UNIVERSITY OF THE STATE OF SANTA CATARINA
CENTER FOR TECHNOLOGICAL SCIENCES
GRADUATE PROGRAM IN APPLIED COMPUTING

MAÍSA FERNANDES GOMES

A DATA-DRIVEN APPROACH FOR AUTISM SPECTRUM DISORDER SCREENING

JOINVILLE
2025

MAÍSA FERNANDES GOMES

A DATA-DRIVEN APPROACH FOR AUTISM SPECTRUM DISORDER SCREENING

Dissertation submitted to the Graduate Program in Applied Computing of the Technological Sciences Center of the University of the State of Santa Catarina, for the attainment of the degree of Master in Applied Computing.

Orientador: Dr. Rafael Stubs Parpinelli

JOINVILLE
2025

Ficha catalográfica elaborada pelo programa de geração automática da
Biblioteca Universitária Udesc,
com os dados fornecidos pelo(a) autor(a)

Fernandes Gomes, Maisa
A data-driven approach for autism spectrum disorder screening /
Maisa Fernandes Gomes. -- 2025.
113 p.

Orientador: Rafael Stubs Parpinelli
Dissertação (mestrado) -- Universidade do Estado de Santa
Catarina, Centro de Ciências Tecnológicas, Programa de
Pós-Graduação em Computação Aplicada, Joinville, 2025.

1. Artificial Intelligence. 2. Machine Learning. 3. Autism
Spectrum Disorder. 4. Diagnostic Assistant. 5. Interpretable Models.
I. Stubs Parpinelli, Rafael. II. Universidade do Estado de Santa
Catarina, Centro de Ciências Tecnológicas, Programa de
Pós-Graduação em Computação Aplicada. III. Título.

Maísa Fernandes Gomes

A data-driven approach for autism spectrum disorder screening:

Dissertation submitted to the Graduate Program in Applied Computing of the Technological Sciences Center at the State University of Santa Catarina as a partial requirement for obtaining the title of Master in Applied Computing, in the field of concentration in Methodology and Techniques of Computing.

Examination Board

Dr. Rafael Stubs Parpinelli
UDESC

Members:

Dr. Avanilde Kemczinski
UDESC

Dr. Hugo Valadares Siqueira
UTFPR

Joinville, February 25, 2025

To my parents, whose love and guidance have been the foundation of everything I am.

"No problem can be solved from the same level of consciousness that created it."

Albert Einstein

RESUMO

A evolução da Inteligência Artificial (IA) e do Aprendizado de Máquina (ML) tem impulsionado avanços significativos no setor da saúde, permitindo a automação de tarefas complexas e aprimorando diagnósticos médicos. Diversos estudos na literatura exploram o uso de modelos de ML para auxiliar no diagnóstico precoce, acelerando o processo de tratamento. Um exemplo relevante é o Transtorno do Espectro Autista (TEA), uma condição atípica sem cura, mas cujo tratamento com Terapias Especiais, especialmente em idades iniciais, traz benefícios significativos ao desenvolvimento de pessoas autistas. Atualmente, o diagnóstico formal de TEA é desafiador, pois não existe uma causa única identificada. O processo envolve uma equipe multidisciplinar que realiza testes psicológicos, exames de imagem, genéticos e de sangue para auxiliar na identificação do transtorno.

Embora trabalhos existentes na literatura utilizem modelos de ML para apoiar o diagnóstico, muitos dependem de dados extraídos de exames médicos, o que nem sempre reflete a realidade de hospitais e provedores de saúde interessados no rastreamento do TEA. Esses provedores, muitas vezes, não dispõem de dados de diagnóstico formal, tornando essencial o desenvolvimento de uma base de dados alternativa que possa viabilizar análises preditivas em contextos reais. A adoção de metodologias estruturadas, como o CRISP-DM, é fundamental no desenvolvimento de modelos de ML, pois contribui para maior assertividade, eficiência e organização ao longo de todo o processo.

Neste estudo, foi aplicada a metodologia CRISP-ML para o desenvolvimento de um banco de dados voltado à identificação de beneficiários com risco de TEA, sem a necessidade de dados de diagnóstico. Foram identificadas 68 variáveis relevantes a partir de registros de uso de serviços médicos e realizada uma análise exploratória para compreender melhor o perfil dos beneficiários que solicitaram Terapias Especiais.

Para o treinamento dos modelos de ML, foram selecionadas 25 variáveis, e a classificação foi abordada de duas formas: classificação binária e classificação de uma classe. Os modelos de classificação binária Random Forest, XGBoost e CatBoost foram empregados para distinguir beneficiários que solicitaram Terapia Especial entre TEA e Não TEA. Já os modelos de classificação de uma classe, Isolation Forest e One-Class SVM, buscaram identificar beneficiários com indicação de solicitação futura de Terapia Especial. Todos os modelos foram comparados em termos de desempenho, utilizando técnicas de tuning de hiperparâmetros e seleção de variáveis.

Os modelos de classificação binária XGBoost e Random Forest, com ajuste de hiperparâmetros e seleção de variáveis, apresentaram os melhores resultados, alcançando respectivamente 73.49% e 73.61% de acurácia. Entre os modelos de uma classe, Isolation Forest obteve 85,90% de acurácia, enquanto One-Class SVM atingiu 83,07%.

Os modelos Random Forest e Isolation Forest, que demonstraram melhor desempenho, foram submetidos a um processo de interpretabilidade por meio do SHAP, em conjunto com especialistas da área. A análise revelou que as variáveis mais relevantes para os modelos estavam associadas a consultas com psicólogos, fonoaudiólogos e terapeutas ocupacionais, serviços frequentemente buscados no processo de diagnóstico do TEA.

Os resultados demonstram que a aplicação de modelos interpretáveis, combinada com técnicas de seleção de características e ajuste de hiperparâmetros, pode contribuir significativamente para a identificação precoce de beneficiários com risco de TEA. Ademais, a colaboração entre especialistas técnicos e profissionais de saúde reforça a validade e a aplicação dessas abordagens no contexto de operadoras de saúde. Essa validade foi assegurada por meio da participação ativa dos especialistas na revisão das variáveis e na análise dos resultados gerados pelos modelos, garantindo coerência com a prática clínica.

Palavras Chave: Inteligência Artificial, Aprendizado de Máquina, Transtorno do Espectro Autista, Auxilio ao Diagnostico, Modelos Interpretáveis.

ABSTRACT

The evolution of Artificial Intelligence (AI) and Machine Learning (ML) has driven significant advancements in the healthcare sector, enabling the automation of complex tasks and enhancing medical diagnostics. Various studies in the literature explore the use of ML models to assist in early diagnosis, accelerating the treatment process. A relevant example is Autism Spectrum Disorder (ASD), an atypical condition with no cure, but for which treatment with Special Therapies, especially at an early age, brings significant developmental benefits for autistic individuals. Currently, the formal diagnosis of ASD is challenging, as no single cause has been identified. The process involves a multidisciplinary team conducting psychological tests, imaging exams, genetic tests, and blood tests to aid in the identification of the disorder.

Although existing studies use ML models to support diagnosis, many rely on data extracted from medical examinations, which do not always reflect the reality of hospitals and healthcare providers interested in ASD screening. These providers often lack formal diagnostic data, making it essential to develop an alternative database that enables predictive analyses in real-world contexts. The adoption of structured methodologies, such as CRISP-DM, is essential in the development of machine learning models, as it contributes to greater accuracy, efficiency, and organization throughout the entire process.

In this study, the CRISP-ML methodology was applied to develop a database aimed at identifying beneficiaries at risk of ASD without requiring diagnostic data. A total of 68 relevant variables were identified from medical service usage records, and an exploratory analysis was conducted to better understand the profile of beneficiaries who requested Special Therapies.

For ML model training, 25 variables were selected, and classification was approached in two ways: binary classification and one-class classification. The binary classification models—Random Forest, XGBoost, and CatBoost—were employed to distinguish beneficiaries who requested Special Therapy as either ASD or Non-ASD. Meanwhile, the one-class classification models, Isolation Forest and One-Class SVM, aimed to identify beneficiaries with an indication of future Special Therapy requests. All models were compared in terms of performance, using hyperparameter tuning and feature selection techniques.

The binary classification models XGBoost and Random Forest, with hyperparameter tuning and feature selection, achieved the best results, reaching 73.49% and 73.61% accuracy, respectively. Among the one-class models, Isolation Forest obtained an accuracy of 85.90%, while One-Class SVM reached 83.07%.

The best-performing models, Random Forest and Isolation Forest, underwent an interpretability analysis using SHAP in collaboration with domain experts. The analysis revealed that the most relevant variables for the models were related to consultations with psychologists, speech therapists, and occupational therapists—services frequently sought in the ASD diagnostic process.

The results demonstrate that applying interpretable models, combined with feature selection and hyperparameter tuning techniques, can significantly contribute to the early identification of beneficiaries at risk of ASD. Moreover, the collaboration between technical experts and healthcare professionals reinforces the validity and applicability of these approaches within the context of healthcare providers. This validity was ensured through the active participation of specialists in the review of the selected variables and the analysis of the model outputs, ensuring alignment with clinical practice.

Key-words: Artificial Intelligence, Machine Learning, Autism Spectrum Disorder, Diagnostic Assistant, Interpretable Models.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Inteligence
ASD	Autism Spectrum Disorder
CRISP-DM	Cross-Industry Standard Process for Data Mining
CAT	Catboost
IF	Isolation Forest
ML	Machine Learning
OC SVM	One Class SVM
RF	Randon Forest
SHAP	SHapley Additive exPlanations
ST	Special Tearapy
WHO	World Health Organization
XGB	XGBoost

CONTENTS

1 INTRODUCTION

The advancement of artificial intelligence (AI) has significantly transformed various fields, including sectors such as technology, finance, logistics, and healthcare, where it has provided intelligent tools to improve practices and processes (ALZUBI; NAYYAR; KUMAR, 2018). The development of Machine Learning (ML) models has played a crucial role in this progress, helping automate complex tasks, predictive analysis, and more accurate decision-making (RUSSELL; NORVIG, 2020). These advancements have sparked the interest of companies seeking to integrate this innovation into their systems and processes, harnessing the potential of AI models to optimize operations and increase efficiency. To ensure that these solutions can be implemented in a structured way and with the best chances of success, the use of methodologies for structuring becomes essential. One such methodology used in data mining project developments within companies is the Cross-Industry Standard Process for Data Mining (CRISP-DM) model, which enables a systematic organization of the development process, from problem understanding and data collection to model implementation and evaluation (WIRTH; HIPP, 2000).

CRISP-DM was developed in the early 2000s, at a time when data mining was gaining more recognition and being widely applied across various industries. Organizations created it in response to the need for a standardized process that would guide them in conducting data mining projects systematically and repetitively. The model consists of a hierarchical process with six main phases, each containing both generic and specialized tasks. This flexible approach allows organizations to adapt CRISP-DM to the specific needs of each application, ensuring its effectiveness in different contexts and industries (SCHRÖER; KRUSE; GÓMEZ, 2021).

Emerging from CRISP-DM, other derived methodologies have been developed and created as an extension of the original model, such as CRISP-ML developed by Kolyshkina e Simoff (2019). CRISP-ML aims to integrate interpretability throughout all stages of the development of ML solutions, with an emphasis on contexts where model explainability is crucial, such as in healthcare. In addition to the stages related to interpretability, CRISP-ML adds steps that evaluate the predictive potential of the data and enrich the data, which ensures that the ML models are not only accurate but also interpretable. Moreover, ML projects must pay special attention to data because teams spend 80% of the time preparing the data to ensure that it is consistent and ready for ML applications. Projects utilizing CRISP-ML have demonstrated the flexibility of this methodology, which can be adapted to various application areas, adjusting to the specific needs of each context. By involving all stakeholders at each stage of the process, CRISP-ML ensures that the models meet the demands of the sector while also guaranteeing a clear and accessible interpretation of the results for professionals (KOLYSHKINA; SIMOFF, 2021), (FRAŇO, 2024).

Using methodologies like CRISP-ML is essential in ML projects, as it provides a structured process that ranges from identifying and defining the theme to implementing and creating documents and code reports. This approach stands out for promoting the interpretability of solutions, an aspect that is increasingly important in sectors where transparency is a fundamental pillar (KOLYSHKINA; SIMOFF, 2021). Healthcare, for example, benefits from advances in ML models, which contribute to the improvement of medical practices and administrative processes (STIGLIC et al., 2020). By integrating interpretability and fostering collaboration between technical teams and healthcare experts, CRISP-ML facilitates the joint development of models with hospitals and healthcare institutions. This collaborative work can optimize diagnostic procedures, improve patient management, and strengthen clinical decision-making, making it safer, more efficient, and responsible.

A healthcare operator refers to an entity directly involved in providing services related to the care and management of health, including doctors, medical procedures, exams, clinics, and hospitals. Beneficiaries are individuals who receive these health services or benefits. They include patients, individuals covered by health insurance plans, or those eligible for medical assistance programs. Healthcare providers play a vital role in the diagnosis, treatment, prevention, and monitoring of diseases, while also ensuring efficient and ethical care for patients. In the context of ML models applied to healthcare, healthcare operators collaborate to integrate technological solutions, offering clinical expertise and contributing to the interpretation of results. This collaboration ensures that solutions meet the needs of patients effectively. The area in which healthcare providers contribute is in Special Therapies (UNIMED SANTA CATARINA, 2021), which cover treatments for complex conditions such as autism spectrum disorder, intellectual disabilities, learning disorders, cerebral palsy, language and communication disorders, and neurological disorders.

Special Therapies(TS) refer to treatments designed to address the specific needs of individuals with medical, neurodevelopmental, psychological, or motor conditions. These therapies aim to promote well-being, autonomy, rehabilitation, and personal development. They involve an interdisciplinary team of professionals, including occupational therapists, psychologists, physiotherapists, speech therapists, and pediatricians. Each condition requires different therapeutic modalities tailored to the beneficiary's needs and daily routine. These therapies are essential to improve quality of life. Research shows that early starting therapy, particularly in children, leads to better outcomes (GURALNICK, 2017) (DAWSON et al., 2010) (ESTES et al., 2015).

One of the neurodevelopmental conditions that benefit from early therapeutic intervention is Autism Spectrum Disorder (ASD) (SUKIENNIK; MARCHEZAN; SCORNAVACCA, 2022), which presents challenges in communication, social interaction, and behavior, with symptoms that can vary significantly between individuals. Although the causes of ASD are not fully understood, research shows that a combination of genetic and non-genetic factors contributes to its development. ASD has no cure, and therapeutic interventions remain the gold standard for treatment. However, to begin treatment, professionals must formalize a diagnosis and create a therapy plan that is tailored to the specific needs and reality of the autistic individual (ARAUJO et al., 2019). Since there are no biological markers, clinicians diagnose ASD through a clinical process in which a team conducts investigations and requests various tests. These tests include blood work, imaging, and interviews or observations made by psychologists (ONZI; GOMES, 2015). The formal diagnosis process can be time-consuming, mainly due to two factors: the delay in seeking a professional diagnosis and the time required for the investigation process (LORD et al., 2020).

The delay in seeking a professional for the diagnosis is often associated with the time it takes parents to notice the first signs that the child's development is deviating from the expected pattern. This is because ASD has some signs that appear in the early years of a baby's life and continue to manifest as the child grows. The search for professional advice is triggered by the parents' recognition of these early symptoms, which is crucial for starting the diagnostic investigation. The delay in this search causes delays in the diagnostic process and, subsequently, in the initiation of therapy. The delay in the process is related to the fact that the diagnosis involves a team of professionals and tests, which can take time before a result is reached (LORD et al., 2018) (SHARMA; GONDA; TARAZI, 2018).

Given the importance of investigation and detection, some institutions have emphasized the need for health and education professionals to be equipped to identify children with potential early signs of ASD. To meet the need for early detection, the American Academy of Pediatrics advises universal screening for ASD at 18, 24, and 30 months. The goal is to compare developmental milestones and identify any anomalies as early indicators, leading to the pursuit of a formal diagnosis (MUKHERJEE, 2017). Some studies focus on the use of AI in genetic and imaging medical tests to identify ASD. In contrast, others emphasize the use of psychological interview data to detect individuals with ASD. All of these efforts are intended to enable early identification of the condition (HYDE et al., 2019) (THABTAH, 2019). The application of these AI solutions to help diagnose ASD could be linked to a health insurance provider looking to enhance its diagnostic processes. To achieve this, the company can supply data to help develop a system that supports the diagnostic process. Among the studies conducted, those based on classification techniques aim to distinguish between ASD and non-ASD, using imaging data such as ultrasounds, electroencephalograms, and MRI. Some rely on videos and photos of children to identify facial features that may indicate ASD. In contrast, others are based on tabular data derived from genetic test results or interview data (WASHINGTON; WALL, 2023a).

Databases based on medical test results do not reflect the reality of many healthcare providers, who typically have access only to utilization histories and personal data of beneficiaries without access to test results. This limitation renders the application of models that depend exclusively on test-based data impractical. To effectively implement machine learning (ML) models within a healthcare provider, it is crucial to evaluate the available data and identify which aspects are most relevant to accurately represent the problem. The use data, which include information on tests, consultations, and procedures requested by beneficiaries, can serve as valuable resources to identify the usage patterns associated with beneficiaries diagnosed with ASD.

This study aims to apply the CRISP-ML methodology to develop a database for ASD and subsequently train ML models using the created database while employing an interpretability technique on the resulting models. The database was designed in collaboration with a team of experts from the UNIMED Santa Catarina healthcare provider, who provided the initial data set. From these data, variables were selected and analyzed to better represent the problem without relying on a formal diagnosis. Using the created database, two ML-based approaches were developed: One-class models to identify beneficiaries at risk of requiring Special Therapies and binary classification models to identify beneficiaries at risk of ASD.

Five models were selected for training and analysis of the results. Two for the One-Class approach: Isolation Forest (IF) and One-Class SVM (OC-SVM), and three for the Binary Classification Approach: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and CatBoost. The performance of these models was compared by applying hyperparameter tuning techniques and feature selection methods. In the final stage, the models were analyzed for interpretability to facilitate discussions with the healthcare provider's team. These discussions focused on the potential of models for their intended tasks and the feasibility of integrating them into the company's internal processes.

The contributions of this work are as follows.

- Analysis of variables available in a healthcare provider for the ASD problem: Analysis of variables present in the database of a healthcare provider together with specialists to create a database that can assist in exploratory analyses to visualize the ASD theme and application of ML models within the company's system;

- Development and consolidation of a database for ASD: Consolidation of a database that is not based on a diagnosis as is most found in the literature. This is crucial for healthcare providers who do not have access to diagnostic data, allowing them to screen for ASD using these variables;
- Analysis of the prediction made by ML models trained with the created database: evaluation of ML models trained with the created database, assessing the potential of this approach for ASD screening; and
- Model interpretability in collaboration with specialists: Application of interpretability techniques to trained ML models to identify the most relevant variables and decision factors. This step enabled discussions with healthcare professionals, facilitating validation and refinement of the results, as well as exploring the feasibility of integrating the models into the company processes.

1.1 MOTIVATION

The use of Special Therapies (TS) significantly improves the quality of life of individuals with specific needs. When applied in the early years, these interventions tend to be more effective, yielding better long-term results. These therapies are typically prescribed when developmental conditions or difficulties are identified. However, access to these interventions is often delayed due to factors such as late identification of atypical developmental signs or delays in the diagnostic and referral processes.

In healthcare systems, providers are generally informed about the need for special therapies for a beneficiary only after submitting a formal request accompanied by a medical prescription. These requests are often made after the optimal intervention window has passed, limiting patients' access to therapies whose effectiveness depends on early implementation. Research on the early diagnosis of atypical conditions has sought to address these challenges, as seen in the case of ASD, where early intervention is critical to improving developmental outcomes. Standard ASD diagnostic processes are primarily clinical and include imaging tests, blood tests, and structured interviews, which collectively help to formalize a diagnosis. To accelerate and improve this process, recent studies have explored the use of ML models to detect ASD based on diagnostic data.

The timeliness of the diagnostic process is crucial not only to improve the quality of life of the patient but also to maximize the effectiveness of available resources. However, in the context of healthcare providers, access to diagnostic data is often restricted, making it challenging to monitor beneficiaries, develop action plans, and accelerate diagnostic and referral processes. This lack of information limits the provider's ability to recommend customized therapies suited to the individual's needs on time. It also restricts the adoption of ML-based solutions based on diagnostic data.

To implement effective ML solutions, it is essential to develop data sets that take advantage of the information already available within the provider's systems. These data sets must integrate administrative and operational data contextualized to the specific application, enabling both detailed analysis and the development of ML-based systems. The development of such solutions can contribute to the establishment of a diagnostic protocol within the healthcare provider, fostering the acceleration and better monitoring of the diagnostic formalization process. Another advantage of these tools is their ability to visualize data from beneficiaries using ST, allowing for the identification of patterns that can guide more effective strategies for care and treatment.

Once analyzed, these patterns can not only optimize the provider's internal processes but also offer valuable insights for early interventions, resource allocation, and personalized action planning. In this way, ML-based solutions have the potential to transform the way healthcare providers manage the care of beneficiaries with specific needs, driving improvements in both operational efficiency and service quality.

1.2 OBJECTIVES

The primary goal of this work is to apply the CRISP-ML methodology to create a database focused on ASD, using beneficiary utilization data rather than diagnostic data. This database served as the foundation for training ML models that can assist in monitoring and identifying beneficiaries at risk. The following specific goals are outlined to accomplish this goal.

