines present in the three databases.

Table 3
Results with the different supervised learning techniques for the adolescent base

Algorithm	Results					
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Naive Bayes	58.654%	41.346%	0.04	65.385%	34.615%	0.243
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	58.654%	41.346%	0.022	62.5%	37.5%	0.068
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Decision Tree	65.385%	34.615%	0.281	61.538%	38.462%	0.166
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	57.692%	42.308%	0.083	52.885%	47.115%	−0.008
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
SVM	63.462%	36.538%	0.215%	64.423%	35.577%	0.252
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	66.346%	33.654%	0.298	61.538%	38.462%	0.039
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
MLP	55.769%	44.231%	0.089	58.654%	41.346%	0.1
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	61.538%	38.462%	0.159	58.654%	41.346%	0.049
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Random Forest	63.462%	36.538%	0.215	59.615%	40.385%	0.109
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	62.5%	37.5%	0.169	58.654%	41.346%	0.04
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Tree Ensemble	58.615%	40.385%	0.147	56.731%	43.269%	0.005
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	64.423%	35.577%	0.219	58.654%	41.346%	0.049
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
PNN	65.385%	34.615%	0.275	58.654%	41.346%	0.108
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	57.692%	42.308%	0.114	55.769%	44.231%	0.024
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Gradient Boosted	56.731%	43.269%	0.09	57.692%	42.308%	0.083
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	55.769%	44.231%	0.058	59.615%	40.385%	0.132

Figure 12 visualizes the dataset consisting of 704 data points related to adult respondents who completed the Thabtah (2017) questionnaire. The plot represents various attributes, including gender, jaundice, autism in the family, relation, continent, ethnicity, age, result (ranging from 0 to 10), and the classes YES and NO. Figure 12(a) provides an overview of the data for both classes, YES and NO. Figure 12(b) specifically focuses on the data points belonging to the NO class, while Figure 12(c) focuses on the YES class, which corresponds to positive cases for ASD.

By examining the parallel coordinate plots, we can observe patterns and distinctions between the two classes. Notably, the attribute that shows the most differentiation between ASD-positive and ASD-negative cases is the "result" attribute. For values greater than or equal to 7, the data points tend to indicate a

Figure 10
Box plot for adult's age and test result binning by ASD class

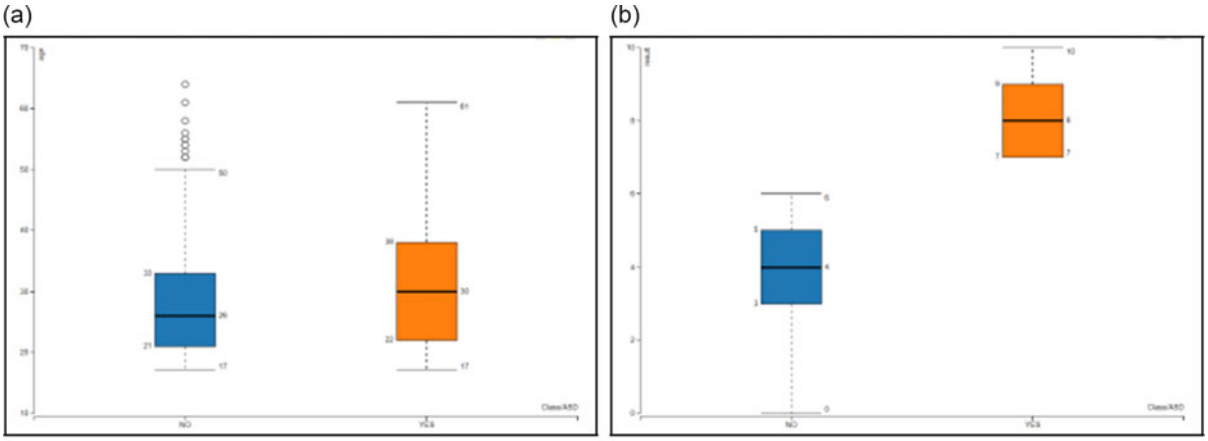


Figure 11
Features and results histogram for adults database

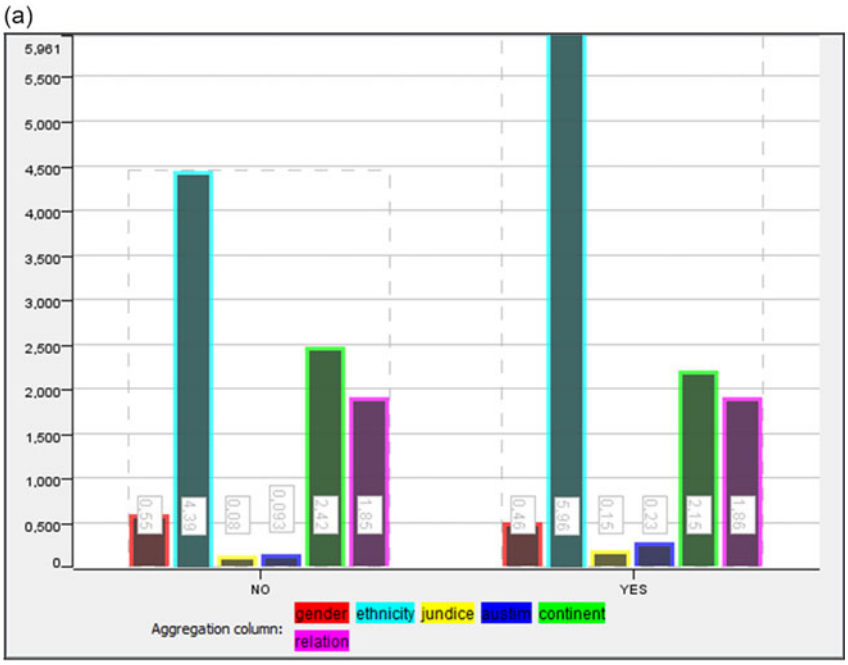
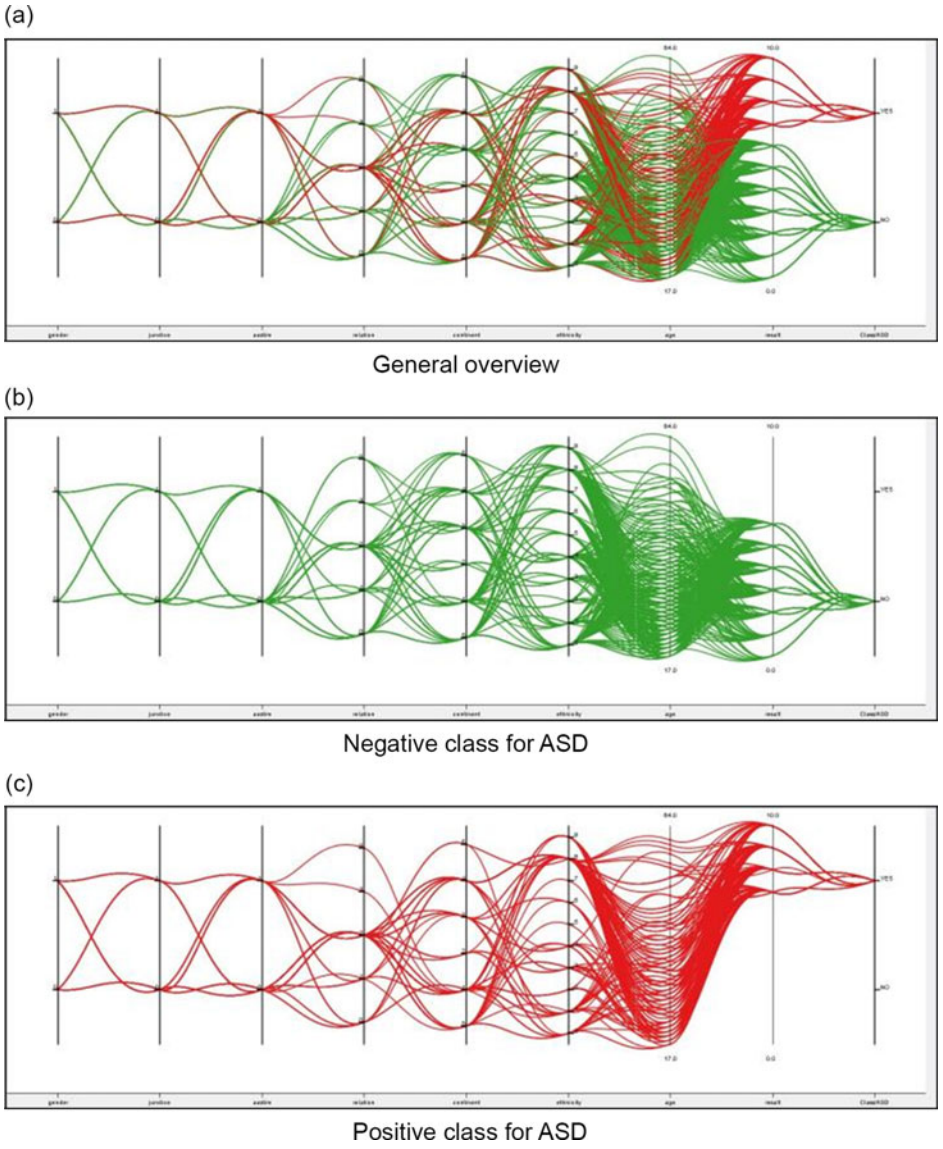