- Understanding the problem in the company context: Hold meetings and discussions with the healthcare provider's expert team to understand the context of the problem, the company's internal processes, and the specific demands related to the diagnosis of ASD. This step is fundamental within the CRISP-ML methodology to align the proposed solutions with the organization's practical needs and objectives.
- Explore and select relevant variables: Analyze the usage data provided by the health insurance company, identifying and selecting the variables that best represent the ASD problem. This step seeks to ensure the relevance of the information used.
- Develop a database specific to the context of ASD: Consolidate a robust and representative database designed to be used in both exploratory analyses and in

the training of ML models. The database will be built from the selected variables and adjusted to the needs identified in the previous step.

- Develop approaches based on ML models: Implement and evaluate two approaches using the created database: One-Class Models, which identify beneficiaries at risk of needing Special Therapies, and Binary Classification Models, which classify beneficiaries at risk of presenting ASD.
- Compare the performance of trained models: Evaluate the performance of models with statistical analysis, comparing models with hyperparameter tuning and feature selection techniques.
- Incorporate interpretability techniques: Apply interpretability methods to trained models to understand the factors that most influence the predictions made. This step aims to facilitate discussions with the healthcare provider team and provide insight for clinical decision-making.
- Evaluate the feasibility of practical integration: Discuss with the healthcare provider team the possibility of integrating the predictive models developed into the company's internal processes, considering the results obtained and the challenges identified.

1.3 DOCUMENT STRUCTURE

This work follows a structured organization. Chapter 2 explains the essential concepts for this study, establishing the foundation for understanding the problem domain. Chapter 3 reviews and discusses relevant studies on ASD detection using ML. Chapter 4 describes the methodology and techniques used to develop the solution, structured around the CRISP-ML phases. Chapter 5 presents the experimental setups, discusses the results, and evaluates performance using the chosen metrics. Chapter 6 summarizes the contributions of this work and outlines potential directions for future research.

2 BACKGROUND

This chapter introduces the central themes of this work. Section 2.1 conceptualizes and describes Autism Spectrum Disorder, emphasizing the importance of early diagnosis and specialized therapies. Section 2.2 outlines the theoretical foundations of machine learning, highlighting its main approaches and general applications. Section 2.3 presents the CRISP-ML process, explaining its concepts and stages in the context of machine learning projects.

2.1 AUTISM SPECTRUM DISORDER

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that can be associated with other medical and psychological conditions. It is characterized by difficulties in social interaction and communication, as well as repetitive and restricted behaviors (HODGES; FEALKO; SOARES, 2020). According to the World Health Organization (WHO), approximately 70 million people worldwide are autistic [1]. ASD is a heterogeneous disorder, meaning that individuals with autism have different symptoms and can be classified into three levels of support: mild, moderate, and severe.

The disorder is characterized by two domains: social communication and repetitive patterns. The social domain relates to difficulties in socio-emotional reciprocity, non-verbal communication, understanding body language and gestures, and understanding and maintaining relationships. The domain of repetitive patterns is associated with the repetitive use of objects, such as arranging toys, difficulties with routine changes, insistence on the sameness, and hyperfocus on specific objects (MUKHERJEE, 2017).

The causes of autism are primarily genetic factors. Currently, over a thousand genes have been mapped as risk factors for the disorder. There are also syndromes and other genetic neurological disorders related to autism, such as Rett Syndrome and Fragile X Syndrome. Furthermore, there is a higher prevalence in male children compared to females, although the reason for this discrepancy is still unknown (SHARMA; GONDA; TARAZI, 2018).

In addition to genetic factors, the manifestation of autism can be related to environmental factors considered risk factors (SUKIENNIK; MARCHEZAN; SCORNAVACCA, 2022), such as advanced paternal age, the use of medications during pregnancy, developmental disorders in family members, the presence of genetic syndromes, prematurity, low birth weight, diabetes and hypertension during pregnancy.

[1] <www.who.int>

The first signs of autism are noticeable in the first months of life, becoming more evident around 18 months. Parents are often the first to notice these signs, and the first doctor consulted is often the pediatrician (ARAUJO et al., 2019). This initial suspicion is usually based on signs such as:

a) Absence of vocal sounds;

b) The baby does not imitate human actions such as smiling, yawning, or sticking out the tongue;

c) Does not show a preference for smiling, sad, or angry faces;

d) The gaze does not follow the mother's departure;

e) Has not spoken any comprehensible words by 12 months;

f) Lack of curiosity about people;

g) The child seeks to isolate themselves from others.

Parents are strongly encouraged to screen their children, as they are in the best position to detect signs of autism. In addition to detecting signs, it is essential to observe the regression of acquired developmental milestones. Many children show signs of autism around the second year of life, with signs such as loss of speech skills, reduced interest in socialization, and the appearance of repetitive behaviors (SUKIENNIK; MARCHEZAN; SCORNAVACCA, 2022).

To identify these traits, interviews with parents and child development monitoring methods are conducted. One of the tools used for the analysis in Brazil is the Child Health Booklet [2], which, since its third edition, has incorporated a screening instrument for ASD based on the Modified Checklist for Autism in Toddlers Scale (M-CHAT). The M-CHAT scale is presented in Annex A and can assist physicians in early screening and identification of autism. In addition to the M-CHAT test, there are other tests such as the Autism Diagnostic Interview Revised (ADI-R), Gilliam Autism Rating Scale 3rd Edition (GARS), Autism Diagnostic Observation Schedule (ADOS), Childhood Autism Rating Scale (CARS), and Autism Spectrum Quotient (AQ-10), which are based on clinical observations and information from parents and caregivers.

It is important to note that a positive result in the screening does not mean that the child has ASD, but rather indicates a risk factor for the possibility of having it. Early suspicion allows the child to be referred for specialized care, allowing early interventions and treatments that can aid in development, promoting independence, quality of life, and accessibility (ONZI; GOMES, 2015).

Since there is no specific cause for autism, there is also no standard test for its diagnosis. A team monitors the development of the child and makes the clinical

[2] <https://www.gov.br>

diagnosis. To help in this diagnosis, imaging tests, genetic tests, and evaluations by psychologists and occupational therapists are performed. A reliable diagnosis of autism can typically be made around two years of age. In Brazil, the average age for diagnosis is six years, a delay that represents a loss in the child's developmental potential (LORD et al., 2018).

There is no cure for autism, but the standard treatment is early intervention, which should be initiated as soon as there is suspicion of a diagnosis. Within ASD, support levels are identified, ranging from mild, with slight difficulties in adaptation and total independence, to severe levels, which require assistance with daily activities throughout life. Thus, the needs of each autistic individual vary, and the therapeutic modalities recommended for each child are related to these needs, the family dynamic, available resources, and other factors, resulting in an individualized plan created from a therapeutic evaluation involving professionals from various fields (SUKIENNIK; MARCHEZAN; SCORNAVACCA, 2022). Examples of therapeutic modalities include psychotherapy, physical therapy, occupational therapy, speech therapy, behavioral therapy, and sensory integration.

Since the initial search for an autism diagnosis is based on observation, there may be a delay in seeking medical help. Often, parents and caregivers notice signs such as communication difficulties and repetitive behaviors but can attribute these characteristics to typical developmental phases, leading to uncertainty about the need for professional evaluation. An example of this observation-dependent detection failure is that girls with autism tend to take longer to diagnose, especially those with mild autism, compared to boys.

Girls often do not fit the typical stereotypes associated with autism and tend to have more masked symptoms compared to boys. Some studies suggest that girls must have severe behavioral and social difficulties for autism to be considered the cause of these deficits. These discrepancies in diagnosis between girls and boys highlight the importance of a more sensitive and individualized approach to the identification and treatment of autism in all children (HALLADAY et al., 2015).

Currently, there are no specific medications to treat the core symptoms of autism, which are social communication and repetitive behaviors. The standard treatment for autism involves therapeutic modalities. An advantage of early intervention is that, in the early years of life, neurons have a greater capacity to form new connections, making therapies more effective when applied at these ages (LORD et al., 2020).

Early intervention has proven to be a fundamental approach in the development of children with ASD. Studies show that when initiated during the early years of life, this intervention can significantly improve children's intellectual abilities, communication, language, and adaptive behavior. These improvements remain notable even after the intensity of services is reduced, indicating that the gains achieved are sustained over the long term. In addition, early interventions have the potential to influence educational placement and the support necessary for each individual, ensuring continuous progress in various areas of development. (DAWSON et al., 2010) (ESTES et al., 2015)

Among the recommended therapies, Applied Behavior Analysis (ABA) stands out for its proven effectiveness and endorsement by the WHO. Evidence-based ABA seeks to understand and modify behaviors to maximize learning and minimize the challenges faced by people with ASD. In addition, psychopedagogical support plays a crucial role in personalizing interventions, respecting the individuality and needs of each child, and ensuring that they reach their full potential. (UNIMED SANTA CATARINA, 2021)

Another essential aspect of comprehensive care is the inclusion of physical therapies, such as physiotherapy and occupational therapy. Professionals in these areas can help develop fine and gross motor coordination, which is fundamental for everyday activities such as playing, dressing, and maintaining personal hygiene. Therapeutic strategies that actively involve parents and make adjustments to the family environment have proven decisive in optimizing the outcomes of interventions.

Given the importance of early and individualized interventions, it is crucial that the diagnosis of ASD occurs as early as possible. Early detection enables these therapies to be implemented in a critical stage of child development, significantly increasing the chances of progress and improved quality of life. Therefore, it is imperative that parents, educators, and healthcare professionals remain attentive to the signs of autism and seek specialized guidance to confirm the diagnosis and plan the most effective intervention strategies. (GURALNICK, 2017)

Early detection and diagnosis are crucial, requiring a multidisciplinary approach to diagnose and treat ASD, involving professionals from various fields, such as psychologists, speech therapists, occupational therapists, and educators. However, diagnosing a person with autism requires a long process of observations and tests, which is why many organizations emphasize the importance of preparing health and education professionals to deal with the specific characteristics of ASD and help with early identification.

Intensive research is ongoing aimed at assisting parents and healthcare professionals in the early detection of autism, focusing on combining advanced technologies with existing clinical assessments to effectively accelerate the diagnosis process. Artificial intelligence (AI) algorithms and machine learning (ML) have been utilized in research to analyze large volumes of data and identify subtle patterns, helping to develop predictive models that provide quick and accurate evaluations (HYDE et al., 2019) (WASHINGTON; WALL, 2023a).

Research also focuses on analyzing genetic tests for autism diagnosis, examining brain imaging data, and using parent-led interviews to train models. In addition, some studies explore diagnostic approaches based on facial expressions or the analysis of movements in videos. These studies have significant potential to exploit computational methods to help diagnose and detect autism (WASHINGTON; WALL, 2023b).

2.2 MACHINE LEARNING

Machine learning is a subfield of artificial intelligence dedicated to developing systems that learn and improve their performance autonomously, without the need for explicit programming. These algorithms estimate parameters or structures with the aim of finding complex patterns in the provided data. The main advantage of this approach is the ability of the system to extract knowledge from the data, allowing it to adapt and continuously improve as more data are processed, without relying on constant manual adjustments (RUSSELL; NORVIG, 2020). In machine learning algorithms, it is not necessary to provide specific instructions for the tasks to be performed. Instead, the system automatically learns from the data fed in it, adjusting its parameters to optimize task execution (ALZUBI; NAYYAR; KUMAR, 2018).

This ability to learn and adjust without direct human intervention is one of the most powerful characteristics of this approach, which allows for the solution of complex problems in various fields. Furthermore, machine learning is particularly effective in situations where it is difficult or impractical to explicitly program all the rules that govern a system. Three points outlined in Russell e Norvig (2020) demonstrate reasons to use machine learning instead of algorithms designed without it:

a) Anticipation of potential situations: A machine learning algorithm can learn various situations without being explicitly programmed for each one;

b) Adaptation to changes: A machine learning algorithm can learn and adapt to changes in the system;

c) Finding solutions: For certain problems, such as facial recognition, where software designers may not know how to create a solution, machine learning algorithms are used to find ways to solve these tasks.

Machine learning can be carried out in three main ways: supervised learning, reinforcement learning, or unsupervised learning. In unsupervised learning, the system learns to recognize patterns in the dataset without receiving explicit feedback. In reinforcement learning, the system learns through rewards or punishments based on its output for a given dataset. In supervised learning, the system receives a labeled dataset with the expected output and learns to map a function to that data. If the output is categorical, this task is called classification, and if it is numerical, it is called regression (SARKER, 2021).

Supervised learning techniques are based on labeled datasets, where each training example consists of an input and a corresponding output. During the training process, the model learns to identify patterns in the input data that are associated with the outputs. Subsequently, this model can be applied to make predictions on unseen data. Supervised machine learning is used to solve two main tasks: classification and regression (NASTESKI, 2017).

In regression tasks, the goal is to predict continuous numerical values based on a labeled dataset. In regression, the goal of the model is to estimate a quantitative value within a continuous range. Practical examples of regression include predicting real estate prices based on features such as location, size, and number of rooms, or estimating future temperatures based on historical climate data. Regression models use mathematical functions to establish a relationship between the independent variables and the dependent variable. This relationship is adjusted to minimize the error between the predicted values by the model and the actual values in the training dataset. The output of a regression model is typically a continuous number, such as a price, a quantity, or a physical measurement (BISHOP; NASRABADI, 2006).

In contrast, classification tasks aim to predict a class or category based on a labeled dataset. Common examples include identifying emails as spam or not and determining medical diagnoses based on clinical data. Classification can be subdivided into various categories, such as one-class classification, where the model is trained exclusively on examples from a single class, learning the normal pattern of the data and identifying deviations as anomalies or outliers. Binary classification limits the model's output to two possible categories, while multiclass classification allows the model to predict more than two categories. The output of the classification models is often represented as a probability associated with each class. This enables the model not only to provide the most likely prediction, but also to indicate the confidence level of that prediction.

To evaluate these models, metrics are employed to assess the model's ability to make accurate predictions and generalize to new data. Regression models use evaluation metrics aimed at calculating model errors by comparing predicted values with actual values. For example, mean absolute error (MAE) measures the average of absolute errors, Mean Squared Error (MSE) calculates the average squared error, and R^2 computes the proportion of data variability that the model can explain.

Classification models use metrics such as accuracy, recall, and precision to assess their performance. These metrics are calculated by comparing the model's predictions with the actual classifications, often summarized in a confusion matrix. As shown in Table 1, the confusion matrix compares the predicted results of the model with the actual values, providing a clear visualization of its performance and allowing the calculation of key evaluation metrics.

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

Table 1 – Confusion Matrix for Binary Classification

Accuracy, as defined in Equation 2.1, represents the proportion of correct predictions out of the total number of predictions. Recall, given in Equation 2.2, assesses the model's ability to correctly identify positive instances, while precision, shown in Equation 2.3, measures the proportion of correctly classified positive instances among all positive predictions. Beyond accuracy, recall and precision are particularly crucial in scenarios where different types of errors carry varying consequences (HOSSIN; SULAIMAN, 2015).

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$\frac{TP}{TP + FN} \tag{2.2}$$

$$\frac{TP}{TP + FP} \tag{2.3}$$

Popular models for binary or multiclass classification include support vector machines, decision trees, and artificial neural networks. However, commonly used regression algorithms include linear regression, polynomial regression, and tree-based models, such as random forests.

2.2.1 Randon Forest

Random Forest is a machine learning model that combines multiple classifiers to optimize the decision-making process. The model relies on decision trees, where each tree is built using a random subset of data, ensuring a low correlation between the models. Decision trees are structured as a tree-like model, where each node represents a decision on how to split the data based on its features. The distribution tree model is constructed sequentially regarding the features, aiming to group instances of the same class within the same node. This recursive process continues until the nodes reach a predefined minimum size or all possible splits are completed. Technically, Random Forest models consist of a collection of decision trees H(X,K), where K represents independent trees, and each tree casts a single vote for the most popular class for input X (BREIMAN, 2001).

The Random Forest model incorporates fundamental techniques to enhance accuracy and generalization, with the Bagging method being one of its primary features. In this method, each tree is trained using a subset of data obtained by sampling with replacement from the original data set. This approach reduces the risk of overfitting by promoting diversity among individual trees. Additionally, Random Forest employs the Out-of-Bag (OOB) technique to evaluate model performance. This technique leverages data not selected during the sampling process to validate each individual tree. Thus, the OOB error provides an accurate estimate of the model's performance without requiring a separate test dataset.

One of the significant advantages of Random Forest lies in its training methodology. The individual strength of each tree and their low correlation are two key factors that determine the effectiveness of the ensemble (BIAU; SCORNET, 2016). Figure 1 provides a high-level visualization of the Random Forest model. In the figure, each tree operates on distinct subsets of features, producing individual outputs that are combined to generate the model's final prediction.

During implementation, several parameters can be tuned to adapt the model to different datasets and tasks. The key parameters include (BIAU; SCORNET, 2016) (PEDREGOSA et al., 2011):

- n_estimators: Specifies the number of trees in the forest. A higher value tends to improve the accuracy of the model, but increases training time;
- max_depth: Controls the depth of each tree. Deeper trees capture more complex patterns, but may lead to overfitting, especially with noisy data;
- min_samples_split: Determines the minimum number of samples required to split a node. Larger values create less complex trees, reducing the risk of overfitting, while smaller values allow capturing more detailed splits, which is useful for large or complex datasets;
- max_features: Controls the number of features considered at each split. Limiting this number increases diversity among the trees, as each tree will make decisions based on different subsets. This strategy also reduces training time;
- max_leaf_nodes: Sets a limit on the number of leaf nodes in each tree, simplifying the model and improving generalization, especially for small datasets;

Figure 1 – Random Forest Model

Figuras/rf_model.png

- min_weight_fraction_leaf: Specifies a minimum fraction of the total sample weight required for a node to be considered a leaf. This parameter is helpful in handling data imbalance, ensuring that leaf nodes represent significant proportions; and
- bootstrap: Indicates whether the Bagging method will be used. When enabled, each tree is trained on a different subset of data, promoting greater diversity and reducing the risk of overfitting;

2.2.2 eXtreme Gradient Boosting

The eXtreme Gradient Boosting (XGB) algorithm, also based on decision trees, utilizes an optimization method known as gradient boosting of decision trees (CHEN; GUESTRIN, 2016).

Boosting is a sequential method designed to improve the accuracy of the model. It begins by training simple decision trees, where each subsequent tree is adjusted to correct the errors made by the previous ones. While bagging trains trees in parallel, boosting trains them sequentially. The goal of boosting is for each new tree to focus on the most challenging aspects of the dataset. At the end of the process, the trees are combined through a weighting mechanism, giving more influence to the trees that performed better. Gradient boosting incorporates the use of gradient descent to minimize model error. For each example in the training set, the loss function gradients are calculated with respect to the prediction of the current model. This involves computing both the first-order gradient and the second-order gradient, which are used to determine the direction and magnitude of the necessary adjustments (FRIEDMAN, 2001).

The first order gradient, which represents the derivative of the loss function L with respect to the prediction $F(x)$, is given in equation 2.4. Similarly, the second-order gradient, which corresponds to the second derivative (Hessian) of the loss function, is expressed as demosntraded in question 2.5. These gradients guide the optimization process, allowing the model to update the predictions in a direction that minimizes the overall loss function.

$$g_i = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \tag{2.4}$$

$$h_i = \frac{\partial^2 L(y_i, F(x_i))}{\partial F(x_i)^2} \tag{2.5}$$

XGB optimizes these methods by implementing parallel processing in tree construction and includes lasso and ridge regularization to prevent overfitting. Figure 2 illustrates the architecture of an XGBoost model, where trees are created based on the residuals generated by their predecessors. When building XGB models, several parameters are defined:

- learning_rate: Controls the contribution of each new tree to the final model. A lower learning rate means the model will add new trees with less influence, which can improve accuracy but requires a larger number of iterations to converge;
- n_estimators: Refers to the total number of decision trees built sequentially in the boosting model. A larger number of trees can improve the model performance, but can lead to overfitting and increased training time;
- max_depth: Defines the maximum depth of each decision tree. Deeper trees can capture more complex interactions between variables, but are more prone to overfitting; and

Figure 2 – XGB architecture model

Figuras/xgb_model.png

- subsample: Represents the fraction of the training data sampled for each tree. This hyperparameter helps prevent overfitting, as each tree is trained on a subset of the data, introducing a form of regularization.

XGB was developed as an open-source package compatible with widely used languages such as Python and R. In addition to integrating with popular tools in data science, the model stands out for its high processing capacity and innovative functionalities, such as sparse pattern recognition for handling sparse data and the use of a weighted quantile sketch based on theoretical principles for approximate learning. These features make XGB an efficient model for solving large-scale problems while optimizing computational resource utilization.

2.2.3 Categorical Boosting

The Categorical Boosting (CatBoost) algorithm employs the gradient boosting method applied to decision trees, standing out for its efficient handling of categorical data. This advantage arises from its design to process categorical features without requiring extensive preprocessing. The strategy involves using statistics calculated from label values, leveraging a random permutation of the data to compute the mean label value for preceding categories in the permutation. This approach helps reduce overfitting (PROKHORENKOVA et al., 2018).

CatBoost introduces additional techniques, such as the concept of ordered boosting, which uses multiple random permutations of the training data to calculate less biased gradient statistics. This approach aims to improve the model's accuracy by reducing the variation in gradient updates, which can occur due to the specific order of the data during training. Furthermore, CatBoost adopts a unique decision tree model called oblivious trees, as shown in Figure 3. Unlike the decision trees used in XGBoost, oblivious trees apply the same splitting criterion across all nodes at a given tree level. This results in more balanced trees, which are less prone to overfitting and, at the same time, enable faster model evaluation execution. Oblivious trees are also particularly advantageous in terms of simplicity and computational efficiency, as the uniform application of splits at each level allows for greater parallelization and reduced calculation complexity. This characteristic makes CatBoost not only robust against overfitting but also efficient in terms of training time. (DOROGUSH; ERSHOV; GULIN, 2018).