Figure 12
Parallel coordinates for adults database including positive and negative classes for ASD



positive case of autism (YES class), while values below 7 suggest a negative diagnosis for ASD (NO class). These visualizations enable a comprehensive analysis of the dataset, allowing us to identify key attributes that contribute to the classification of individuals into YES and NO classes for ASD.

Table 4 shows the mean accuracy data of the 10 simulations using 8 attributes and 8 algorithms for the adult database. The adult database has a total of 704 data. Four correlation filters were used, one of which (= 1) considers all 8 attributes, the other 3 filters according to Table 1. The PNN was the algorithm with the best performance in the analysis performed with data from adult patients (considering Table 1) when a correlation filter (= 1) is not used with the value of 73.864% accuracy in the results, the algorithm also obtained a good result in relation to the Cohen kappa coefficient ($k = 0.135$), but the best Cohen's kappa is $k = 0.198$, for DT learner when no filter is applied. The worst performance was given by the algorithm RFL with an accuracy of 58.654% and Cohen kappa $k = 0.04$, when using the correlation filter (= 0.05), but the worst Cohen's kappa is $k = 0.023$, for SVM algorithm when no filter is applied.

4.4. Complete database

In this section, we will work considering the results obtained in the previous subsections. In this sense, we unite the three public bases of Thabtah (2017). By joining the three previous bases, we got a total of 1100 given. So, in an attempt to find the most relevant personal attributes, we used the correlation filters from the previous section. In this sense, it was shown that they are presented in Table 5.

From this merger, we started to consider one more attribute that is the Age_desc (age description). In other words, we started to consider 9 attributes for our analysis, as can be seen in Table 5. Using the complete database, the MLP algorithm performed better (see Table 5) with the correlation filter at 0.15 (considering ethnicity, genetic inheritance, continent, and age group) with the

Table 4
Results with the different supervised learning techniques for the adult database

Algorithm	Results					
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Naive Bayes	71.165%	28.835%	0.018	72.727%	27.273%	0.105
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	72.443%	27.557%	0.147	73.58%	26.42%	0.023
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Decision Tree	70.17%	29.83%	0.198	71.591%	28.409%	0.211
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	69.176%	30.824%	0.172	68.182%	31.818%	−0.014
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
SVM	71.733%	28.267%	−0.023	73.153%	26.847%	0
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	71.733%	28.267%	−0.018	73.153%	26.847%	0
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
MLP	69.886%	30.144%	0.111	73.153%	26.847%	0.117
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	73.295%	26.705%	0.187	72.727%	27.273%	0.006
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Random Forest	70.881%	29.119%	0.086	72.443%	27.557%	0.019
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	59.615%	40.385%	0.109	58.654%	41.346%	0.04
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Tree Ensemble	71.165%	28.835%	0.095	72.443%	27.557%	0.019
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	72.585%	27.415%	0.022	72.869%	27.131%	−0.001
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
PNN	73.864%	26.136%	0.135	71.023%	28.977%	0.167
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	72.159%	27.841%	0.098	70.028%	29.972%	−0.003
		No filter			**Filter = 0.1**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
Gradient Boosted	69.602%	30.398%	0.146	70.028%	29.972%	0.156
		Filter = 0.15			**Filter = 0.05**	
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	69.744%	30.256%	0.148	69.318%	30.682%	−0.007

value of 67.00% accuracy in its result, also having the best result in relation to the Cohen kappa coefficient ($k = 0.248$).