CatBoost positions itself as an advanced solution for machine learning tasks involving categorical features, with developers emphasizing significant acceleration using GPUs. Like other models, CatBoost features parameters that can be adjusted to improve performance for specific tasks. Due to its similarity with XGB, some parameters overlap:

- learning_rate: Controls the step size in weight adjustments, impacting training speed and precision;
- n_estimators: Defines the number of trees in the model, influencing capacity and computational cost;
- max_depth: Sets the maximum tree depth, where greater depth increases model complexity;
- l2_leaf_reg: Applies L2 regularization to leaves to control overfitting;
- subsample: Specifies the proportion of samples used for training each tree, introducing randomness and improving generalization; and

Figure 3 – Difference between oblivious trees and asymmetric trees

Figuras/cat_model.png

- colsample_bylevel: Determines the proportion of columns used at each tree level to reduce dimensionality and prevent overfitting.

2.2.4 Isolation Forest

Some classification scenarios require algorithms designed to identify anomalies or patterns of interest within a single data class. Unlike traditional classifiers, these methods are tailored to model a single majority class and recognize examples deviating from that pattern, ensuring robust performance in scenarios with imbalanced distributions or the absence of minority-class examples. Common algorithms include a one-class SVM, which learns to delineate a boundary around target-class data, isolation Forest, which detects anomalies by exploring isolation characteristics, and autoencoders, which use neural networks to reconstruct majority-class data and flag significant deviations.

Although one-class classification and anomaly detection share similarities, they differ in focus. The one-class classification focuses on modeling the distribution of a single class to identify outliers. In contrast, anomaly detection identifies rare or significantly different patterns or observations, encompassing both specific-class data and adversarial-class data. Classification of a class exclusively distinguishes the target class from unknown data, while anomaly detection applies to various contexts and data types, with the aim of identifying any out-of-pattern behavior regardless of predefined classes (KHAN; MADDEN, 2014).

The Isolation Forest (IF) algorithm builds decision trees to isolate anomalies. Unlike traditional methods that create normal instance profiles to identify deviations as anomalies, IF explicitly focuses on isolating anomalies. The premise of IF is that anomalies are "few and different," making them more likely to be isolated near the tree root, while normal instances are isolated at deeper levels (LIU; TING; ZHOU, 2008).

In the architecture illustrated in Figure 4, anomalous examples are isolated at shallow levels due to their distinct characteristics compared to normal data. Meanwhile, normal instances tend to be isolated only at deeper levels, as they share similar features with most of the dataset's samples. This behavior reflects the idea that anomalies are easier to isolate due to their uniqueness, while normal instances, being more common, require greater depth to be isolated.

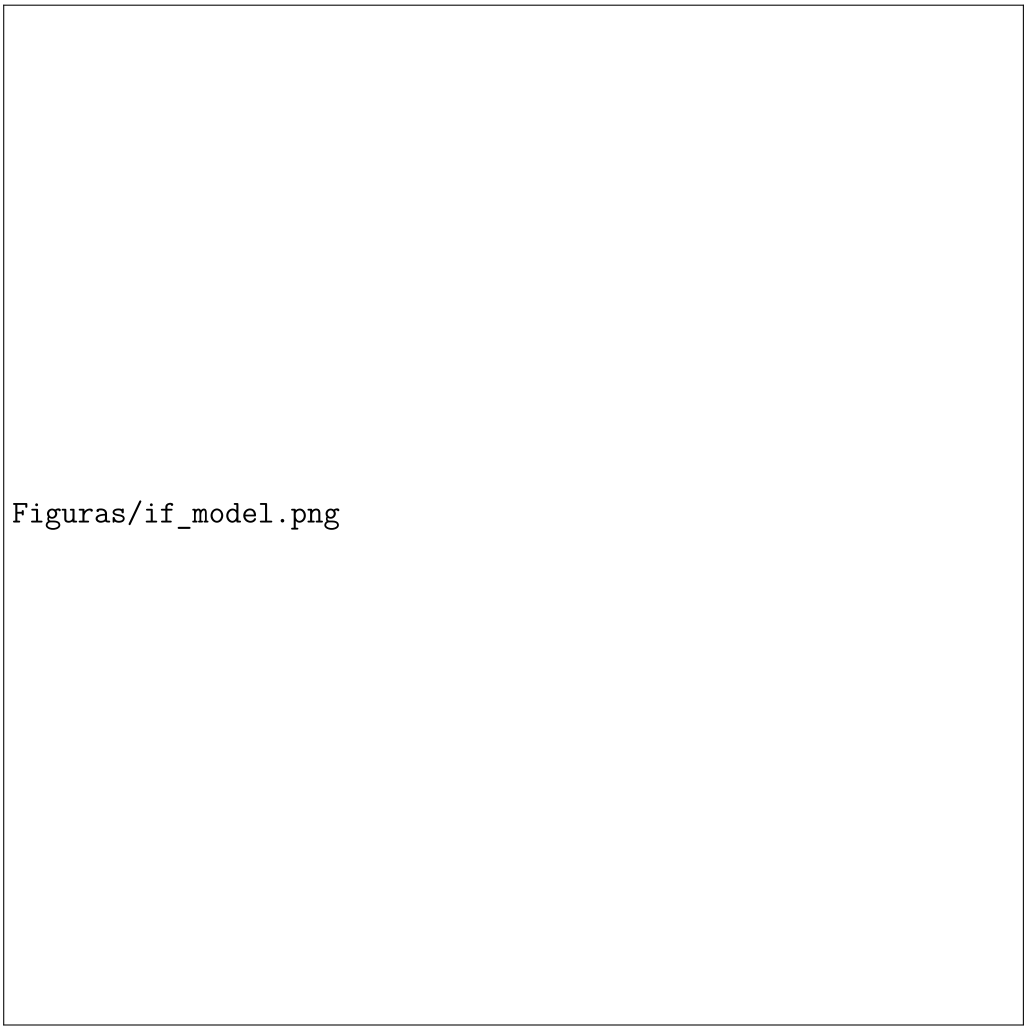
IF adopts an ensemble approach, repeatedly isolating elements with various tree configurations and calculating an anomaly score based on the average depth required to isolate each data point. The smaller the average depth, the higher the probability that a data point is an anomaly, as it was isolated more quickly compared to normal data. This strategy allows IF to be particularly effective at detecting anomalies in large, high-dimensional datasets, where traditional detection methods may be less efficient. Moreover, using multiple trees in an ensemble improves the robustness of the algorithm, ensuring that the model has a broader view of the data patterns.

The IF training process involves creating a tree ensemble using dataset subsamples. For each iteration, a subsample is selected and a tree is built using random partitions. At each node, an attribute and split point are randomly chosen between the attribute's minimum and maximum values. A key advantage of IF is its ability to perform well even when the training set lacks anomalies, which is common in many practical scenarios (LIU; TING; ZHOU, 2012).

Key parameters include:

- n_estimators: Specifies the number of trees in the forest, with more trees improving anomaly detection precision but increasing computational cost.
- max_samples: Determines the maximum number of samples per tree, set as an

Figure 4 – Isolation Forest Architecture Model



integer or fraction of the total data.

- contamination: Indicates the expected anomaly proportion in the dataset, used to set the threshold separating normal and anomalous instances.
- max_features: Sets the maximum features per node during tree construction, reducing model complexity but potentially impacting precision.
- bootstrap: Indicates whether sampling is performed with replacement. If set to "True," samples can repeat; otherwise, trees use unique samples.

2.2.5 One-Class SVM

The One-Class SVM (OC-SVM) algorithm is a variation of the Support Vector Machine (SVM) algorithm. Its goal is to construct a hyperplane that separates similar samples from those that are dissimilar, acting as a decision boundary f that returns +1 for data samples on one side of the hyperplane and -1 otherwise (SCHÖLKOPF et al., 2001).

In its operation, the objective is to create a boundary around normal instances by encapsulating them in a region of familiarity. This boundary is strategically placed to maximize the margin around normal data points, enabling a clear distinction between what is considered common and what might be considered uncommon. The mathematical formulation of OC-SVM involves minimizing an objective function that seeks to find the smallest possible boundary around normal samples while penalizing samples that fall outside this boundary.

The OC-SVM uses a dual solution that allows for the incorporation of kernel functions, such as the Gaussian (RBF) kernel, to handle non-linear data distributions. This approach adds flexibility to the model, allowing it to capture complex patterns and efficiently model the training data distribution. The use of kernels is a key feature of SVMs, as they map the input data to a higher-dimensional space, where the separation between classes, or in this case, the separation between normal and anomalous instances, becomes more manageable. The RBF kernel function is particularly useful for anomaly detection problems where the relationship between instances does not follow a simple linear distribution, allowing the OC-SVM to better adapt to the complexities of the data (CHEN; ZHOU; HUANG, 2001).

The Gaussian kernel function is given by equation 2.6, where $K(x, x')$ is the kernel value between instances x and x', $\|x - x'\|^2$ is the squared Euclidean distance between data points x and x', and σ is the kernel bandwidth parameter, which controls the "thickness" of the kernel distribution.

$$K(x, x') = \exp\left(-\frac{|x - x'|^2}{2\sigma^2}\right) \tag{2.6}$$

Moreover, the kernel parameter have a significant impact on the model's generalization ability. The improper choice of these parameters can lead to a model that does not adequately capture anomalies, either by overfitting the training data or by not being flexible enough to identify more subtle anomaly patterns.

Figure 5 illustrates an example of novelty detection using OC-SVM with a Gaussian kernel. The red line of the learned decision boundary encapsulates the white circles of the training observations in high-density regions. New regular observations, such as green circles, are correctly classified within the known regions, while anomalous observations yellow circles fall outside the familiarity regions. This separation demonstrates the kernel's effectiveness in modeling complex distributions and identifying nonconforming instances. In addition to the kernel, several hyperparameters can be configured to improve the model's performance:

Figure 5 – SVM One Class Architecture Model

```
Figuras/svm_model.png
```

- kernel: Defines the type of kernel function to be used for transforming the data. Kernels enable modeling non-linear patterns;

- degree: Specifies the degree of the polynomial kernel (`poly`). It is ignored if another kernel is used. A higher value allows us to model more complex relationships;

- gamma: Defines the coefficient for RBF, polynomial, and sigmoid kernels. Controls the impact of each sample on the data transformation. Small values create smoother decision boundaries, while larger values make the model more sensitive to the data;

- coef0: An independent term used in polynomial and sigmoid kernels. Adjusts the impact of non-linear interaction in the transformed data;

- tol: Determines the tolerance for the convergence of the optimization algorithm. Lower values increase the solution's precision but may increase execution time;

- nu: A parameter that controls the expected fraction of outliers in the training data and the upper fraction of support vectors. It is a value between 0 and 1. Larger values allow for more outliers and influence the flexibility of the model;
- shrinking: A Boolean value that enables or disables the use of heuristics to speed up convergence. When True, it reduces the number of calculations needed during optimization; and
- cache_size: Specifies the size of the cache for storing kernel values, in megabytes. A higher value can speed up training, especially for large datasets.

One of the main challenges in classification models is avoiding the model's tendency to overfit the training data, known as overfitting, which compromises its ability to generalize to new data. This occurs when the model learns irrelevant patterns or noise present in the training set, leading to poor performance on unseen data. To address this issue, techniques such as hyperparameter tuning and feature selection play a crucial role in improving model performance.

Hyperparameter tuning involves optimizing the model's hyperparameters, such as the learning rate, decision tree depth, or regularization, to find the ideal combination that minimizes error and prevents overfitting. Additionally, feature selection is essential for reducing the model's complexity by retaining only the most relevant features for the classification task. Removing irrelevant or redundant features not only improves the model's generalization but also reduces computational costs and simplifies interpretation.

When effectively applied, these approaches are vital for enhancing a classification model's ability to generalize to new data, resulting in more robust and efficient models.

Before exploring technical approaches to optimize models, selecting the appropriate algorithm is crucial for the task's success. The choice of model depends on various factors, such as the dataset size, the type of variables involved, and performance requirements. Thus, understanding the problem to which the models will be applied is of great importance, as it guides the selection of the most effective methodology and ensures better results in the specific application context.

2.3 CRISP-ML

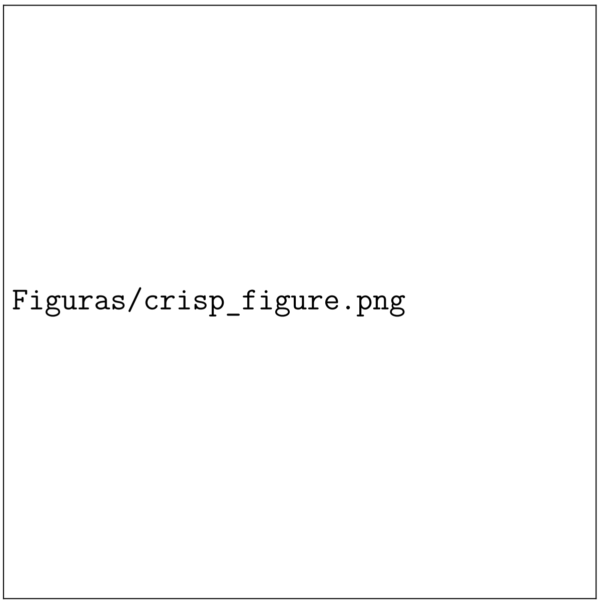
In 1996, interest in Data Mining grew significantly. However, there was a lack of a process to assist the industry in applying and standardizing this practice. Four leaders in the emerging data mining market created CRISP-DM at the end of 1996. By

2000, the first version was developed with the aim of establishing a standard process model to serve the data mining community. CRISP-DM acts as a guide to help anyone conduct a data mining project. The authors emphasized that the model is not a manual, but rather a source of inspiration for best practices and a way to enable organizations to implement data mining more effectively. The process comprises six cyclical phases, which illustrate the natural flow of data mining (SCHRÖER; KRUSE; GÓMEZ, 2021).

Figure 6 presents all phases of CRISP-DM, which are divided into business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The time distribution across these phases is approximately 50% to 70% for data preparation, 20% to 30% for data understanding, and 10% to 20% for the phases of modeling, evaluation, and business understanding. The deployment planning phase occupies about 5% to 10% of the total time.

The first phase is business understanding, focusing on defining the project's objectives and requirements. The second phase, data understanding, involves the initial collection and evaluation of data. Data preparation refers to the preprocessing needed to prepare the data for modeling. In the modeling phase, models are selected and applied. Evaluation includes analyzing and validating the model's predictions and results. Finally, the deployment phase involves planning, implementation, and project monitoring.

Figure 6 – CRISP-DM Phases



Source: Chapman (2000)

As shown in Figure 7, the methodology is a hierarchical process model. Each task has four levels of abstraction: phase, generic task, specialized task, and process instance. At the highest level are the six process phases; the second level includes the generic tasks corresponding to each phase. The third level consists of specialized tasks that detail the actions needed to complete each task. Finally, the fourth level, the process instance, records the actions, decisions, and results (WIRTH; HIPP, 2000).

Figure 7 – Hierarchical Process Model of CRISP-DM

Figuras/hierarchical_process.png

Source: Chapman (2000)

Currently, CRISP-DM is considered the standard methodology, but its usage has declined because it does not fully meet the community needs. This gap has led to the development of new methodologies based on CRISP-DM aimed at addressing these shortcomings. CRISP-ML, an extension of CRISP-DM, is one such methodology. Its primary innovation is the addition of steps to apply interpretability to models (KOLYSHKINA; SIMOFF, 2021).

Model explainability is crucial in various contexts, such as validating results, increasing user trust, and continuously improving the modeling process (STIGLIC et al., 2020). In critical fields such as healthcare, finance, and security, explainability is essential to justify automated decisions and ensure models are transparent and ethical. Furthermore, interpretable models help identify issues such as biases or errors in data, enabling adjustments before implementation. The application of interpretability to models has become an increasingly relevant topic. The CRISP-ML methodology supports planning from the project's early stages, allowing for a better understanding of the need and level of interpretability each model must meet based on specific context requirements.

Figure illustrates the organization of the CRISP-ML model with respect to its phases. The model adopts a hierarchical structure, where each phase contains specific goals and follows a cyclical system. In step (1), all major stakeholders collaborate to understand the objectives of the project. Steps (2) to (4) focus on data: making it comprehensible, cleaning, and enriching it. Step (5) involves selecting, building, and evaluating ML models. Step (6) focuses on interpreting and validating the insights generated by the proposed solution. Finally, Step (7) is dedicated to deploying the model and preparing detailed technical reports. This cyclical approach ensures project flexibility, allowing phases to be revisited and adjusted as needs evolve (KOLYSHKINA; SIMOFF, 2021).

Another innovation of CRISP-ML is the Interpretability Matrix (IM), which structures the project stages and defines the responsibilities of each stakeholder. Figure 9 shows an example of a high-level IM, where the project phases are aligned with key stakeholders, such as executive teams, data engineering teams, IT teams and modeling teams. The matrix highlights the differentiated participation of each group throughout the project stages. For instance, in the "Project Initiation and Planning" phase, the executive team focuses on establishing requirements and domain understanding, while the data engineering team contributes to data quality checks in subsequent phases. This hierarchical and collaborative organization makes the process more efficient and well-directed, promoting clarity about each stakeholder's responsibilities throughout the project lifecycle.

Figure 8 – CRISP-ML Flowchart

./Figuras/crisp_ml_flow.png

Source: Kolyshkina e Simoff (2021)

The CRISP-ML methodology, as shown in Figure 8, consists of seven distinct phases: Project Initiation and Planning; Data Audit, Exploration, and Cleaning; Evaluation of the Predictive Potential of Data; Data Enrichment; Model Building and Evaluation; Business Insights Extraction; and Deployment and Reporting.

The first phase, Project Initiation and Planning, involves understanding business problems and collecting related data. This phase is considered the most impor-

Figure 9 – Example of an Interpretability Matrix

Figuras/interpretability_matrix.jpeg

Source: Kolyshkina e Simoff (2021)

tant and crucial for the overall performance of the project. During this stage, developers and stakeholders meet to plan the execution of the project and analyze the available data. General objectives and intermediate goals are defined, resulting in a detailed implementation plan and flowchart.

In the second phase, Data Audit, Exploration, and Cleaning, data is analyzed to verify its representativeness for the problem. Its characteristics are reviewed with specialists to ensure that the team fully understands the data. During this phase, exploratory analysis, cleaning, and adjustments are performed to make the data consistent and usable. The Data Predictive Potential Evaluation phase focuses on determining whether the data have sufficient quality and relevant variables for ML models. In this phase, issues such as missing, inconsistent, or redundant data are addressed.

The next phase, Data Enrichment, aims to improve data quality using techniques such as variable selection, feature transformation, and specific pre processing for ML models. This step is critical to ensure that the data are adequately prepared for model training. In the fifth phase, Model Building and Evaluation, predictive models are developed, trained, and evaluated using metrics related to the problem, such as accuracy, precision, recall, and R2. Advanced techniques like hyperparameter optimization and cross-validation are applied to ensure the effectiveness and robustness of the model.

Subsequently, in the Business Insights Extraction phase, model results are interpreted to generate actionable insights that can guide strategies and operational decisions. This step bridges technical results with business problem solving, transforming data into practical value. The final phase, Deployment and Reporting, focuses on integrating models into the production environment and documenting the entire process. Detailed reports on model performance and insights are generated, serving as references for stakeholders and supporting strategic decision making.

By organizing ML project development into structured and interdependent phases, CRISP-ML provides a systematic approach to solving complex problems. Each phase contributes to the project's success, from defining objectives to delivering models and final reports. This methodology, both flexible and robust, ensures that results are consistent, explainable, and aligned with business needs, fostering a more efficient integration of data, technology, and organizational strategy.

2.4 CONSIDERATION OF THE CHAPTER

In this chapter, the main theoretical foundations related to the approach developed in this dissertation are presented. Initially, ASD was addressed, highlighting its definition, key signs, and the therapies used, providing the necessary context to understand the problem. The concepts of ML were then discussed, with an explanation of the models to be used in this work, emphasizing their characteristics and applications in the context of supervised models for classification. Finally, the CRISP-ML model was introduced, with a detailed explanation of its stages, which guide the structuring and implementation of the solution developed in this dissertation. This approach enables us to address the challenges of the topic while ensuring a practical and well founded application.

3 RELATED WORK

In this chapter, a systematic literature review will be presented on studies that employ Machine Learning (ML) models to identify Autism Spectrum Disorder (ASD). The primary objective of this review is to investigate existing approaches, identify methodological trends, highlight the most commonly used models, and evaluate the results achieved in these studies. Additionally, the review aims to explore the challenges faced in this field and identify gaps in the literature.

3.1 RESEARCH METHOD

The use of machine learning (ML) for the diagnosis of Autism Spectrum Disorder (ASD) has proven to be a promising approach to improve the diagnostic process, assisting doctors and professionals in early detection. However, it is evident that most existing studies are primarily based on clinical data and specific tests. Some research, on the other hand, explores the potential of biomarkers, such as brain imaging and genetic markers, for the early identification of autism signs.

This reliance on clinical data limits the application of models in scenarios where diagnostic information is unavailable, such as in the case of healthcare operators. Therefore, it is crucial to understand how to structure datasets that can be applied to ML models, enabling their use in this context. In addition, there is a notable gap in the literature regarding the identification of the most relevant variables and the most effective models for detecting risk factors associated with autism. This lack of information highlights the need for more in-depth investigations and a systematic approach to identify and validate ASD predictors through machine learning techniques.