Among the tested algorithms, the PNN algorithm exhibited the lowest performance, achieving an accuracy of 62.147% and a Cohen's kappa value of $k = 0.115$ when utilizing a correlation filter threshold of 0.05. Conversely, the SVM algorithm yielded the lowest Cohen's kappa value of $k = 0$ when correlation filters of 0.1 and 0.15 were employed. These findings highlight the significance of various attributes in the classification process, including genetic predisposition to autism, prior app usage, continent, ethnicity, and age description, all of which contribute to identifying the target class of ASD.

Although the previous use of the app is a relatively interesting attribute, we decided not to use it in our next approach, since this may not be a universal attribute to define the characteristics of an individual.

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%	0.071	64.273%	35.727%	0.000
	Filter = 0.15			**Filter = 0.05**		
	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%	0.202	64.727%	35.273%	0.162
	Filter = 0.15			**Filter = 0.05**		
	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%	0.198	65.727%	34.273%	0.186
	Filter = 0.15			**Filter = 0.05**		
	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%	0.196	65.182%	34.818%	0.177
	Filter = 0.15			**Filter = 0.05**		
	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%	0.2	64.422%	35.578%	0.151
	Filter = 0.15			**Filter = 0.05**		
	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%	0.213	64.727%	35.273%	0.192
	Filter = 0.15			**Filter = 0.05**		
	Accuracy	Error	Kappa	Accuracy	Error	Kappa
	65.727%	34.273%	0.224	63.273%	36.727%	0.159

4.5. 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	0.916318	0.824349
	YES	334	46	542	34	0.907609	0.921769		
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

Currently, ASD has significantly increased its number of cases, and because of this, many families have difficulties in getting the diagnosis at the appropriate time for treatment to begin, since late diagnosis can cause greater damage to development of the child, while early diagnosis has enormous advantages, preventing even the complete emergence of ASD. Taking these aspects into account, it has recently been possible to observe that one of the fundamental points in research on ASD is precisely the improvement in the development of diagnostic tools to reduce their waiting time, thus providing patients, quick access to treatment. In view of these goals, researchers have been adopting methods based on machine learning; however, it should be remembered that the application of these methods for the discovery and diagnosis of ASD is still in an introductory phase, yet they are able to demonstrate very favorable responses (Thabtah et al., 2022).

In this paper, we took on the challenge of demonstrating the importance of data science for the classification of ASD, due to the possible expansion of diagnostic methods and the development of applications capable of tracking using specifically appropriate attributes that can increase the predictive accuracy without affecting its effectiveness.

In the DM stage, it was demonstrated using several data mining algorithms in databases of up to 1100 data (children: 292 data, teenagers: 104 data, adults: 704 data, and complete: 1100 data), that the behavioral and individual characteristics related to the performance of screenings that make up the bases used are extremely relevant. Thus, it was shown that characteristics related to ethnicity, continent, and jaundice at birth help in the accuracy of the diagnosis, and these variables should be considered both by health specialists and by application developers. The factors that should be disregarded by people in the field of neuroscience and health are the attributes linked to the patient's age and previous use of the application, which presented the worst results.

Considering the factors mentioned thus far, including the economic impact of ASD and the substantial costs of medical assistance, along with the limitations of existing tools that have not yielded satisfactory outcomes, there is an evident need for accessible screening procedures. These procedures would facilitate self-assessment (or assessments guided by family or educators) and support healthcare professionals in their evaluations. As a result, it was proposed in this work the isolation of variables (demographic, geographical and individual) and behavioral characteristics to carry out the identification of autism traits through more accurate screenings, capable of providing a timely prior diagnosis that can be used by a series of users. Thus, it will provide those who need faster guidance to specialist doctors. The isolated variables (ethnicity, continent, autism genetic predisposition, and jaundice) can be incorporated into the ASD Tests application proposed by Thabtah (2017). When these variables are combined with the 10 behavioral questions from the application, they can improve effectiveness by approximately 72% in diagnosing ASD across different age groups. Furthermore, when the AQ-10 is integrated with classification algorithms, the SVM learner achieves a perfect accuracy rate of 100%.

In future research, it would be valuable to delve deeper into the individual impact of each variable (demographic, geographical, and individual) on behavioral features within the dataset. Understanding the independent contributions of these variables can provide valuable insights into the intricate relationship between various factors and the manifestation of ASD. Furthermore, investigating the potential interactions and correlations between these variables can enhance our understanding of the complexity of ASD diagnosis. Additionally, exploring the longitudinal aspects of the dataset and considering the influence of time could shed light on the evolution of behavioral patterns in individuals with ASD. These future endeavors can contribute to a more comprehensive understanding of the intricate dynamics between demographic, geographical, and individual factors and their influence on the behavioral characteristics of individuals on the autism spectrum.