In order to identify the variables used in models that are not based on diagnostic data and the ML models applied, a systematic literature review (SLR) was conducted, focusing on articles published from January 2020 to January 2025. The methodology followed was PRISMA (MOHER et al., 2010).

The overall objective of this systematic literature review is to investigate studies that utilize Machine Learning algorithms for the detection of Autism Spectrum Disorder (ASD) without relying directly on clinical diagnostic data. Through a comparative analysis, the aim is to understand existing approaches and extract relevant insights that can serve as a foundation for the development of this project. To achieve this objective, five research questions were formulated to guide the review and analysis of the selected studies. These questions are as follows.

- Which datasets are used for ASD detection with machine learning and what variables are present in these datasets?;
- Which Machine Learning models are being applied to structured datasets?;
- What techniques are being used for pre-processing these datasets?;
- What interpretability strategies have been adopted in the models?; and
- What limitations have been identified in the analyzed datasets?.

These questions allow for a detailed analysis of the methodological and technical aspects utilized in the existing literature, providing insights into the identification of gaps and challenges that still need to be addressed.

3.2 PLANNING THE REVIEW

After defining the research objective and establishing the investigation questions, the next step was to determine the search phrase to be used in the selection of articles. This phrase was applied to the title and abstract fields of scientific databases. Through several iterations, a final formulation was reached that reflects the most relevant terms for the investigation.

During this process, it was observed that no studies used the term "risk factors," leading to the selection of the keywords "diagnosis" and "identification," as they are more widely employed. Additionally, to avoid including articles relying on data sets derived from medical exams, videos, audio recordings, or genetic data which fall outside the scope of this research, an exclusion clause was incorporated into the search phrase.

The final search phrase used was:

a) ((AUTISM DETECTION OR AUTISM RISK FACTOR) AND (MACHINE LEARNING OR DEEP LEARNING)) NOT (audio OR speech OR video OR genetic OR exam)

This formulation allows the search to focus on articles aligned with the proposed objective while excluding studies that do not correspond to the methodological approach of this research. Four platforms recognized for their scientific relevance were selected to search: SpringerLink, IEEE Xplore, JAMA, and ScienceDirect, using the specified search string.

3.3 THE REVIEW PROCESS

The research was conducted between December 2024 and January 2025. Several filters were applied during the search process, including a time range from January 2020 to January 2025, a restricted document type to articles, and a language limited to English or Portuguese. The search phrase was applied on the four selected scientific publication platforms.

In addition to the initial filters, specific inclusion and exclusion criteria were established to ensure that the articles identified were aligned with the research objectives. These criteria were used to select relevant articles and eliminate those that did not contribute meaningfully to the study.

The following inclusion criteria were adopted:

Articles that applied ML methods to detect ASD in structured data without explicit diagnostic information. Articles with objectives included comparing different ML models to detect ASD in structured data without explicit diagnostic information. The focus on articles using structured data was intended to identify variables not related to medical tests or diagnoses used in these datasets, as well as to understand which ML models are applied to such data structures.

The exclusion criteria were applied in three stages:

Inaccessible Articles: Documents found in inaccessible databases were discarded. Articles using image, video, or audio datasets: These were excluded as they were outside the scope of the study. Articles using structured data derived from medical tests: These were also eliminated.

These criteria ensured that the analyzed data closely resembled the types of data available in hospitals and healthcare providers without access to diagnostic information to create ASD datasets. By excluding articles outside the scope, the research focused more intensely on studies relevant to the topic, improving the understanding of datasets and ML models.

The selection process was structured to ensure objectivity and consistency, following the following criteria:

a) Objective Criteria:
   - Time range: 01/2020 to 01/2025;
   - Document type: Articles; and
   - Language: English or Portuguese.

b) Inclusion Criteria:
   - Articles that applied ML for ASD detection; and

  - Articles that compared ML algorithms for ASD detection.

c) Exclusion Criteria:

  - Articles using datasets containing images, videos, or audio; and
  - Articles using tabular datasets derived from medical tests.

By applying these criteria, the selected set of articles was refined, excluding those that were irrelevant or misaligned with the research objectives. This method ensured that the analyzed data were consistent and relevant to the scope of the study.

3.4 SUMMARY AND DISCUSSIONS

Using the search string in the selected databases defined in Section 3.2, 152 articles were found. After applying the inclusion criteria, 23 articles were selected. Subsequently, with the application of the exclusion criteria described in Section 3.3, 15 articles remained for the final review. Table 2 provides a summary of the articles found on each platform, those included based on the inclusion criteria, those excluded after applying the exclusion criteria, and finally, the articles retained for the literature review.

Table 2 – Summary of articles found in SRL, included, and selected after applying the criteria.

Database	**Articles Found**	**Included**	**Excluded**	**Selected**
SpringerLink	25	1	0	1
IEEE Xplore	50	15	4	11
JAMA	27	4	3	1
ScienceDirect	50	3	1	2
Total	**152**	**23**	**8**	**15**

A table was created to document the articles selected for review and facilitate the creation of graphs, presented in Table 3, which consolidates the information extracted from each analyzed article. This table simplifies and organizes the comparative analysis process. The information was collected through a complete reading of the articles and the table was structured according to the questions defined in Section 3.1. Below is a description of the columns and their respective contents.

- Article: Identification of the article and its year of publication;
- Objective: General objective of each article;
- Database: Description of the database used in the studies;
- Variables: Number of variables analyzed in the study;
- Limitations: Restrictions or challenges associated with the databases used;

- Pre-processing: Strategies adopted to prepare the data before applying the models;
- ML: Machine learning algorithms used; and
- Interpretability: Methods used to explain the results of the models.

Table 3 – Table model for comparative analysis of the reviewed articles

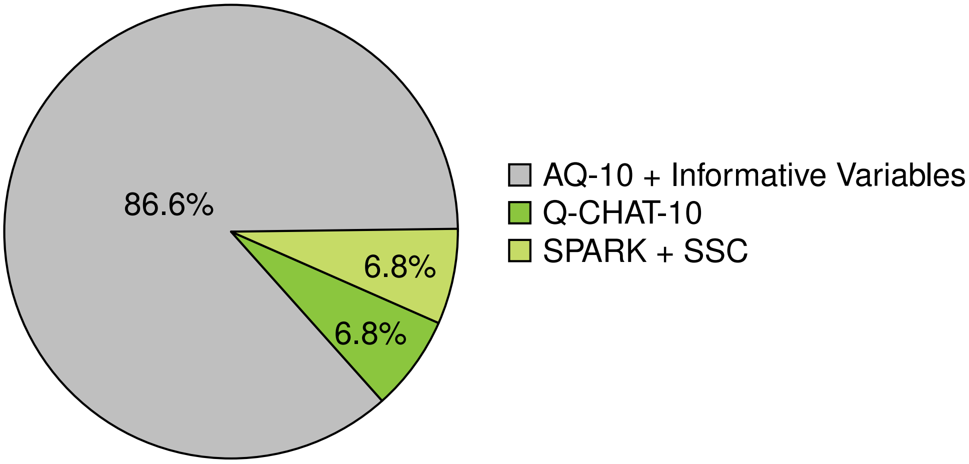
Article	Objective	Database Used	Variables	Database Limitations	Pre processing	ML Models	Interpretability
Raj e Masood (2020)	Compare ML models	AQ-10 + Informative Variables	21	Database size	Removal of null values, Outlier treatment, one-hot-encoding	SVM, NB, CNN, LR, KNN, ANN	-
Romero-García et al. (2021)	ML model Creation	Q-CHAT-10	6	Database size	Removal of irrelevant data, one-hot-encoding	DT, RF, XGB	Feature Importance
Vakadkar, Purkayastha e Krishnan (2021)	Compare ML models	AQ-10 + Informative Variables	18	Database Size, Availability of public datasets, imbalance, lack of diversity	Removal of null values, Outlier treatment, Feature Selection, one-hot-encoding	LR, NB, SVM, KNN, RF	-
Saranya e Anandan (2021)	ML model Creation	AQ-10 + Informative Variables	10	outdated	-	SVM, ANN, KNN, WOEM	-
Hasan et al. (2022)	Compare ML models	AQ-10 + Informative Variables	21	Database size, Data imbalance	one-hot-encoding, Removal of null values	AB, RF, DT, KMM, GNB, LR, SVM, LDA	Info Gain Attribute Evaluator(IGAE), Gain Ratio Attribute Evaluator(GRAE), Relief F Attribute Evaluator(RFAE), Correlation Attribute EvaluatoR(CAE)
Manoj e Praveen (2023)	Compare ML models	AQ-10 + Informative Variables	21	Database size, Data imbalance	one-hot-encoding, Removal of null values, Removal of irrelevant data	LDA, NB, SVM, LR, KNN, XGB, DT, RF, MLP	Feature importance
Kavitha e Siva (2023)	Compare ML models	AQ-10 + Informative Variables	21	Database size	Feature selection, Removal of irrelevant data, Outlier treatment	SVM,NB,LR,PSO-CNN	-
BalaKrishna et al. (2023)	Compare ML models	AQ-10 + Informative Variables	15	-	Removal of null values, Data Normalization	SVM, LR, DT, LDA	-
Naik et al. (2023)	Compare ML models	AQ-10 + Informative Variables	-	Database size, Only data for diagnosis	Removal of null values	KNN, LR, DT, RF, NB, XGB	-
Bose e Seth (2023)	Compare ML models	NDAR e AGRE	50	data redundancies	Removal of null values, Removal of irrelevant data	LR, XGB, SVM, NB	-
BalaKrishna et al. (2023)	Compare ML models	AQ-10 + Informative Variables	21	-	Removal of null values, Removal of irrelevant data, Feature Selection	SVM, NB, CNN, ANN, KNN, LR	-
Mittal et al. (2024)	Analysis of ML models	AQ-10 + Informative Variables	21	-	Feature Selection, Data Normalization	AdaBoost	-
Gill et al. (2024)	Analysis of ML models	AQ-10 + Informative Variables	21	-	Removal of null values, Feature Selection	NB	-
Rajagopalan et al. (2024)	Compare ML models	SPARK + SSC	28	lack of diversity, Add exam data	Feature Selcetion, Removal of null values, Data balancing	LR, DT, RF, XGB	SHAP
Dholakia et al. (2024)	Analysis of ML models	AQ-10 + Informative Variables	21	lack of diversity	Removal of null values, Feature Selection	ANN	-

This structure provides a systematic and detailed overview of the articles analyzed, facilitating the identification of gaps, trends, and relevant insights for research.

The analysis of the table shows that out of the 15 reviewed articles, 12 use the AQ-10 database. This distribution is illustrated in Figure 10, where 86.6% of the studies use this database with the addition of informative variables. Subsequently, the Q-CHAT-10 and SPARK + SSC databases appear, each representing 6.8%, used only once.

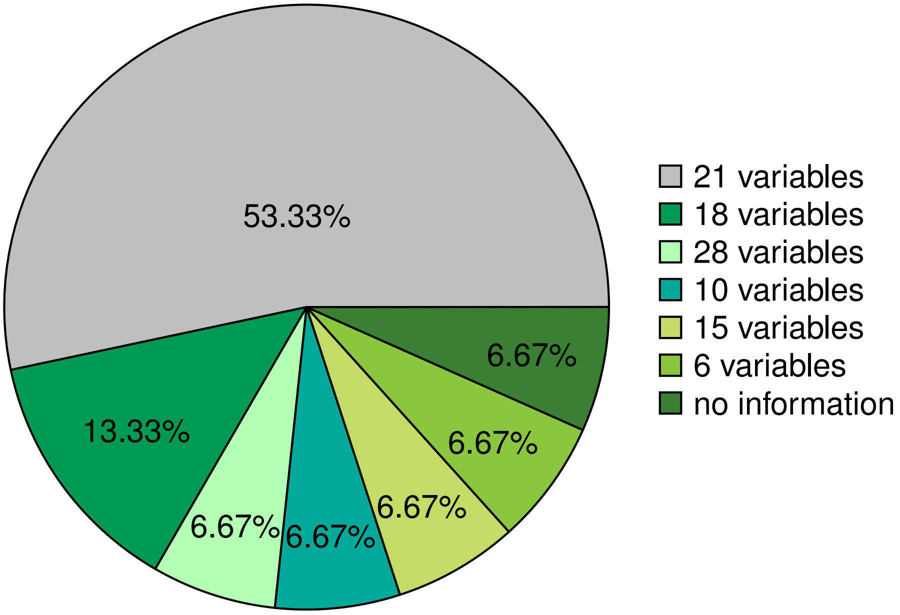
The AQ-10 database is detailed in annex B. This public database contains a total of 10 variables related to questions formulated for autism screening. In addition to

Figure 10 – Proportion of databases used



these, other variables are of a personal nature, such as age, ethnicity, and family history of psychological conditions. This difference in the number of variables is illustrated in the chart in Figure 11, where, despite 80% of the articles using the same database, there is a variation in the number of variables, reflecting the additional questions related to personal data. Another important feature of this database is its division by age group: child, adolescent, and adult. The questions remain consistent across the three age groups, with the only variation being the method of response for children, which is usually provided by their parents.

Figure 11 – Proportion of the number of variables in the database

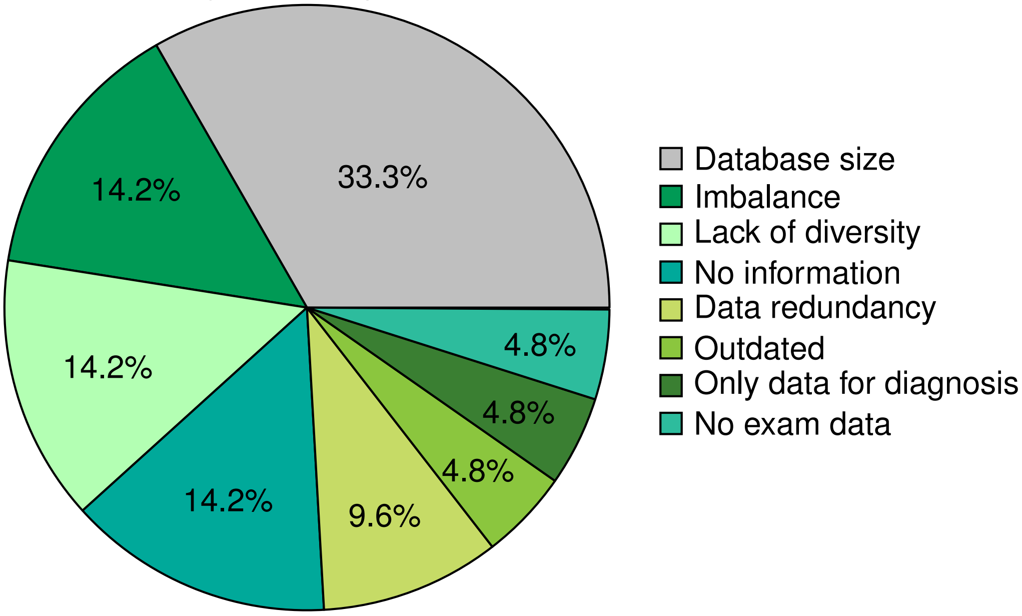


The other databases are also based on questionnaires. CHAT-10, used in the study by Romero-García et al. (2021), is similar to AQ-10. However, the database contains only questions for screening behaviors related to autism and is applied exclusively to children.

On the other hand, the study by Rajagopalan et al. (2024), which uses the SPARK + SSC, adopts a different approach. The database used contains medical and behavioral information on individuals with ASD and their families in the United States. It stands out because, in addition to including data related to the AQ-10 questions, it encompasses a wide variety of variables related to medical, familial, behavioral and developmental conditions, such as gestational age, feeding problems, neurological conditions, and the educational background of parents. The variables used in this database are presented in the Annex C.

Figure 12 presents the limitations found in the databases. The limitation most cited was the size of the databases, which affected 33% of the studies, followed by class imbalance and lack of diversity. This is related to the fact that publicly available databases often do not receive updates. Since they are based on questionnaires, they depend on new tests to populate these databases. Given that the databases used in the study come from similar sources, the lack of diversity due to this dependency is noticeable.

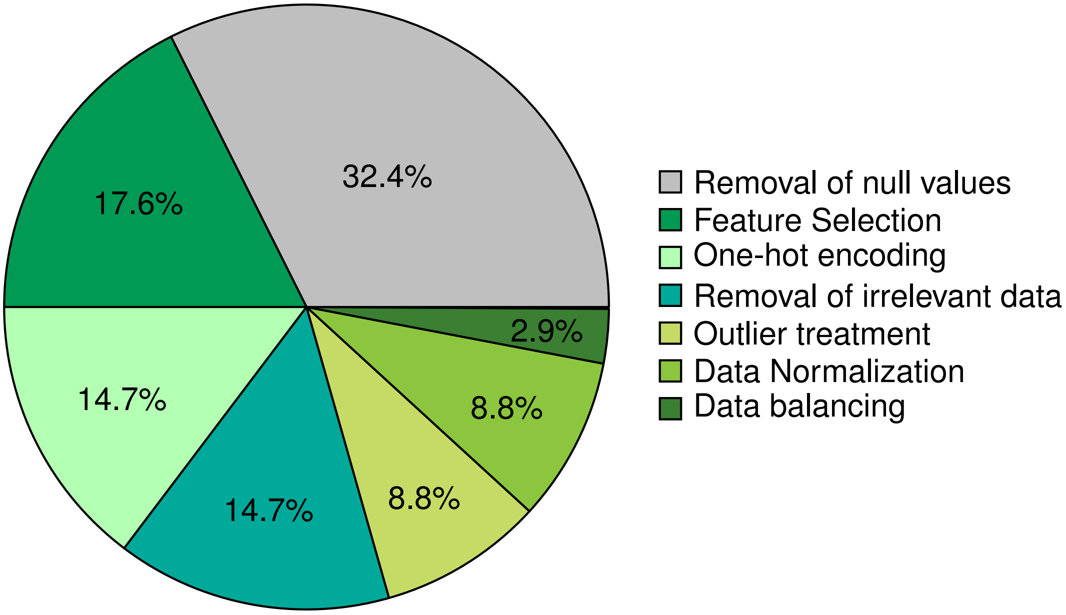
Figure 12 – Proportion of limitations found in databases



Although studies such as the one by Rajagopalan et al. (2024) suggest that one way to improve the diversity of databases would be to include diagnostic tests, such as imaging data and genetic information, the work of Naik et al. (2023) presents a divergent view. This study suggests that relying on diagnostic data could make it more difficult to obtain new data, as the process is more time consuming. An interesting point raised by Saranya e Anandan (2021) is the lack of updates in the databases. Since they are based on questionnaires, sometimes updates are made to these instruments by medical boards, but these changes are not reflected in the databases because they are outdated.

The pre-processing methods used by the studies were also analyzed, as shown in Figure 13. The technique most commonly applied was the removal of null values, used in 32.4% of the studies. Feature selection came second, with 17.6%, while other techniques such as outlier treatment 8%, hot encoding 14% and data normalization 8% were also used. Furthermore, 14.7% of the studies focused on removing irrelevant data and 2.9% performed data balancing.

Figure 13 – Proportion of preprocessing techniques used



The objectives of the studies analyzed reveal that 81. 25% focuses primarily on comparing machine learning (ML) models. This trend is illustrated in the graph presented in Figure 14, which also explains the wide variety of models shown in Figure 15.

Only the studies by Mittal et al. (2024) and Gill et al. (2024) had the primary objective of creating new models. However, even in these cases, traditional ML models were applied for comparison purposes. In addition, three studies focused on the implementation of only one specific ML model.

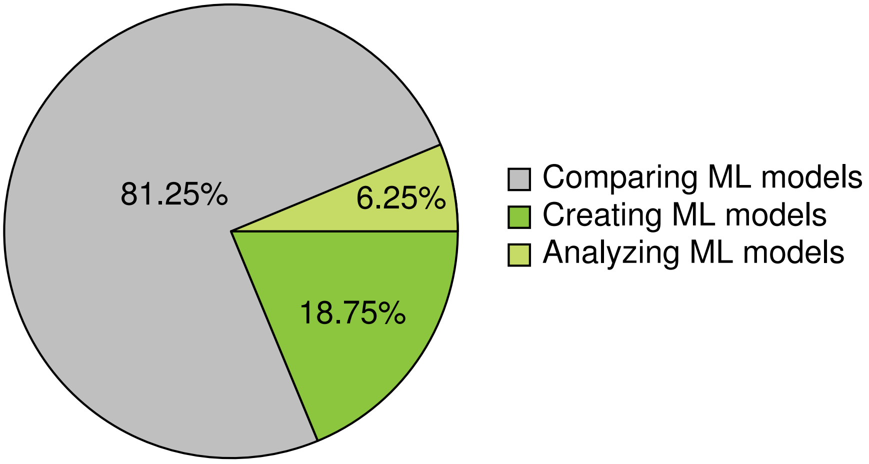
Despite differences in general objectives, all studies share a common secondary goal: evaluating the performance of ML models as a tool to assist in the early identification of Autism Spectrum Disorder (ASD).

Among the models used, the Support Vector Machine (SVM), Logistic Regression (LR) and Naive Bayes (NB) stand out, each used in at least 20% of the studies. Models such as K-Nearest Neighbors (KNN), Random Forest (RF), and Decision Tree (DT) appear in 13% of the studies. XGBoost (XGB) and Artificial Neural Networks (ANN) were used in 5 studies each, while Convolutional Neural Networks (CNN) and Linear Discriminant Analysis (LDA) were applied in only three studies. Algorithms such as AdaBoost (AB), Particle Swarm Optimization combined with Convolutional Neural Networks (PSO-CNN), Optimized Extreme Learning Machine (WOEM), Gaussian Naïve Bayes (GNB), and Multi-Layer Perceptron (MLP) appear just once, collectively comprising the "Other" category, representing 10% of the studies.