Funding Support

The research project received funding from various sources, including PROPI, CNPq, CAPES, and FAPEMIG. These organizations played a vital role in supporting the research conducted by author DAL, providing essential financial resources to carry out the project successfully.

Authors' Contributions

RSD took charge of executing and simulating the results within the database. Additionally, RSD was responsible for collecting and analyzing the data, conducting simulations and tests, and generating the results showcased in this paper. On the other hand, DAL played a pivotal role in writing the paper, translating it, and providing guidance and mentorship to RSD throughout the research process.

Availability of Data and Materials

We are not dealing with humans in our research, as the databases used in our research are open, public, scientific database available on the UCI Machine Learning Repository website[10]. In this article, we used an open public database deposited in the repository.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202–212.

Assumpção Jr, F. B., & Pimentel, A. C. M. (2000). Autismo infantil. *Brazilian Journal of Psychiatry*, *22*, 37–39.

Aziz, M. Z., Abdullah, S. A., Adnan, S. F., & Mazalan, L. (2014). Educational app for children with autism spectrum disorders (ASDS). *Procedia Computer Science*, *42*, 70–77.

[10]Adult Data Set https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult, Children Data Set https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult and Adolescents Data Set https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++.

Baruh, L., & Popescu, M. (2017). Big data analytics and the limits of privacy self-management. *New Media & Society*, *19*(4), 579–596.

Biswas, M., Kaiser, M. S., Mahmud, M., Al Mamun, S., Hossain, M. S., & Rahman, M. A. (2021). An XAI based autism detection: The context behind the detection. In *Brain Informatics: 14th International Conference Proceedings, 14,* 448–459.

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. D. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, *137*.

Chagas, F. (2016). Autismo: Caracterização e classifica-ção do grau de severidade dos alunos da associação maringaense dos autistas (AMA) com base no método cars. *Brazilian Journal of Surgery and Clinical Research*, *15*(3), 37–41.

Chakraborty, S., Thomas, P., Bhatia, T., Nimgaonkar, V. L., & Deshpande, S. N. (2015). Assessment of severity of autism using the Indian scale for assessment of autism. *Indian Journal of Psychological Medicine*, *37*(2).

Crane, L., Chester, J. W., Goddard, L., Henry, L. A., & Hill, E. (2016). Experiences of autism diagnosis: A survey of over 1000 parents in the United Kingdom. *Autism*, *20*(2), 153–162.

Davis, M., Otero, N., Dautenhahn, K., Nehaniv, C. L., & Powell, S. D. (2007). Creating a software to promote understanding about narrative in children with autism: Reflecting on the design of feedback and opportunities to reason. In *IEEE 6th International Conference on Development and Learning*, 64–69.

de Carvalho, T. B. A., Sibaldo, M. A. A., Tsang, R., & da Cunha Cavalcanti, G. D. (2017). Principal component analysis for supervised learning: A minimum classification error approach. *Journal of Information and Data Management*, *8*(2), 131–131.

de Morais, J. I., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., & Barbon Jr, S. (2020). A multi-label classification system to distinguish among fake, satirical, objective and legitimate news in Brazilian Portuguese. *iSys - Brazilian Journal of Information Systems*, *13*(4), 126–149.

Eder, M. S., Diaz, J. M. S., Madela, J. R., Magusara, M., & Sabellano, D. D. (2016). Fill Me App: An interactive mobile game application for children with autism. *International Journal of Interactive Mobile Technologies*, *10*(3), 59–63. https://doi.org/10.3991/ijim.v10i3.5553

Farias, E., & Cunha, M. (2013). Protótipo de uma ferramenta de software para apoio no tratamento de crianças com autismo. In *Anais do IX Simpósio Brasileiro de Sistemas de Informação*, 332–342.

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). Knime for reproducible cross-domain analysis of life science data. *Journal of Biotechnology*, *261*, 149–156.

Fletcher-Watson, S., Pain, H., Hammond, S., Humphry, A., & McConachie, H. (2016). Designing for young children with autism spectrum disorder: A case study of an ipad app. *International Journal of Child-Computer Interaction*, *7*, 1–14.

Garai, S., & Paul, R. K. (2023). Development of MCS based ensemble models using Ceemdan decomposition and machine intelligence. *Intelligent Systems with Applications*, *18*.