It is noteworthy that traditional models predominate, likely due to their established reliability, low computational cost, robustness, ease of use, and proven effectiveness across various contexts. Most studies rely on the same dataset, and even with simpler models, they achieve significant metrics, often surpassing 90

In contrast, the only study that used a distinct data set, incorporating information beyond standard interview data, was Rajagopalan et al. (2024), which reported an average precision close to 80%. This contrast suggests that the inclusion of more complex variables may impact results, potentially due to the increased difficulty of modeling diverse data or the need for more advanced techniques to leverage the dataset's potential fully.

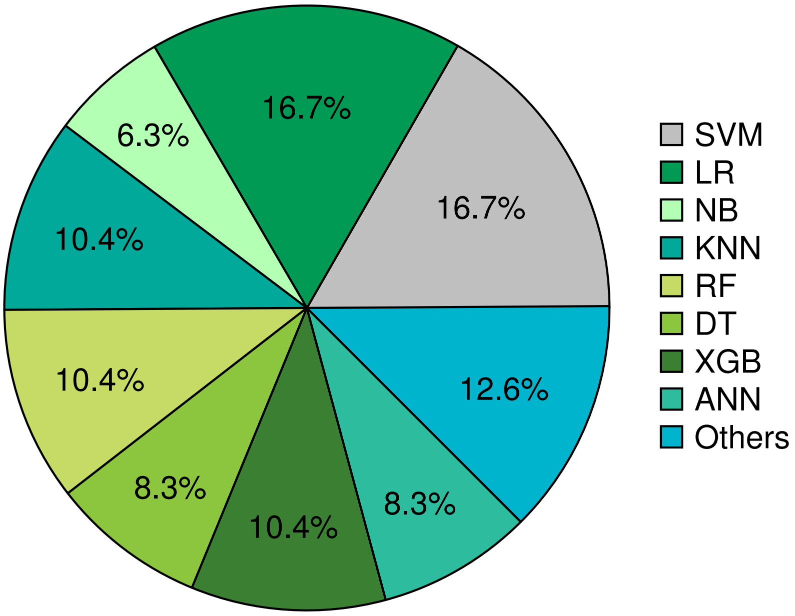
Figure 14 – Objectives of the studies



Although the importance of applying interpretability techniques is widely mentioned in almost all articles, both for model development and operational relevance, only four articles, representing 26.6% of the reviewed articles, applied interpretability techniques to their models, as shown in Figure 16.

Two articles utilized the Feature Importance technique to identify the most relevant features for classification. The study by Manoj e Praveen (2023) does not provide much detail on the implementation of "Feature Importance," only mentioning that the version available in the Extra Tree classifier was used. As a result, it was observed that the A4 and A9 features were the most important for the model. These questions are related to behavior, with A4 being "If there is an interruption, can I quickly get back to what I was doing?" indicating difficulties with routine disruptions, and A9 being "I think it is easy to tell what someone is thinking or feeling just by looking at their face", which

Figure 15 – Proportion of machine learning models used



is related to difficulties with socialization in ASD. The article by Romero-García et al. (2021) also mentions the use of Feature Importance but without much detail. In the three models presented, the variable that stood out the most was A7, followed by A9. The A7 question is also related to socialization, asking "When I am reading a story, I find it hard to understand the intentions of the characters."

In the study by Rajagopalan et al. (2024), which used a dataset with more features, the SHAP technique was applied using the TreeExplainer of the SHAP module in Python. The results showed that the variables related to feeding problems and neurological problems in children were the most significant.

The article in Hasan et al. (2022) tested four different techniques to calculate the importance of characteristics and identify the most prominent variables for the prediction of ASD. The techniques used were IGAE, GRAE, RFAE, and CAE, and were applied to three data groups: children, adolescents, and adults. For children, the most important variable was A4; for adolescents, A5, which is related to socialization, asking 'I find it easy to read between lines' when someone is talking to me"; and for adults, the variables A9 and A5.

These results highlight the importance of evaluating characteristics so that healthcare professionals consider the most relevant variables when screening for ASD, as identified by the models.

After analyzing the selected articles, it is possible to conclude that the use of machine learning models to detect autism spectrum disorder (ASD) is promising. However, some significant limitations were identified. Most of the models relied on the same dataset, which reduced the diversity in analyzing limitations and performance.

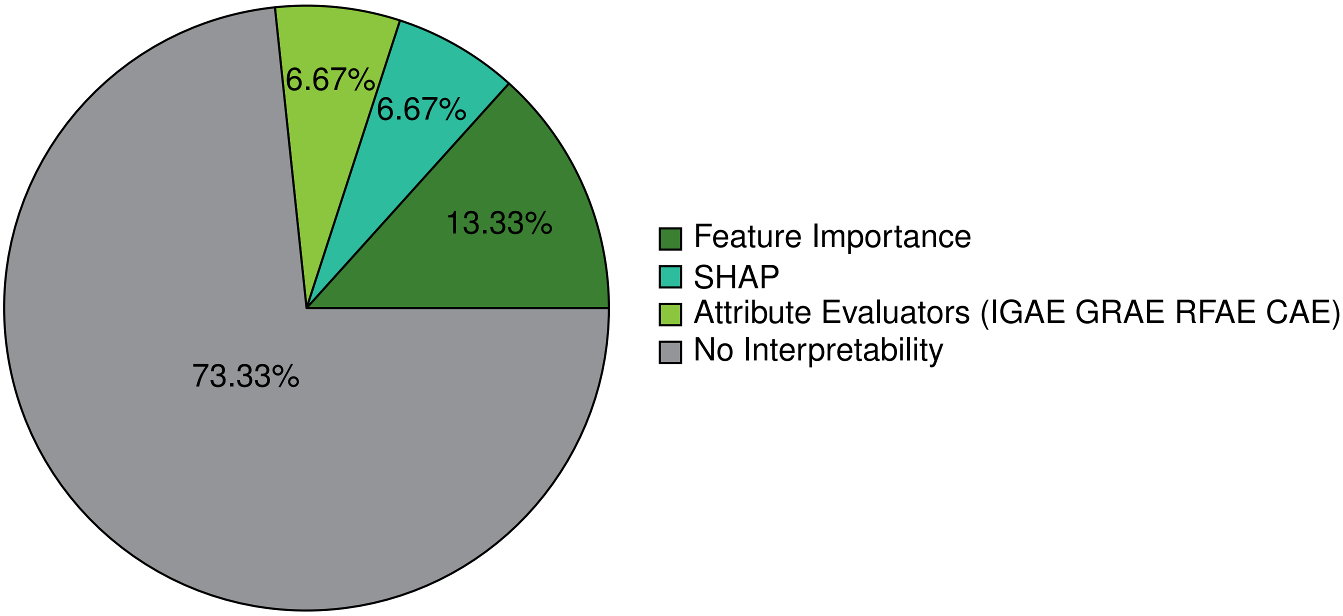


Figure 16 – Proportion of applied interpretability techniques

This lack of diversity affected the comparative analysis between the articles, as the models focused primarily on similar data sets and development objectives.

All the reviewed studies were based on data derived from questionnaires. Although these approaches do not involve imaging exams or genetic testing, they rely directly on the responses of the respondents to feed the datasets. This dependency was highlighted in several studies that noted the limited amount of data available to train and test models. One of the reviewed studies presented a more comprehensive data set that combined questionnaire data with additional variables. However, even this data set still relied on interviews for data collection, underscoring the need to diversify data sources.

This analysis was crucial for the proposed work, as the goal was to create a data set that was not based exclusively on diagnostic data. It became essential to understand which variables would be relevant and necessary. Although none of the data sets analyzed fully aligned with the objective of this study, additional questions and variables from existing data sets helped identify potential key elements for developing a new data set.

The limitations identified in the reviewed studies provided valuable information for addressing these challenges when creating a new data set. Among the most recurring issues were the lack of data and class imbalance. Despite these challenges, all reviewed models employed binary classification approaches within supervised learning, without exploring alternative ML techniques, such as unsupervised learning. Innovations observed in the studies focused on two areas: the development of new learning techniques and the application of alternative models, such as deep learning, in some cases.

Interpretability was highlighted as an important factor in half of the articles analyzed, emphasizing its value for the development and application of models. However, only four studies (26.6% of the reviewed articles) effectively applied interpretability techniques to their models. These techniques generally aimed to identify the most relevant features for model prediction, which is essential for future practical implementations, especially in the medical field. Combining interpretability with models can provide significant support for medical decision-making, adding reliability and transparency to predictions.

Despite the exclusive reliance on interview data, all the reviewed studies prioritized autism detection as the primary objective. This somewhat contrasts with the broader ASD literature, which typically considers interviews as tools for identifying autism suspicion rather than definitive diagnostic methods. For this reason, the present work chooses to adopt the term "risk factor," ensuring that the model is understood as a tool to assist in identifying risk factors rather than as a definitive diagnostic solution.

Furthermore, it was noted that none of the analyzed studies applied specific technical methodologies to implement ML. Since this work is conducted in collaboration with a healthcare provider, the CRISP-ML methodology was chosen to ensure an efficient process in all stages, from project creation to implementation. CRISP-ML was selected as an evolution of the well-established CRISP-DM, incorporating aspects related to model interpretability. Given the relevance of the topic in the medical context, interpretability plays a crucial role, contributing to a better understanding of predictions and increasing confidence in the results.

Even with careful planning, some limitations must be acknowledged in the literature review. The choice of databases, for example, may have excluded important studies published elsewhere or in other languages. Additionally, the search terms used might not have captured all relevant works, especially in a broad topic like autism, which involves various areas of knowledge.

There is also the so-called publication bias, where studies with positive results are more likely to be published than those with negative or inconclusive findings. This can influence what is most visible in the literature. Finally, it’s important to note that even with objective criteria, the selection and interpretation of studies involve researcher judgment, which introduces a degree of subjectivity.

Recognizing these limitations is essential to ensure transparency and demonstrate the care taken throughout the analysis.

3.5 CONSIDERATION OF THE CHAPTER

This chapter presents a systematic literature review conducted using the PRISMA methodology, based on studies that applied Machine Learning (ML) techniques for the detection of Autism Spectrum Disorder (ASD). The analysis reveals that most studies focus on comparing ML models, favoring traditional approaches due to their reliability and low computational cost. However, reliance on the same dataset limits result diversity. Additionally, few studies explore model interpretability, a crucial factor for medical applications. Therefore, diversifying data sources and adopting more robust methodologies, such as CRISP-ML, are essential to enhance future research and applications in this field.

4 PROPOSED METHOD

The objective of this chapter is to present the proposed methodology to achieve the main objective of this work, which is to apply the CRISP-ML methodology to create a database focused on ASD and train an ML model to predict beneficiaries at risk of ASD. The proposed method is structured into six main phases, each contributing to the overall objective, with specific steps outlined within each phase.

Section 4.1 details the project planning phase, describing the selected problem, the objectives to be achieved, and the database to be used. The following three sections, 4.2, 4.3, and 4.4, focus on data exploration and cleaning, analyzing the predictive potential of the data, and enriching the data through feature engineering. Section 4.5 describes the planning for the construction and evaluation of the machine learning model for the proposed objective, describing the models chosen and the metrics for their evaluation. The stage of prediction analysis, along with interpretability techniques, is described and planned in Section 4.6. Section 4.7 presents the planning for the application and implementation within the company.

4.1 PROJECT INITIATION AND PLANNING

The first phase of the project involves meetings with the company's team of specialists to understand the existing process, identify where the ML model will be applied, define objectives, and clearly outline the problem to be solved. For this study, the focus was on the topic of ST requests, with an emphasis on ASD.

Currently, the company does not have a formal process in place to track ST requests or monitor patients with ASD. As a result, the company faces several challenges related to the diagnostic and treatment process. This is why this topic was chosen: to investigate how ML methods can be applied to address these gaps and improve the assistance provided. The healthcare plan provider's process involves offering beneficiaries access to medical consultations, exams, and various procedures. To access these services, beneficiaries must make requests that are recorded in the company's systems. These requests and records form the database through which the company monitors and manages the use of the health services offered.

The current process for monitoring beneficiaries with ASD involves analyzing ST requests. ST refers to interventions based on practical methods and techniques designed to assist individuals with atypical development who show delays and impairments in social interaction and language, as well as in the emotional, cognitive, motor, and sensory domains. When recorded in the company's systems, these ST requests are made through medical requests, where the physician describes the reason for the therapy request. This description allows for the identification and monitoring of beneficiaries with ASD.

However, this process has some gaps. Health operators only become aware that a beneficiary has ASD when ST is requested, and, often, by this point, the beneficiary and their family have already gone through the entire investigation process. This leads to the need for more integrated support, which results in care disparity and fragmentation of the services provided. Another critical point is that there can be a delay between initial signs, the search for help, and diagnosis, causing a significant delay in the start of diagnosis and treatment, potentially compromising crucial early interventions for the proper development of children with ASD. This situation also reflects an inefficient use of available resources, as late detection prevents optimization of early interventions. Another challenge caused by the lack of a structured process is the difficulty in creating a database that allows one to evaluate the effectiveness of the services provided, implement improvements in the systems, and analyze and control whether the ST aligns with the patient's situation.

Analyzing the topic, two problem domains were identified for seeking solutions with ML: addressing pre-diagnostic and post-diagnostic scenarios. The post-diagnostic domain involves investigating methodologies for analyzing recommended medical therapies and detecting potential anomalies. On the other hand, the pre-diagnostic domain focuses on researching techniques to identify individuals who present risk factors for ASD. Subsequently, an analysis of the available data was performed to determine which problem would be the most suitable for the research. The available data consisted of the beneficiary usage database related to exams, consultations, and procedures requested by health plan beneficiaries. This data contains detailed information on the requests, including procedure codes, request dates, and personal information about the beneficiaries, such as age and sex.

After observing and discussing the data, it was decided to focus on pre-diagnostic solutions, first creating a representative database of the problem to assist in data analysis and later developing an ML model with the created database, aiming to predict patients at risk of developing ASD. The application of ML solutions in this area can benefit both beneficiaries and the company by allowing better monitoring of the beneficiaries and their family throughout the diagnostic formalization process, better utilization of resources, understanding usage patterns, and improving overall care.

Following the CRISP methodology, the project's overall goal was defined as improving the efficiency of ASD care, diagnosis, and treatment, with a focus on optimizing and speeding up the process. The system's main goal was to develop a system capable of identifying potential beneficiaries at risk for ASD. In addition to the main objectives, secondary goals were also listed, including:

- Creation of a database with important variables for ASD identification;
- Labeling of consolidated data;
- Exploratory data analysis;
- Search for usage patterns;
- Creation of an ML system based on the created database; and
- Creation of a system capable of identifying beneficiaries at risk for ASD.

Figure 17 illustrates the project flow, highlighting the key stages in the development of the predictive model and the implementation of pre-diagnostic solutions. The first phase involves defining the objectives and goals of the project. The second phase focuses on creating the list of variables in collaboration with the specialists, followed by the creation of the raw database, which is then processed through a pre-processing step. This process results in the final database, which is used in phase 5 to train the selected ML models. After training, the models are evaluated and the best performing models proceed to interpretability analysis in collaboration with the specialists.

4.2 DATA AUDIT, EXPLORATION, AND CLEANSING

The objective of this stage is the construction of a database related to the problem, with the aim of understanding how variables are distributed, identifying potential flaws and biases, and analyzing the overall distribution of the data. This stage was divided into two phases, as illustrated in the flow chart in Figure 18. Phase (A) involves the creation of the dataset, including the definition of the variables to be analyzed, the preparation of a list of these variables, and the request for the raw version of the database from the responsible team. Phase (B) refers to the exploratory analysis and pre-processing of the database.

Due to the fact that the company did not have a specific database for the problem they wanted to solve, it became necessary to create a custom database. To build this database, meetings with ASD specialists were held to identify the most appropriate variables from the request database to represent the topic effectively. The first step consisted of selecting relevant and representative variables in the context of ASD. For this selection, the knowledge of experts on the topic of ASD, professionals involved in the company's request process, and variables found in public databases from the literature review were used. The specialists involved in this phase included a physician with experience in autism and specialized therapies, an occupational therapist who works

Figure 17 – Project Flowchart

./Figuras/process-flowchart.png

directly with interventions for children with ASD, and an analyst responsible for reviewing requests for specialized therapies. These professionals, who deal with the topic on a daily basis, actively contributed to ensuring that the selected variables reflected the clinical reality and the challenges faced by both healthcare providers and beneficiaries. The final list included 68 variables related to risk factors for the manifestation of ASD, excluding data directly related to diagnoses. The selected variables are detailed in the appendix A.

The second step consisted of refining the initial list to establish the logic necessary to extract data from the company's database. During this process, some variables had to be adjusted. For example, when considering the risk factor that the mother has gestational diabetes, and since the diagnosis was unavailable, it was decided to check whether the mother had undergone a specific diabetes test. A similar approach was adopted for all the listed variables, seeking viable alternatives within the available data. However, some variables had to be discarded because they could not be accessed with the company's current database.

The query to the database returned 34,180 records of beneficiaries aged 0 to

Figure 18 – Pre Process Flowchart

./Figuras/pre_process.png

18 years. Various exploratory analyses were performed to identify inconsistencies and missing data. Adjustments were made to correct for these inconsistencies and, during this process, some variables were discarded due to lack of data. Furthermore, records with ages incompatible with the scope of the project were excluded, as well as those of beneficiaries without request records.

By the end of this stage, the database consisted of two types of data: numerical and binary. Binary data related to the request for exams, procedures, or consultations. A value of 1 was assigned when there was a request and 0 when no request was recorded. The numerical data represented the beneficiary's age at the time of the request. In addition to information about exams, consultations, and procedures, records of the current age, sex, and age of the parents at birth were included.

During this phase, the data was also labeled. This process was performed manually by a team consisting of two ASD specialists and one authorization specialist using a blind labeling procedure. The data were labeled as follows: in order to request a TS, the beneficiary needed a medical request indicating the necessity of follow-up, and the doctor also specifying whether the child had any condition, such as ASD. Based on this information, an analysis was performed to determine whether the child had ASD according to the medical request. If there was no request for a TS exam or if the request was made but the doctor indicated a condition other than ASD, the beneficiary was classified as not having ASD.

4.3 DATA PREDICTIVE POTENTIAL EVALUATION

In this stage of the CRISP-ML process, an analysis of the potential application of ML is performed, with the aim of evaluating the representativeness of the database and the quality of the information it contains.

The first analysis focused on the number of records available in the database. As illustrated in Figure 19 (a), a significant imbalance is observed in the distribution between the groups labeled ASD and non-ASD. The graph shows that only 4.27% of the beneficiaries requested TS, while 95.73% did not make this request, resulting in a considerable imbalance in the data set for analysis.

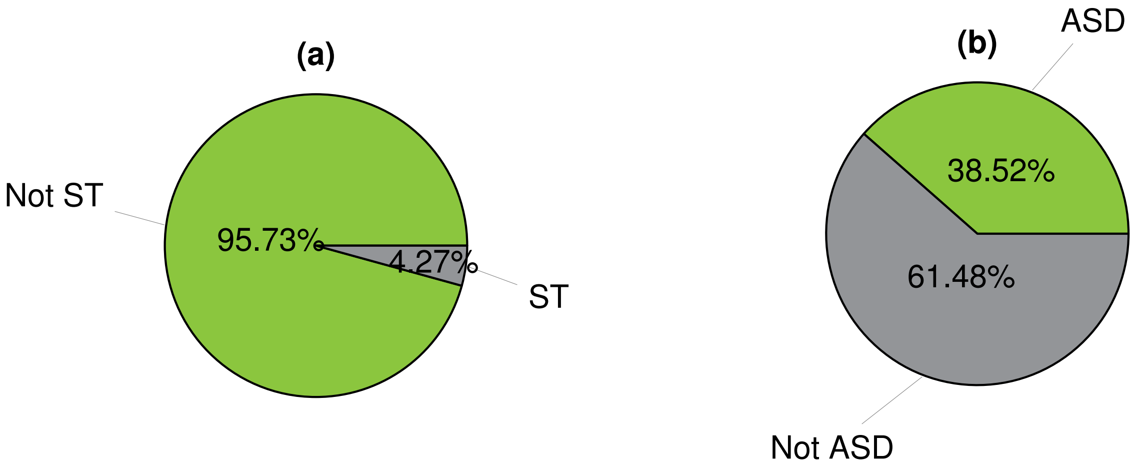
In addition to the imbalance, another concern was raised during discussions with the team of specialists: The cases labeled as "non-ASD" may not, in fact, be true negatives. That is, there is the possibility of false negatives, where patients with ASD did not request TS and were incorrectly labeled. To mitigate this situation and improve the accuracy of the analysis, it was decided to use only beneficiaries who requested TS, as this group contains medical evidence that confirms the absence of ASD, thus reducing the risk of false negatives.

The graph presented in Figure 19 illustrates the proportion of data in two different contexts. The first part, shown in (a), shows the proportion of beneficiaries who requested special therapy (TS). In this graph, 4.27% of the beneficiaries requested TS, while 95.73% did not. The second part of the chart, shown in (b), focuses on beneficiaries who requested TS and shows the proportion between those diagnosed with ASD and those not diagnosed. In this case, 38.52% of the beneficiaries who requested TS were diagnosed with ASD, while 61.48% were not diagnosed with the condition.

After dividing the dataset, two potential approaches for applying machine learning algorithms were analyzed: binary classification and one-class classification.