Garai, S., Paul, R. K., Rakshit, D., Yeasin, M., Paul, A., Roy, H., . . . , & Manjunatha, B. (2023). An MRA based MLR model for forecasting Indian annual rainfall using large scale climate indices. *International Journal of Environment and Climate Change*, *13*(5), 137–150.

Goulart, P., & de Assis, G. J. A. (2002). Estudos sobre autismo em análise do comportamento: Aspectos metodológicos. *Revista Brasileira de Terapia Comportamental e Cognitiva*, *4*(2), 151–165.

Guandaline, V. H., & de Campos Merschmann, L. H. (2017). Hcaim: A discretizer for the hierarchical classification scenario applied to bioinformatics datasets. *Journal of Information and Data Management*, *8*(2), 146.

Kanner, L. (1965). Infantile autism and the schizophrenias. *Behavioral Science*, *10*(4), 412–420.

Klin, A. (2006). Autismo e síndrome de asperger: Uma visão geral. *Brazilian Journal of Psychiatry*, *28*, s3–s11.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Articial Intelligence*, *14*(2), 1137–1145.

Lima, D. A., Ferreira, M. E. A., & Silva, A. F. F. (2021). Machine learning and data visualization to evaluate a robotics and programming project targeted for women. *Journal of Intelligent & Robotic Systems*, *103*(1), 1–20.

Lima, D. A., & Isotani, S. (2022). Systematic map and review of google classroom usage during the covid-19 pandemic: An analysis by data clustering approach. *Revista Brasileira de Informática na Educação*, *30*, 20–49.

Lopes, H. J., & Lima, D. A. (2021). Evolutionary tabu inverted ant cellular automata with elitist inertia for swarm robotics as surrogate method in surveillance task using e-puck architecture. *Robotics and Autonomous Systems*, *144*.

Lozano-Martínez, J., Ballesta-Pagán, F. J., & AlcarazGarcía, S. (2011). Software para enseñar emociones al alumnado con trastorno del espectro autista. *Comunicar: Revista Científica de Comunicación y Educación*, *18*(36), 139–148.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Paul, R. K., & Garai, S. (2022). Wavelets based artificial neural network technique for forecasting agricultural prices. *Journal of the Indian Society for Probability and Statistics*, *23*(1), 47–61.

Shahamiri, S. R., & Thabtah, F. (2020). Autism AI: A new autism screening system based on artificial intelligence. *Cognitive Computation*, *12*(4), 766–777.

Siegel, B. (2021). Pervasive developmental disorders screening test (PDDST). *Encyclopedia of Autism Spectrum Disorders*, 3447–3451.

Sousa, F. R. M., Costa, E. A. B., & de Castro, T. H. C. (2012). Worldtour: Software para suporte no ensino de crianças autistas. In *Brazilian Symposium on Computers in Education, 23*(1).

Stone-Heaberlin, M., Rouse, M. L., Blake, H. S., Fodstad, J. C., Smith, J., Kerswill, S., & Bushnell, E. (2022). Measuring feeding disorders in individuals with autism and pervasive developmental disorders. In J. L. Matson, & P. Sturmey (eds), *Handbook of Autism and Pervasive Developmental Disorder.* Springer.

Thabtah, F. (2017). Autism spectrum disorder screening: Machine learning adaptation and dsm-5 fulfillment. In *Proceedings of the 1st International Conference on Medical and Health Informatics*, 1–6.

Thabtah, F., Spencer, R., Abdelhamid, N., Kamalov, F., Wentzel, C., Ye, Y., & Dayara, T. (2022). Autism screening: An unsupervised machine learning approach. *Health Information Science and Systems*, *10*(1), 1–13.

Vieira, N. M., & Baldin, S. R. (2017). Diagnóstico e intervenção de indivíduos com transtorno do espectro autista. *Encontro Internacional de Formação de Professores e Fórum Permanente de Inovação Educacional*, *10*(1).

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, S., Saxena, S., . . . , & Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. *Autism Research*, *15*(5), 778–790.

How to Cite: Dornelas, R. S. & Lima, D. A. (2023). Correlation Filters in Machine Learning Algorithms to Select Demographic and Individual Features for Autism Spectrum Disorder Diagnosis. *Journal of Data Science and Intelligent Systems* 1(2), 105–127, https://doi.org/10.47852/bonviewJDSIS32021027