The binary classification approach aims to identify beneficiaries at risk of ASD. To achieve this, the 1912 records of beneficiaries who requested ST were divided into two groups according to the labels ASD and NON-ASD. However, this approach has a limitation: the model learns only the patterns of beneficiaries who requested ST without incorporating information or patterns from beneficiaries who did not make such requests.

Figure 19 – Graph of the proportion of the database. (a) Proportion among beneficiaries who requested special therapy. (b) The proportion of beneficiaries in the ASD group among beneficiaries who requested special therapy.



To address this limitation, the possibility of using a one-class classification model was considered. This approach aims to learn the patterns of a specific group and identify similar instances in a larger group. In this context, all ST request records were used to create a model capable of identifying beneficiaries with characteristics similar to those already receiving ST, enabling predictions about beneficiaries who might request ST in the future.

In addition to the definition of the predictive models, it was essential to assess the validity of certain variables that had a low number of records in the dataset. Although these variables demonstrated technical relevance based on the literature and domain knowledge, their limited frequency required a careful evaluation to determine the most appropriate handling strategy. This step was crucial to preserve the overall consistency, and representativeness of the dataset.

To address this, an exploratory data analysis was conducted with a specific focus on identifying and quantifying these low-frequency variables. Following this identification, targeted discussions were carried out with the team of experts to assess the clinical and operational significance of each variable, despite its sparse occurrence. These interactions aimed to explore viable alternatives for retaining or enriching the information in a meaningful way. The insights and challenges from this phase were then incorporated into the Data Enrichment process, during which strategic solutions were proposed to enhance the robustness and comprehensiveness of the data used in model development.

4.4 DATA ENRICHMENT

In this final stage of data processing, the dataset was re-analyzed and subjected to additional pre-processing to address issues identified in earlier phases.

During the analysis, it was observed that some variables had a low number of records, compromising their ability to represent the problem adequately. To address this, a technique was applied to group similar variables. However, even after grouping, some variables still exhibited a low record frequency. Consequently, the decision was made to exclude these variables from the dataset, ensuring greater consistency and quality in the final dataset.

The resulting dataset contains 25 columns, including one label column and 1912 records. The variables are distributed as follows:

- Sex (binary): Sex of beneficiary;
- Current Age (numerical): The current age of beneficiary;
- Mother age at birth (numerical): Age of mother at birth of beneficiary;
- Advanced mother age at birth (binary): Checks mother's age at birth of beneficiary is over 35 years;
- Father age at birth (numerical): Age of father at birth of beneficiary;
- Advanced father age at birth (binary): Reviews father's age at birth of beneficiary is over 35 years;
- Mother hospitalized (binary): Searches for any hospitalization item or type of hospitalization guide in the mother's card, analyzing nine months before the child's date of birth;
- Mother diabetes (binary): Checks if the mother requested a specific test for diabetes during pregnancy;
- Cardiologist mother (binary): Checks if the mother requested a specific test for a cardiologist during pregnancy;
- Pediatrician (binary): Checks if the patient had a request for a consultation with a Pediatrician;
- Age Pediatrician (numerical): The age of the first request for a pediatrician appointment with the patient;
- Speech therapist (binary): Checks if the patient had a request for a consultation with a speech therapist;
- Speech therapist age (numerical): Age of the first request for consultation the patient had with a speech therapist;

- Occupational therapist (binary): Checks if the patient had a request for consultation with an occupational therapist;
- Occupational therapist age (numerical): Age of the first request for consultation the patient had with an occupational therapist;
- Psychologist (binary): Checks if the patient had a request for a consultation with a psychologist;
- Psychologist age (numerical): The age of the first request for consultation the patient had with a psychologist;
- Neurologist (binary): Checks if the patient had a request for a consultation with a neurologist;
- Neurologist age (numerical): The age of the first request for consultation the patient had with a neurologist;
- Ophthalmologist (binary): Checks if the patient had a request for an ophthalmologist appointment;
- Age ophthalmologist (numerical): of the first request for an ophthalmologist appointment;
- Critical Ophthalmologist (binary): Check if the patient was under three years old at the first request for an ophthalmologist appointment;
- Total Consultations Last 12 Months (binary): Total consultations in the last 12 months; and
- Age special therapy (numerical): Age of the first request for special therapy.

The numerical variables had missing values, as they were associated with the age at which specific procedures were requested. When no request was recorded, the corresponding age was also absent, resulting in missing values. For the binary classification model, both numerical and binary variables were used. For the one-class models, only binary variables were employed to ensure that these variables did not contain missing values.

The organization of each dataset and the respective number of columns are described below.

a) Binary Classification Dataset
   - Columns: 25;
   - Type: Binary and Numerical; and

- Groups: All variables in the dataset.

b) One-Class Classification Dataset

- Columns: 14;
- Variable Type: Binary; and
- Groups: All binary variables in the dataset.

The feature selection technique was applied using the Optuna library. Variable selection is a crucial step in the development of ML models, enabling the identification and retention of only the most relevant features for the problem under analysis. This reduces the dimensionality of the dataset, improves the interpretability of the model, and mitigates issues such as overfitting.

The Optuna library was utilized to perform variable selection in an automated manner. In the context of variable selection, the library was configured to explore different combinations of variables in the dataset, evaluating the impact of each combination on model performance. At the end of the process, Optuna returned the subset of variables that delivered the best performance, resulting in a more efficient model.

4.5 MODEL BUILDING AND EVALUATION

In this phase of the project, the ML algorithms, the metrics to be implemented, the training approach, and the performance analysis criteria are selected.

For the binary classification model, three ML algorithms were chosen: CatBoost, Random Forest (RF), and XGBoost. These models were selected based on the type of data and the presence of null variables. Tree-based models were considered robust, with the potential to deliver better performance to classify this type of data.

For the One-Class classification task, two classic models were chosen: Isolation Forest and SVM-OneClass. These models were selected to allow for an initial analysis of the performance of algorithms trained for this type of task.

All these models were implemented using Python[1] v3. For binary classification models, the following libraries were used: Sci-kit Learn[2] for Random Forest, the XGB library[3] for XGBoost, and the CatBoost library[4] for the CatBoost model. The choice of parameters for these models was made in two ways: first, using the default values from the libraries; and second, through a parameter search using the Optuna model, with the Optuna library[5].

[1] Python: <www.python.org/>

[2] Sci-kit Learn: <www.scikit-learn.org>

[3] XGBoost: <www.xgboost.readthedocs.io>

[4] CatBoost: <www.catboost.ai/>

[5] Optuna: <www.optuna.org/>

Parameter optimization is a method that is used to find configurations that enhance a model's performance relative to its metrics. For this purpose, the Optuna library employs optimization techniques based on search algorithms to efficiently find these configurations. Optuna uses the Tree-structured Parzen Estimator (TPE) algorithm by default for hyperparameter search, which is effective in complex search spaces by modeling the parameter distribution based on previous trials. In addition to TPE, the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) can be used for continuous spaces, and Random Search is available for basic comparisons. The pruning strategy, based on the Asynchronous Successive Halving Algorithm (ASHA), accelerates the process by discarding less promising trials. This execution-driven definition approach enables dynamic search space construction, making Optuna efficient and versatile for a variety of optimization problems (AKIBA et al., 2019).

For model training and validation, subsampling was applied by randomly splitting the data set into 70% for training and 30% for testing. This process was repeated 30 times, generating new models with each iteration. For each model, three metrics were calculated: accuracy, precision, and recall. The average of these metrics was then calculated to obtain a more reliable and consistent assessment of the model's performance.

In Table 4, the number of data points allocated for training and testing in the binary and one-class classification models are presented. In the One-Class Classification model, "PC" stands for Positive Class and "NC" stands for Negative Class, indicating the respective distribution of data between these classes.

Model	Total Data	Training Data	Test Data
Binary Classification	1912	1138	774
One-Class Classification	1912 (PC) + 247 (NC)	1138	774(PC) + 247(NC)

Table 4 – Data distribution for training and testing in the binary and one-class classification models

The evaluation was performed on models trained both with and without Feature Selection and with and without Parameter Optimization. For a more comparative analysis between the models, Critical Difference Diagrams (CDDs) were used. The critdd[6] library was used to generate the graph.

Critical Difference Diagrams (CDDs) were used to compare the performance metrics of ML models. These diagrams are helpful in identifying pairs of classifiers whose performance does not show statistically significant differences. Before applying the CDDs, the Friedman test is performed to determine whether there are significant differences between the classifiers. If the Friedman test does not reject the null hypothesis, it is concluded that there is insufficient evidence to distinguish the classifiers, and

[6] critdd: <www.github.com/hfawaz/cd-diagram>

the analysis is terminated. However, if the null hypothesis is rejected, indicating significant differences, a post hoc analysis is performed. In this study, the Wilcoxon-Holm test was used to make pairwise comparisons and identify statistically significant differences (DEMŠAR, 2006).

In addition to the analyses with the metrics, the performance of the hyperparameter search algorithms and feature selection techniques was analyzed by reviewing the graphs provided by the library. The flow of this stage is represented in Figure 20. Part A of the figure illustrates the training of binary classification models, while Part B represents the training of one-class classification models.

Figure 20 – Model Training Flowchart

./Figuras/models-flowchart.png

4.6 DERIVE BUSINESS INSIGHTS

The planning for this stage focuses on translating the results into actionable insights for the team of specialists, making the model's behavior more understandable and accessible.

To achieve this, the SHapley Additive exPlanations (SHAP) technique was chosen due to its widespread recognition for providing interpretability to machine learning models. Based on Shapley values derived from game theory, SHAP calculates the individual contribution of each characteristic to the predictions of the model (LUNDBERG; LEE, 2017). This approach enables the identification of variables with the most significant positive or negative influence on the model's decisions, promoting a detailed and transparent analysis of its behavior in relation to the data.

The application of this technique aims to validate the consistency of the model and to ensure that its decisions are both understandable and reliable. Furthermore, to improve the interpretation of the results, graphical visualizations were employed to highlight the most important variables between different groups in the data set. The primary objective is not only to validate the model's output, but also to confirm that they align with the problem addressed in the study.

This technique was applied exclusively to the binary classification and one-class classification models that demonstrated the best performance compared to others. For its implementation, the SHAP[7] library for Python was utilized, ensuring robust and efficient support for the execution of this analysis.

4.7 DEPLOYMENT AND REPORTING

The final stage focused on implementing the model and preparing detailed reports to evaluate its performance.

Strategies for integrating this work into the company system were discussed. The main proposal involves implementing a system capable of periodically analyzing data and generating a list of beneficiaries with a higher probability of requesting special therapy and being at risk of autism.

During this phase, all the code developed, including exploratory data analysis and model development, was documented and made available on GitHub, along with detailed descriptions of its functionality.

4.8 CONSIDERATION OF THE CHAPTER

This chapter detailed all phases of CRISP-ML, including the creation of the ASD dataset, which was not based on clinical diagnoses. The pre processing steps used and the variables in the database were also described. In addition, the planning of experiments with machine learning models was presented to detect risk factors for ASD and TS was presented. Finally, an explanation of how interpretability was applied to the process was provided.

[7] SHAP: shap.readthedocs.io/

5 EXPERIMENTS, RESULTS, AND ANALYSIS

This chapter presents the experiments conducted, the results obtained, and the analyses performed based on these results.

Section 5.1 will detail the protocols and configurations used to carry out the experiments. Section 5.2 will present the results achieved by each model, as well as a comparison between them using the selected metrics. This section will be divided into two parts: the first will address the results of the experiments conducted with binary classification models to identify beneficiaries at risk of ASD, and the second will present the results of One-Class classification models aimed at detecting beneficiaries who may require TS services. Finally, the last section, 5.3, will provide a detailed discussion of the results obtained, along with insights derived from the application of the SHAP technique.

5.1 EXPERIMENTS PROTOCOL

The infrastructure used for the experiments was provided by the Laboratory of Computational Intelligence (LABICOM)[1]. The computational environment consisted of a system equipped with an AMD® Ryzen 7 2700X processor, operating at a base frequency of 3.7 GHz and capable of reaching up to 4.35 GHz, with 16 processing cores, 16 GB of RAM, 1 TB of hard disk storage, and an additional 240 GB SSD.

For the binary classification task aimed at identifying beneficiaries with risk factors for Autism Spectrum Disorder (ASD), the algorithms CatBoost (CAT), Random Forest (RF), and XGBoost (XGB), were used. The models were trained on a dataset composed of numerical and binary variables, including missing values. Performance was evaluated under two conditions: using default parameters and after optimization with the Optuna library. Furthermore, feature selection (FS) was applied with and without hyperparameter optimization, resulting in a total of 12 configurations analyzed.

For the one-class classification task, designed to identify beneficiaries at risk of requesting ST, the Isolation Forest (IF) and One-Class SVM (OC_SVM) algorithms were used. These models were trained on a binary dataset with no missing values. Similarly to the binary task, performance was evaluated with default parameters and after optimization using Optuna. Performance analysis also included the application of FS with and without hyperparameter tuning, resulting in eight configurations.

The evaluation metrics included accuracy, precision, and recall, and each ex-

[1] LABICOM: <https://labicom-udesc.github.io/>

periment was repeated 30 times to ensure statistically representative results. The dataset was randomly divided into 70% for training and 30% for testing, using a random sampling technique. In addition to the analysis of the average metrics, the Holm test was used to perform a statistical comparison between the models. This test was applied to determine which models show similar performances in terms of accuracy, recall, and precision. The results showed the average ranking of the models, where a lower value indicates better performance. The Holm test was applied with a significance level (alpha) of 0.05, and the models were grouped according to their performances, highlighting the groups of models with statistically indistinguishable performances.

During the hyperparameter search process, the Optuna library was used to perform 500 trials, exploring predefined ranges for numerical parameters to balance model complexity and execution time. In addition, the available options for categorical hyperparameters were also considered. The default values provided by the CatBoost, Scikit-Learn, and XGBoost libraries were included within the defined search ranges for optimization. After identifying the best parameters, the model was executed 30 times, and the average of the results was calculated for evaluation.

For binary classification models, the parameters chosen for optimization were as follows. The CatBoost algorithm, learning_rate, n_estimators, max_depth, l2_leaf_reg, subsample and colsample_bylevel. For the RF algorithm, n_estimators, max_depth, min_samples_split, bootstrap, max_features, max_leaf_nodes, and min_weight_fraction_leaf. For the XGBoost algorithm, learning_rate, n_estimators, max_depth, subsample, colsample_bytree, and gamma. For one-class classification models, the parameters chosen were: For the Isolation Forest, n_estimators, max_samples, contamination, max_features, bootstrap. For the One Class SVM, kernel, degree, gamma, coef0, tol, nu, shrinking, cache_size, verbose.

The feature selection process was also performed using Optuna. The method involved executing 500 trials to evaluate different combinations of features, with the aim of identifying those that maximize the model's performance. After selecting the best features, the model was tested with 30 executions to calculate the average performance. The same procedure was applied to tests with hyperparameter optimization but using the selected features.

Both feature selection and hyperparameter optimization aim to maximize model accuracy. To ensure consistent results, the average accuracy was calculated based on five executions for each trial. To optimize processing time, native parallelism provided by the Optuna library was utilized.

Table 5 presents the hyperparameters used in each configuration of binary classification models. The first column lists the parameters for each model, while the

Table 5 – Default parameters (D) and tuned parameters (T) for each binary classification algorithm without FS.

CatBoost	**Default (D)**	**Tuning (T)**	**Tuning FS (FS - T)**
learning_rate	0.03	0.02623292011074082	0.34610210440537453
n_estimators	500	458	78
max_depth	6	1	7
l2_leaf_reg	3.0	9.01840557469614	8.819128051583093
subsample	0.8	0.7233460392680002	0.7779136268958362
colsample_bylevel	1	0.5972609521727729	0.9426037542737686
Random Forest	**Default (D)**	**Tuning (T)**	**Tuning FS (FS - T)**
n_estimators	100	356	378
max_depth	None	47	20
min_samples_split	2	6	3
max_features	sqrt	log2	log2
max_leaf_nodes	None	62	93
min_weight_fraction_leaf	0.0	0.05133044227928327	0.00962332125172342
bootstrap	True	False	True
XGBoost	**Default (D)**	**Tuning (T)**	**Tuning FS (FS - T)**
learning_rate	0.3	0.04384030720034894	0.02388950876563794
n_estimators	100	77	83
max_depth	6	3	3
subsample	1	0.24404394491752207	0.6856813712701775
colsample_bytree	1	0.7428446277828783	0.5792353969754332
gamma	0	0.4569748998601463	0.021132551230889297

Table 6 – Default parameters (D) and tuned parameters (T) for each one class classification algorithm without FS.

Isolation Forest	**Default (D)**	**Tuning (T)**	**Tuning FS (FS - T)**
n_estimators	100	196	134
max_samples	auto	0.3	0.7
contaminationt	auto	0.15	0.34
max_features	1.0	0.1492908690338267	0.40111145469559323
bootstrap	False	False	False
One Class SVM	**Default (D)**	**Tuning (T)**	**Tuning FS (FS - T)**
kernel_rate	rbf	sigmoid	poly
degree	3	4	2
gamma	scale	scale	auto
coef0	0.0	0.8456074491162815	0.8699419858410483
tole	0.001	0.0001	0.12105896892865563
nu	0.5	0.22499294740687295	0.0042917173575792045
shrinking	True	True	False

second displays the default values (D) defined by the libraries. The third column contains the parameters optimized through hyperparameter tuning (T) without feature selection (FS), and the last column shows the parameters optimized for models with feature selection applied. The same structure is followed in Table 6, which details the hyperparameters used for the one-class classification models.

Table 7 – Results of Binary Classificartion Algorith obtained with default (D) and tuned (T) hyperparameters, with and without feature selection (FS)

Algorithms	Acc	Recall	Precision
CAT-D	72.02% ±0.01	57.98% ±0.03	65.31% ±0.02
CAT-T	72.14% ±0.01	**57.74**% ±0.03	66.19% ±0.02
CAT (FS)-D	72.15% ±0.01	57.74% ±0.03	66.19% ±0.02
CAT (FS)-T	72.14% ±0.01	56.62% ±0.03	65.71% ±0.02
RF-D	71.35% ±0.01	55.13% ±0.02	64.63% ±0.04
RF-T	72.59% ±0.01	56.43% ±0.03	66.97% ±0.03
RF-FS-D	71.64% ±0.01	55,56% ±0.03	65.49% ±0.03
RF-FS-T	**73.61**% ±0.01	58.77% ±0.03	67.21% ±0.02
XGB-D	68.09% ±0.01	53.35% ±0.03	59.40% ±0.03
XGB-T	73.33% ±0.01	58.56% ±0.02	66.83% ±0.03
XGB (FS)-D	68.31% ±0.01	52.42% ±0.03	59.84% ±0.02
XGB (FS)-T	73.49% ±0.01	58.36% ±0.03	**67.85**% ±0.03

5.2 RESULTS

5.2.1 Binary Classification

Table 7 shows the results obtained by each binary classification model, using both the default (D) and tuned (T) parameters, and the feature selection (FS). The values presented are the averages of each metric for the test data after 30 independent runs, accompanied by their respective standard deviations.

Among the models evaluated, the CAT showed slight improvements compared to the model with optimized parameters or feature selection applied. The CAT-D model, with default parameters, achieved an accuracy of 72.02%. After optimization of the parameters, the CAT-T model showed a slight improvement, reaching an accuracy of 72.14%. The application of FS did not show significant effectiveness, as the models CAT-FS-D and CAT-FS-T maintained accuracy values similar to CAT-T accuracy of 72.14%. Furthermore, the recall and precision values were lower than those obtained by CAT-D, confirming that feature selection was not effective in improving the performance of this particular model.

Unlike the CAT model, the RF and XGB algorithms showed improvements when feature selection and parameter optimization were applied. The RF-D model, with default parameters, achieved an accuracy of 71.35%. After applying parameter optimization, the RF-T model achieved an accuracy of 72.59%, with an increase in recall 56.43% and precision 66.97%. This performance was superior to both RF-D and the model with feature selection without parameter optimization, RF-FS-D, which showed values similar to the default model, with an accuracy of 71.64%. The best performance was observed in the RF-FS-T model, which, with the combination of feature selection and parameter optimization, achieved an accuracy of 73.61%, a recall of 58.77%, and a precision of 67.21%. This model provided the best results in all metrics, standing out as the most efficient model among those evaluated.

The XGB algorithm, with default parameters, achieved a relatively low accuracy compared to the other algorithms evaluated, reaching an accuracy of 68.09%. The recall 53.35% and precision 59.40% metrics were also lower compared to other models. When feature selection was applied without parameter optimization (XGB-FS-D), the results did not show significant improvement, with an accuracy of 68.31%, staying close to the values of the default model. However, hyperparameter tuning proved to be quite effective for the XGB model, resulting in substantial improvements across all metrics evaluated. The XGB-T model, with only parameter optimization, achieved an accuracy of 73.33%, a recall of 58.56%, and a precision of 66.83%, significantly outperforming the default model. The combination of feature selection and hyperparameter tuning XGB-FS-T generated a slight additional improvement, reaching an accuracy of 73.49%. This result shows that parameter tuning was crucial for the significant performance improvement of the XGB model.

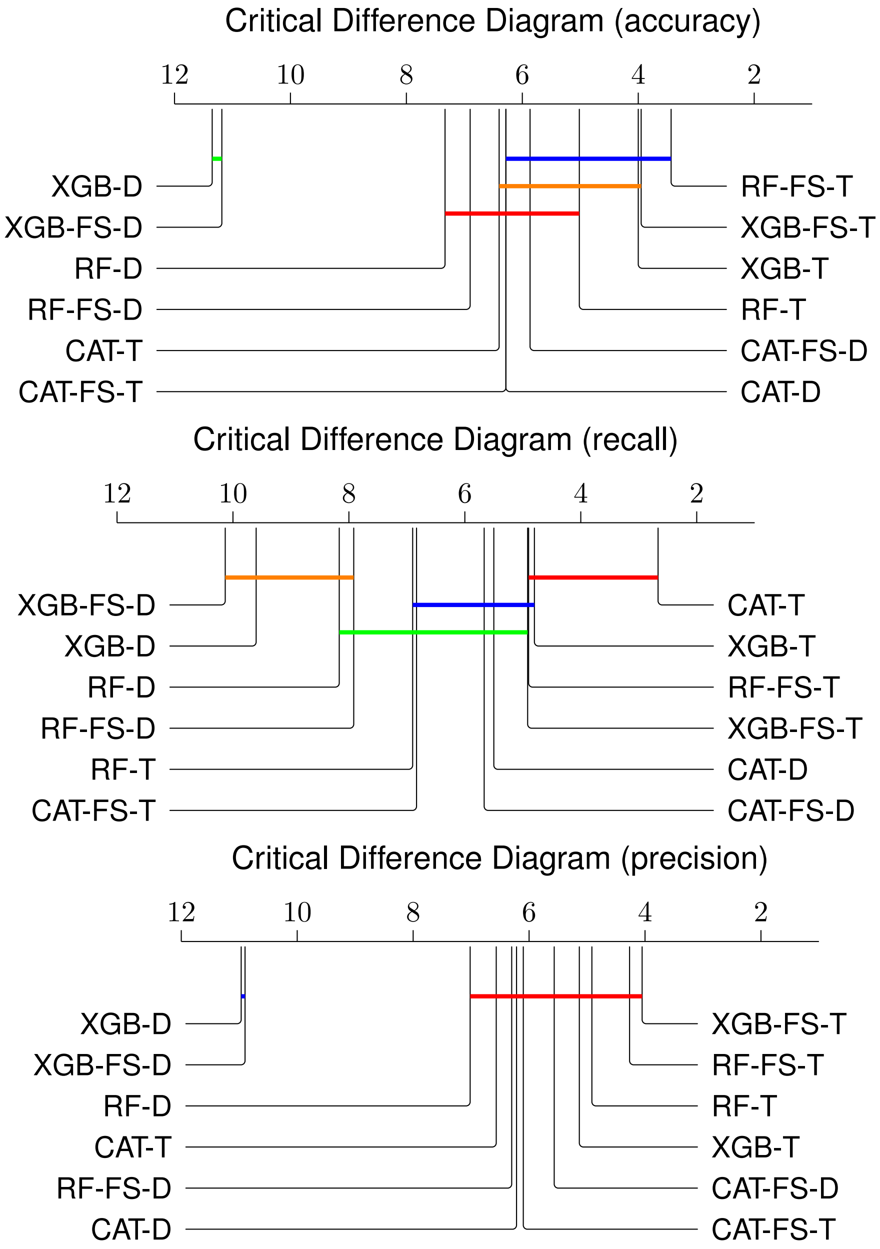
In this first analysis, considering the average values of accuracy, recall, and precision, it is evident that all the algorithms tested, when adjusted with optimized hyperparameters, showed improvements, although small, compared to the default version. Feature selection (FS) did not always result in substantial improvements, but in some cases, such as with Random Forest, the combination of FS with parameter tuning provided a significant performance gain.

The Random Forest model with feature selection and parameter tuning (RF-FS-T) achieved the best overall performance, reaching the highest accuracy of 73.61% and the best recall of 58.77%. Another model that showed similar performance was XGB with XGB-FS-T feature selection and parameter tuning, which achieved 73. 49% precision and 58. 36% recall. These results highlight that, for some models, the combination of feature selection with parameter optimization can lead to efficient performance.

For a better analysis of the algorithms, the Critical Difference Diagram (CDD) was generated from the results achieved by each model. As shown in Figure 22, the diagram provides a clear view of the relationship between the different algorithms, highlighting the behavior of each model in relation to the results obtained. The lines connecting the points on the graph illustrate the relationships between the algorithms, providing a visual understanding of the comparisons.

The CDD graph is presented as follows. On the horizontal axis, the average ranking of the models is shown, with models positioned further to the left indicating better performance. The lines connecting the models indicate whether there are sta-

Figure 21 – Critical Difference Diagram of the twelve binary classification models generated for the three metrics analyzed after 30 runs.



tistically significant differences between them. Models connected by lines do not show significant differences, whereas models not connected indicate that there is a statistically significant difference in their performance.

To facilitate interpretation, the groups shown in Figure 22 are connected by lines of different colors. Models with greater separation in the average ranking tend to show significant performance differences, while small variations may not be significant, depending on the number of experiments conducted.

The first CDD graph compares the model performance in terms of accuracy, identifying similarity groups represented by colored lines. The group with the worst performance consists of XGB-D and XGB-FS-D, connected by the green line. The second group, highlighted by the red line, includes the CAT, XGB-T, and RF-T models. The third group, represented by the orange line, encompasses all models based on CAT, as well as RF, RF-D, RF-T, and RF-FS-D. Finally, the highest-performing group, highlighted by the blue line, includes the models RF-T, RF-FS-T, XGB-T, and XGB-FS-T, as well as some based on CatBoost. The models' accuracy performance was significantly influenced by the application of hyperparameter tuning and feature selection, especially for Random Forest and XGBoost. These adjustments reduced dimensionality and highlighted relevant patterns, resulting in performance improvements. In this analysis, the RF-FS-T approach showed the best average performance in terms of accuracy.

Recall analysis presents a different configuration than accuracy, with a focus on the CAT-T model. The CAT-T and XGB-T models, trained exclusively with hyperparameter tuning, proved to be the most effective in terms of the recall metric. Next, models with feature selection and parameter tuning, such as RF-FS-T and XGB-FS-T, stand out. In the figure, four groups can be identified. The first, represented by the orange line, includes the models with the worst recall performance: RF-D, RF-FS-D, XGB-D, and XGB-FS-D, which show significantly lower values compared to the upper groups. The top group, with the best performance, is highlighted by the red line and includes the models CAT-T, RF-FS-T, and XGB-T, where it is evident that the models tuned with parameter adjustments perform better. For practical applications, the models in the group represented by the red line CAT-T, XGB-T, and RF-FS-T would be the best options.

In the precision analysis, the third diagram in the figure highlights the XGB and RF models, trained with hyperparameter tuning and feature selection, which showed the best performances. These models formed two distinct groups, with one group composed of XGB-D and XGB-FS-D, represented by the blue line, showing the worst performance and being statistically different from the others. All other models were grouped together in a single group, represented by the red line, indicating that statistically they have similar performances.

Comparative analysis of the models, using accuracy, recall, and precision metrics, demonstrated that hyperparameter tuning and feature selection are crucial to improve model performance. The models RF-FS-T and XGB-T stood out as the best options for the given problem. However, statistical differences indicate that, depending on the context, models such as CatBoost can also be competitive. For the continued interpretability analysis of the models, RF-FS-T was chosen, as it stood out as the best in all three metrics, integrating the high performance group in each.

Table 8 – Results of One Classe Classificartion Algorith obtained with default (D) and tuned (T) hyper-parameters, with and without feature selection (FS)

Algorithms	**Acc**	**Recall**	**Precision**
IF-D	51.01% ± 0.01	**94.98**% ± 0.01	39.26% ± 0.005
IF-T	72.07% ± 0.02	44.13% ± 0.08	58.34% ± 0.04
IF(FS)-D	72.30% ± 0.01	91.28% ± 0.001	54.24% ± 0.01
IF (FS)-T	**85.90**% ± 0.009	84.60% ± 0.001	75.0% ± 0.018
OC_SVM-D	60.30% ± 0.01	84.18% ± 0.02	43.97% ± 0.01
OC_SVM-T	76.36% ± 0.01	73.22% ± 0.01	61.22% ± 0.02
OC_SVM (FS)-D	64.63% ± 0.01	94.42% ± 0.02	47.59% ± 0.01
OC_SVM (FS)-T	83.07% ± 0.02	72.29% ± 0.02	74.97% ± 0.05

5.2.2 One Class Classification

Table 8 presents the results achieved by each one-class classification model, considering both the default parameters (D) and the tuned parameters (T), as well as the application of feature selection (FS). The reported values represent the mean of each metric calculated on the test dataset in 30 independent executions, along with their corresponding standard deviations.

In the analysis of the average values of the metrics, we observed that the IF algorithm showed high recall values in its model trained with the default parameters IF-D, reaching 94.98% recall. However, when analyzing metrics such as precision, we noticed a low value of 39. 26%, resulting in a precision of 51. 01%. This initial analysis reveals that, although the model is able to correctly identify most of the positive class examples, it presents a high number of false positives. After parameter tuning, the IF-T model showed a significant improvement in accuracy, which increased to 72.30%, but there was a considerable decrease in recall, which decreased to 44.14%, while precision increased.

The IF models with feature selection showed the best performance. The IF-FS-D reached an accuracy similar to that of the IF-T model, but with a high recall of 91.28% and a precision of 54.24%. Analyzing the accuracy of the IF-FS-T, we observed the best result among the four tests performed, with an accuracy of 85.90%, recall of 84.60%, and precision of 75.00%. This shows an improvement in both the identification of the positive class and the reduction of false positives, making the IF-FS-T model the most balanced between precision and recall among the IF models tested.

The One Class SVM models present performance where parameter tuning had an impact on improving the metrics. The version without tuning OC_SVM-D shows an accuracy of 60.30% with a reasonable recall of 84.18% and a low precision of 43.97%. This performance is similar to the model trained with feature selection using default parameters OC_SVM-FS-D, which shows an accuracy of 64.63%, high recall of 94.42%, and precision of 47.59%.

Models trained with parameter tuning showed a significant improvement in accuracy values. The model without feature selection OC_SVM-D has an accuracy of 76. 36%, a recall of 73. 22%, and a precision of 61. 22%, indicating a better overall performance, especially in balancing recall and precision. When adjusted with OC_SVM-FS-T feature selection, the model shows an accuracy of 83.07%, recall of 72.29%, and precision of 74. 97%, demonstrating balanced performance.

Hyperparameter tuning and feature selection application showed significant improvements in the performance of both the IF and OC_SVM models. In the case of IF models, the combination of feature selection proved to be effective, while hyperparameter tuning was the most impactful technique for OC_SVM models. It is evident that these techniques contribute to a more balanced performance between precision and recall, significantly reducing the impact of false positives.

Among the models analyzed, the IF-FS-T stood out as the best overall performer, achieving an ideal balance between precision and recall. However, for a more comprehensive analysis, the critical diagram was evaluated considering the three metrics: accuracy, recall, and precision, as presented in Figure 22.

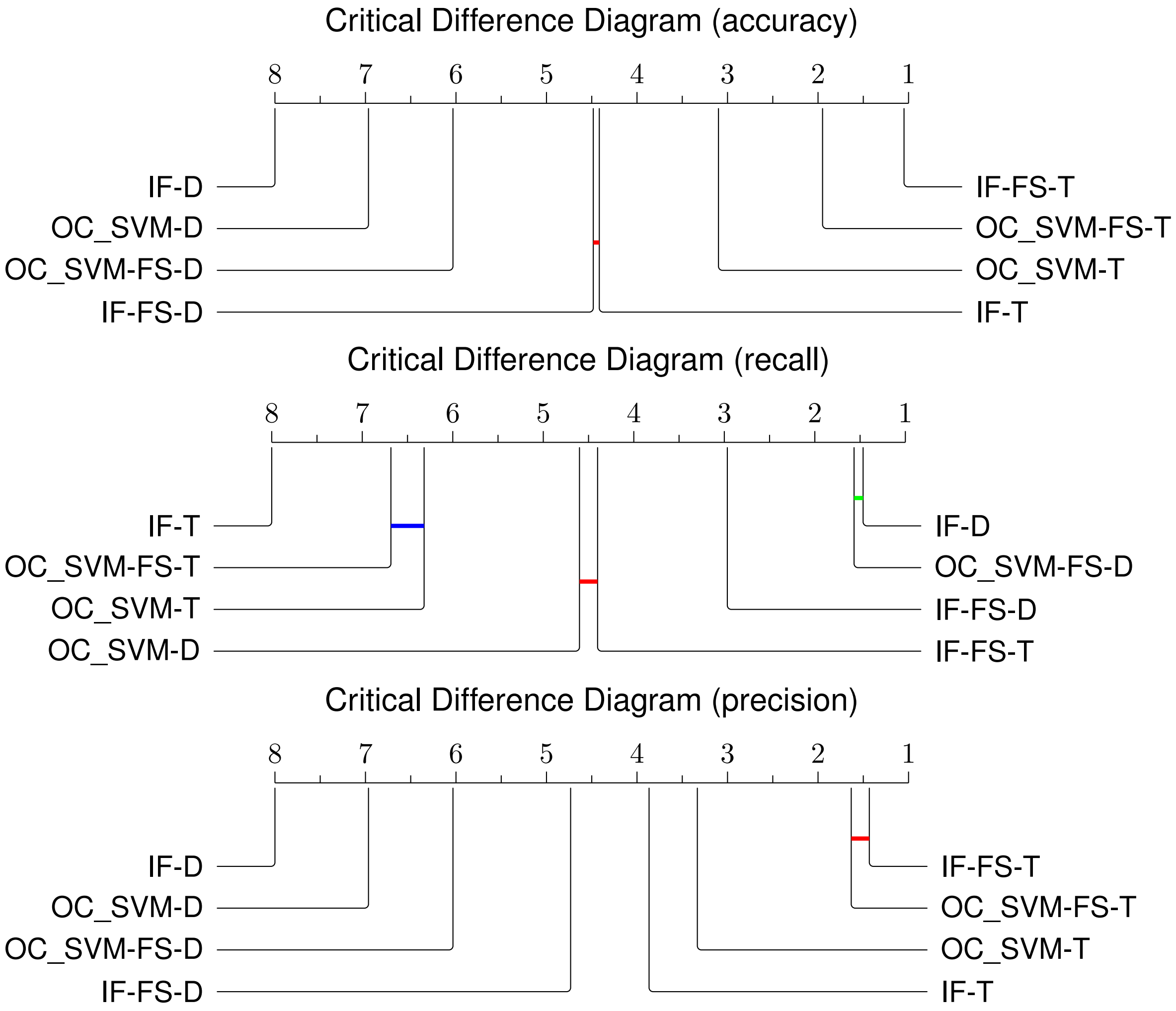
In the first diagram presented, it is possible to analyze the accuracy performance of the models. Among the eight models analyzed, IF-FS-T stands out with the best accuracy, followed by OC_SVM-FS-T. It is observed that models using parameter tuning and feature selection tools yield better results, followed by models with tuning only, with the exception of IF-T and IF-FS-D, which are the only groups connected by the red line, indicating that these models have similar performance.

In the recall diagram, a different configuration is observed, where models trained with default parameters show better results. The diagram presents three groups with no statistical differences: the blue group with OC_SVM-FS-T and OC_SVM-T, the group connected by the red line, which includes OC_SVM-D and IF-FS-T, and the group represented by the green line, which includes the best performances, such as OC_SVM-FS-T.

In the last precision graph, an organization similar to that of the accuracy graph is observed. Models trained with tuning and feature selection again stand out with the best results. The only group formed is represented by the red line, which concentrates the two best models in terms of precision: OC_SVM-FS-T and IF-FS-T.

When analyzing these results, it is essential to adopt a critical perspective regarding the dataset used. The models were trained exclusively with examples from positive class beneficiaries who requested special therapy. When applying the models to the general beneficiary set, the objective was to identify which individuals would be classified as belonging to the trained class. For this reason, the recall and precision

Figure 22 – Critical Difference Diagram of the eigth One Class Classification models generated for the three metrics analyzed after 30 runs.



metrics are of great importance.

Recall reflects the model's ability to correctly identify beneficiaries belonging to the positive class, which is crucial in scenarios where minimizing false negatives is a priority. Precision, on the other hand, evaluates the proportion of correct positive classifications, which are essential to avoid false inclusions of beneficiaries in the positive class false positives. For this problem, it is important to note that for the company, a false positive is less harmful than a false negative, as the system serves as a preliminary identifier and not as a definitive diagnosis.

When analyzing the models with the best recall values, a significant disparity in relation to precision values is observed, leading to a lower accuracy. However, the models OC_SVM-FS-T and IF-FS-T stand out because they show a better balance between precision and recall. Although they are statistically similar in terms of precision, the IF-FS-T model stands out with higher average values of accuracy and precision. For this reason, it was chosen to continue with the analysis.

5.3 ANALYSIS

At this stage, the best-performing models were selected for further analysis using the SHAP technique. The goal of this analysis is to go beyond performance metric values, aiming to understand the model's behavior concerning the variables and validate the insights obtained with the team of experts. This validation is essential to identify the most relevant variables for the problem at hand.

For the binary classification model, the Random Forest model trained with features selected by the feature selection technique and fine-tuned through hyperparameter tuning, referred to as RF-FS-T, was analyzed. Out of the 24 features originally present in the dataset, the model selected twelve, which are as follows:

- Current age;
- Advanced father age at birth;
- Mother diabetes;
- Age Pediatrician;
- Occupational therapist;
- Occupational therapist age;
- Psychologist age;
- Speech therapist;
- Speech therapist age;
- Ophthalmologist;
- Critical Ophthalmologist; and
- Age special therapy

Figure 23 provides a visualization of the importance of the features calculated using SHAP values. It compares the average absolute contributions of different features to the two classes: ASD and NON ASD. The vertical axis displays the features used by the model from the dataset, while the horizontal axis represents the average absolute SHAP values, quantifying the average impact of each feature on the model's predictions. Higher values indicate a greater relevance of the feature to the decision-making process.

Figure 23 – Feature Importance Using SHAP Random Forest Model

./Figuras/shap_rf_18x18.png

Own authorship

The analyzed dataset contains 1912 records, with 1181 beneficiaries classified as NON ASD and 731 classified as ASD. Since the model's objective is to identify beneficiaries at risk of ASD, Class 1 represents the ASD group, and Class 0 represents the NON ASD group.

The variable psychologist age stands out as the most relevant feature in the model, with the greatest impact on Class 0. This suggests that this variable is one of the primary discriminators between the classes. Analyzing the dataset regarding the age at the first psychological consultation request, the ASD group has an average age of 5 years, while the NON ASD group has an average of 7 years. This difference indicates that beneficiaries with ASD statistically tend to seek psychological consultations at a younger age. This is reflected in the histogram and boxplot for the psychologist variable, where the ASD group appears to have a higher frequency of consultations at younger ages Figure 24.

The importance of this variable in the model reflects the reality of the problem, especially considering that 69% of NON ASD group beneficiaries, 821 beneficiaries

Figure 24 – Distribution and Boxplot of Age at Psychological Consultation Request

./Figuras/psychologist_request.png

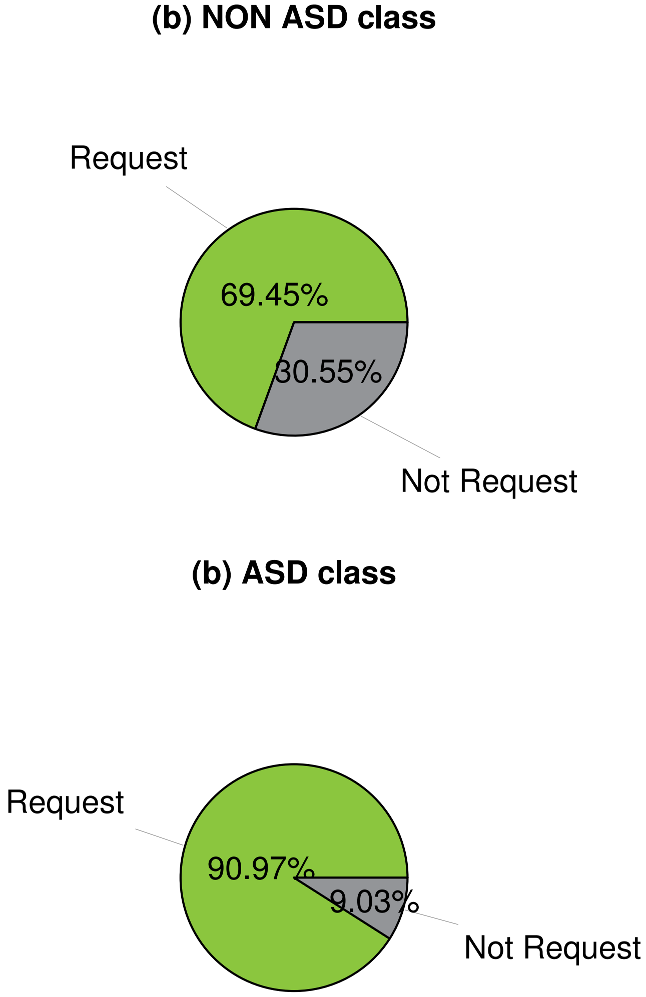
Own authorship

requested psychological consultations, compared to 90% of ASD group beneficiaries 665 beneficiaries, as observed in Figure 25. This high frequency of requests among ASD beneficiaries can be attributed to the fact that psychological consultations are often part of the diagnostic process for the disorder.

Other important features include occupational therapist and occupational therapist age, which also show significant contributions for both classes. Analyzing the data, the average age for the first occupational therapy request is 4 years for both groups as seen in the distribution of requests in the boxplot for occupational therapy Figure 26. However, 75% of the beneficiaries in the ASD group 553 beneficiaries requested occupational therapy, compared to only 37% of the beneficiaries in the NON ASD group. These numbers highlight the role of occupational therapy in the diagnostic journey of ASD as highlighted in figure 27.

The second most important feature is the Speech Therapist Age, which contributes more to Class 0 than to Class 1. Analyzing the dataset, we observe that more than half of the beneficiaries of ASD, 81% 597 beneficiaries, requested speech therapist consultations, compared to 54% 649 beneficiaries in the ASD group, as show in Figure 29. Highlighted by the histogram and boxplot for the age of the ophthalmologist Figure 28, it is possible to verify that the average age for the first consultation with the speech therapist is very similar, with the average for the ASD group being 4 years and that for the NON ASD group being 4.25 years.

Figure 25 – Proportion of psychologist request in NON ASD and ASD classes.



The remaining features show lower relevance. The SHAP analysis was crucial in demonstrating that, although all variables in the data set were related to the problem, the variables the model identified as the most important align with the typical patterns of health service requests, with psychological and occupational therapy being highly requested by individuals on the diagnostic journey for ASD.

For the analysis of the one-class model, the chosen model was IF-FS-T, a variant of the isolation forest that uses features selected by the feature selection technique and adjusts the parameters by tuning. Of the 12 features in the original training dataset, the model was trained on 4 selected features:

- occupational therapist
- psychologist
- speech therapist
- ophthalmologist

Figure 26 – Distribution and Boxplot of Age at Occupational Therapy Consultation Request

./Figuras/occupational_request.png

Own authorship

The training dataset consisted of 1912 records of beneficiaries who requested special therapy. The objective of the model was to learn the characteristics of these beneficiaries to identify individuals similar to the class used in the training.

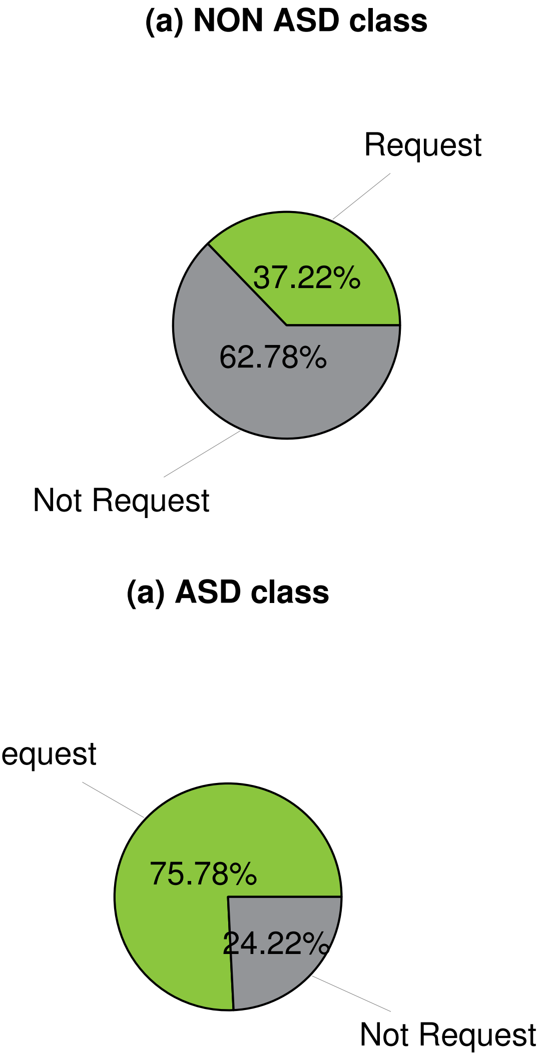
Figure 30 provides a visualization of the features and their respective importance in the model. The structure follows the same format as described for the Random Forest model, except that only one normal class is presented, representing the patterns of beneficiaries who request special therapy (TS).

The graph shows that the most important variable for the model is speech therapist, followed by psychologist, ophthalmologist, and occupational therapist. Analyzing the data for the group of beneficiaries who requested special therapy, it was found that 77% of the group requested psychological consultations, 65% requested speech therapy, 56% requested ophthalmologist consultations, and 51% requested occupational therapy.

In contrast, the negative class used during the model test, consisting of 273 records, presented the following percentages: 14% requested speech therapy, 41% requested ophthalmologist consultations, 11% requested psychological consultations, and 2% requested occupational therapy.

The four variables selected by the model align with the context of the problem, as these consultations are often part of the diagnostic process for conditions that re-

Figure 27 – Proportion of occupational therapist request in NON ASD and ASD classes.



(a) NON ASD class

(a) ASD class

quire TS, such as autism, attention deficit, and Down syndrome. These characteristics are crucial for identifying patterns among beneficiaries with similar behaviors. Figure 31 provides a summary of the number of requests for each of these variables within the database, which contains 1912 records of beneficiaries who requested TS. It is possible to observe the high demand for certain variables, such as psychologists and speech therapists.

The importance values of the features calculated by SHAP for both Random Forest and Isolation Forest may appear low due to the nature of the SHAP values, which are normalized to reflect the average contribution of each feature in the model. This is because, in models like Random Forest, the importance is distributed across many trees and often among various features. Factors such as a high number of features, the specific relevance of variables in the problem's context, and possible dataset imbalances can also reduce the magnitude of these values. Despite this, the presented values are significant as they help identify the most influential variables in the

Figure 28 – Distribution and Boxplot of Age at Speech Therapist Consultation Request

./Figuras/speech_request.png

model's decision-making, emphasizing the need for a deep understanding of the data and model to contextualize the results.

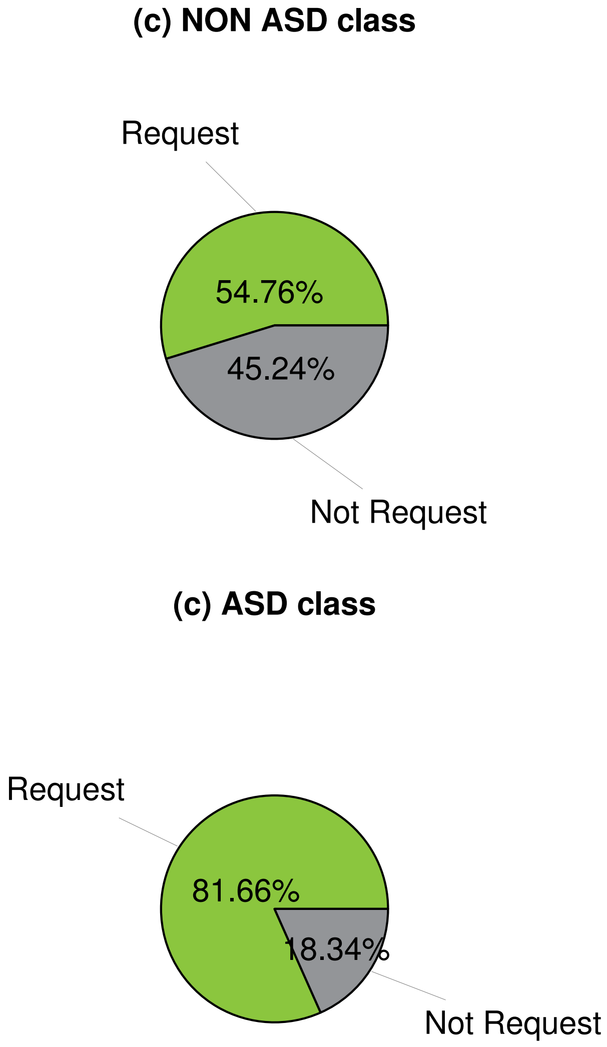
The SHAP analysis was shared with the team of experts and contributed to consolidating the models for implementation and testing in the company's system. The results demonstrated that the most important features identified by the models aligned with critical aspects of the problem, reinforcing the relevance of the models and the selected variables in the context of practical application.

Based on this feedback, it was possible to advance the development of the models and conduct the first tests in a corporate environment. The initial goal of these tests is to evaluate the performance of the models when dealing with new data from beneficiaries in pediatric age groups.

The objective of the One-Class model is to input a dataset of beneficiaries who have not undergone TS with the characteristics requested by the models. From the classification output of the model, the cases it does not identify as anomalies, that is, those similar to the class it was trained on, the beneficiaries who requested TS, will generate a list of potential beneficiaries who may come to request TS.

The One-Class model was designed to take as input a dataset consisting of beneficiaries who have not requested Special Therapy (TS). Using the characteristics identified as relevant by the models, the task of the One-Class model is to classify cases. Beneficiaries not classified as anomalies (those similar to those who requested TS, with whom the model was trained) will be included in a list of possible beneficiaries

Figure 29 – Proportion of speech therapist request in NON ASD and ASD classes.



(c) NON ASD class

(c) ASD class

who may request TS in the future.

Currently, the binary classification model will be applied to a dataset that contains beneficiaries who have already requested TS. Its purpose will be to identify, within this group, beneficiaries with characteristics associated with Autism Spectrum Disorder (ASD). Those classified as ASD will be added to a list that indicates individuals who have risk factors for the disorder.

The lists generated by both models will be submitted to the experts' team for evaluation, assessing the predictive power of the models. This feedback will enable the validation, adjustment, and enhancement of the models, fostering a continuous cycle of improvement. In addition to these first tests, future improvement points have already been identified that could significantly impact the models' performance. The main challenges include the following.

- Dataset: During the planning phase, several variables had to be excluded due to the lack of sufficient records. Future analyses of these variables may allow for

Figure 30 – Feature Importance Using SHAP Isolation Forest Model

./Figuras/shap_if_18x18.png

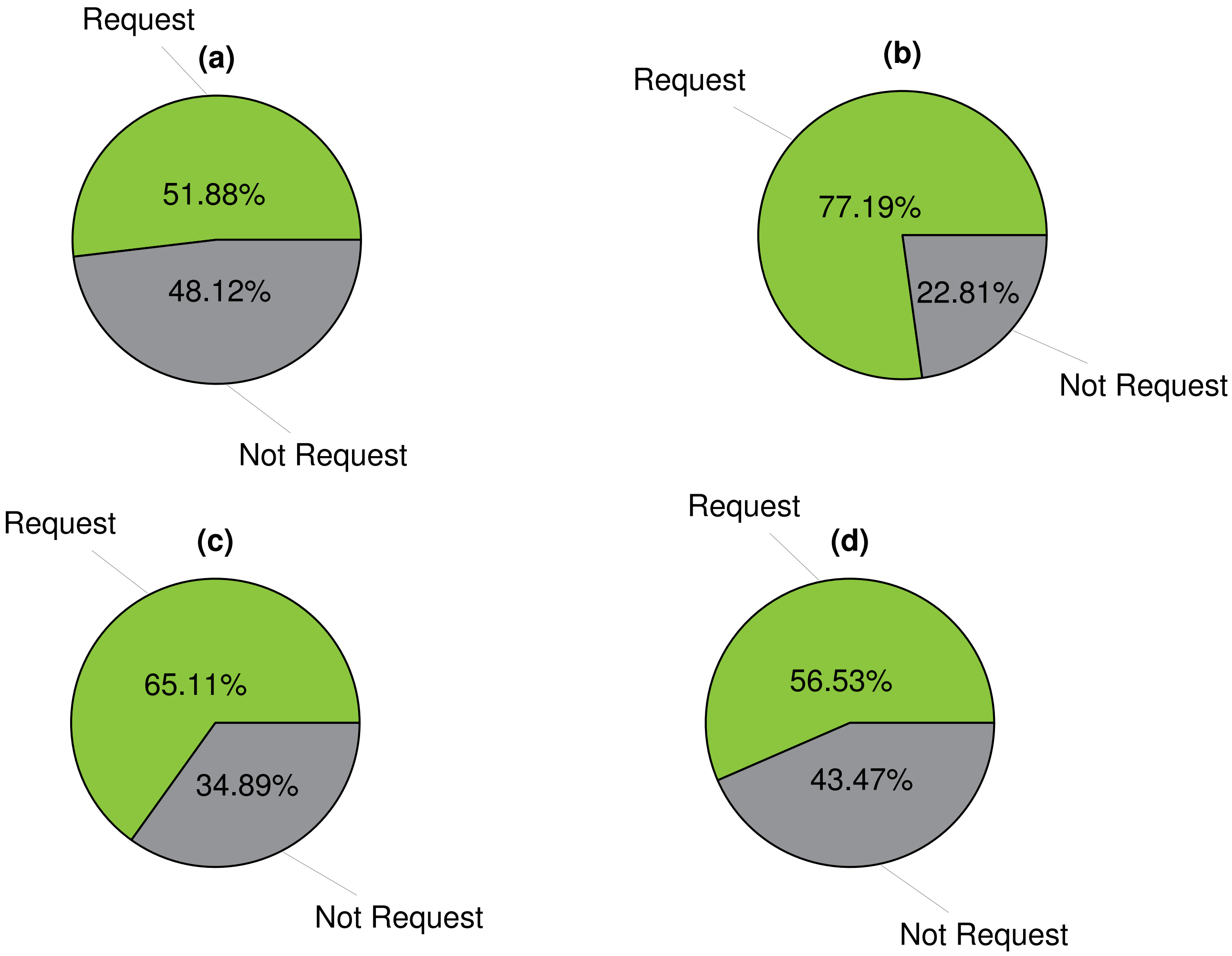
Own authorship

the inclusion of more comprehensive information and reduce the proportion of missing data.

- Sample Increase: It is necessary to consolidate a larger number of cases in both classes, including beneficiaries who have not requested TS, which could provide a more robust training base.
- Model Integration: The analysis suggests that the One-Class model could be better evaluated to act as a prescreening step, implemented in conjunction with the binary classification model, thus forming a more efficient classification pipeline.

The implementation of these models for testing will be crucial to improve the predictive performance of the models, ensuring greater precision in supporting specialists and contributing to the early identification of beneficiaries at risk or in need of specialized support.

Figure 31 – Graph of the proportion of request one class database. (a) Proportion of occupational therapist request. (b) The proportion of psychologist request. (c) The proportion of speech therapist request. (d) The proportion of ophthalmologist request.



The accuracy of the results obtained with both models was considered satisfactory, particularly when observing the improvements achieved through the application of feature selection techniques and hyperparameter tuning. Although accuracy values have not yet surpassed the 90% mark, the optimized models demonstrated significant enhancements compared to their initial versions, highlighting the potential of these approaches for future developments aimed at continuous performance improvement.

Beyond the perspective of healthcare providers, the beneficiaries themselves can also be directly and positively impacted. One of the main goals of this approach is to reduce the time between the first warning signs and the actual initiation of therapies, promoting a faster diagnostic process and more precise referrals. Early identification is essential, especially in cases where certain therapies are more effective when started during specific stages of a child's development. Therefore, accelerating the screening process can have a substantial impact on beneficiaries' quality of life and therapeutic progress.

Furthermore, by incorporating interpretable models to support clinical decision-making, the solution fosters greater transparency and trust among all stakeholders involved. This contributes to a more balanced relationship between healthcare providers and users, promoting more humanized, practical, and patient-centered care. The expert team's participation was crucial to ensuring the validity and practical applicability of the developed models, aligning them with clinical practice and the real needs of the beneficiaries.

5.4 CONSIDERATION OF THE CHAPTER

In this chapter, the results obtained from applying binary classification algorithms such as Random Forest, XGBoost, and CATBoost are presented. Furthermore, the performance of one-class classification algorithms, namely, one-class isolation forest and one-class SVM, was evaluated. All models were trained using a dataset that is not based on clinical diagnostics, and feature selection techniques, along with parameter tuning, were applied to optimize the algorithms' performance.

The initial results have shown promise, with particular emphasis on the models that incorporated feature selection and parameter adjustment techniques. For a more robust validation, the models were subjected to SHAP analysis, which provided a detailed interpretation of the contribution of each feature to the model's predictions. This analysis was carried out in collaboration with a team of experts in order to support the validity and relevance of the results obtained.

However, the need for continuous adjustments and improvements in both the dataset and the models themselves was highlighted. The constant evolution of the dataset and the optimization of the algorithms are crucial to refine the accuracy of the classifications and ensure the reliability of the decisions supported by the machine learning system.

6 CONCLUSION

The evolution of artificial intelligence (AI) and machine learning (ML) has brought significant transformations in various sectors, including healthcare. AI technologies offer innovative tools that improve practices and processes, particularly in the medical field, by automating complex tasks and providing more accurate analyses, which are essential for decision making and diagnosis. From classic ML models to deep learning approaches, these technologies have demonstrated great potential in helping diagnoses and improving medical processes. However, many companies still want to adopt these models but lack the knowledge necessary to implement them effectively.

In this context, methodologies such as CRISP-DM and its derivative, CRISP-ML, play a crucial role in ensuring the success and integrity of ML projects in healthcare. CRISP-DM offers a robust framework for organizing model development processes, while CRISP-ML focuses on model interpretability, ensuring that solutions are accessible and understandable for healthcare professionals.

Model interpretability has become an increasingly relevant topic in ML and deep learning research. With the growing importance of understanding how models make predictions and which variables are used for classification or regression, this technique proves essential not only for model development—helping configure and fine-tune them—but also for their application in the real world. Many situations require a deep understanding of the models for potential audits, in addition to providing valuable insights to solve problems and improve business decisions.

The application of methodologies such as CRISP-ML requires close collaboration between the technical team and field specialists, which is essential for the successful implementation of the system. With this objective in mind, this study proposes the application of CRISP-ML in developing a database aimed at diagnosing Autism Spectrum Disorder (ASD) and using ML models to assist in diagnosing and treating autism within a healthcare provider.

The adoption of advanced technologies in autism diagnosis has become increasingly sought after due to the difficulties and high costs of diagnosing this condition. The use of interpretable models contributes to improving diagnostic processes and enabling more efficient clinical decision-making, leading to faster and more accurate diagnoses, which can optimize the start of treatments.

Autism Spectrum Disorder (ASD) is a neurological condition with no cure, but it can be treated with therapeutic modalities that help in the development and quality of life of autistic individuals. However, delays in diagnosis and the prolonged investigative process result in delays in starting therapies, which compromises the development of affected children.

In the literature, various models have addressed the use of ML for early detection of ASD, relying on clinical data to accelerate diagnosis and identify signs of ASD more quickly.

One of the main challenges faced by healthcare providers—responsible for offering consultations, exams, and procedures to health plan beneficiaries—is the lack of formal diagnostic data. This limitation makes it challenging to identify and profile beneficiaries with Autism Spectrum Disorder (ASD). Constructing this profile is crucial for providers to align their resources with the needs of these individuals, enabling quicker identification and more accurate diagnosis.

Based on this need and the goal of this work, this study addressed the creation of a database tailored to the reality of healthcare providers. The proposal was to use information about beneficiaries, such as their medical history and procedures, to identify usage patterns that may be associated with the risk of ASD, without relying exclusively on formal clinical diagnoses.

To achieve this, relevant variables related to ASD were selected, considering usage data and information available in the provider's database. The selection of these variables was based on a literature review, analyzing studies that also used tabular data for ML applications in the ASD context, as well as specialized knowledge from medical professionals in the field. As a result, 68 relevant variables were identified, associated with risk factors for the disorder.

From these variables, a database was built, enabling exploratory analyses to better understand the profile of beneficiaries who requested Special Therapies,interventions based on effective methods and techniques to assist people with atypical development. After this analysis, the database was consolidated, containing a total of 1912 beneficiary records, of which 731 were identified as having ASD.

With this structured database, it became possible to evaluate the feasibility of applying ML models to identify beneficiaries at risk of ASD.

Three ML models were chosen for binary classification to classify beneficiaries in risk factors for ASD, and the One-Class technique was also evaluated with a different objective of identifying beneficiaries at risk of requesting TS. The application of this one-class classification model aims to better understand the data and create a model capable of analyzing data from beneficiaries who did not request TS, unlike the binary classification model, which was trained only with data from those who requested TS, separating them into ASD and non-ASD categories.

Among the models evaluated were Random Forest, XGBoost, and CatBoost for binary classification models, and Isolation Forest and One-Class SVM for one-class classification models. All models were assessed in their standard configurations as well as with parameter tuning and feature selection. The models were analyzed by comparing the mean metrics from 30 executions and using a critical difference diagram to assess whether there were statistical differences between the models.

For the XGB and RF models, significant performance gains were observed when trained with feature selection techniques and hyperparameter tuning. The XGB-FS-T and RF-FS-T models achieved accuracies close to 73%, although this result is still far from the above 90% accuracy levels. However, the performance improvement of these models compared to their non-optimized versions highlights the importance of selecting the appropriate variables and tuning parameters for developing more robust solutions. In addition to accuracy, metrics like recall and precision were also analyzed to evaluate the models' effectiveness in detecting relevant patterns.

CCD analysis showed that many models are statistically similar, suggesting that, while the models could still be improved, the approach used in this study represents a significant advance in building a tool to support the identification of beneficiaries at risk of ASD without depending on diagnostic data.

The RF-FS-T model was chosen for the application of SHAP techniques, providing significant contributions to both model understanding and validation. Among the most important variables were: Age of the first request for a psychologist consultation Age of the first request for a speech therapist consultation Age of the first request for an occupational therapist consultation, Request for an occupational therapist consultation.

The results obtained align with the reality of healthcare providers, considering that these professionals are the most sought after in the ASD diagnosis process. Thus, it can be seen that the model is learning correctly, capturing relevant patterns for identifying beneficiaries at risk of ASD.

For the One-Class classification model, both algorithms showed promising results. One-Class SVM performed better with hyperparameter tuning, while Isolation Forest stood out when feature selection was applied.

However, in both models, it was observed that precision values were significantly higher than recall values. This indicates that the models are more conservative in identifying beneficiaries at risk of ASD, meaning that the cases classified as positive have a high probability of being correct high precision, but there is a high rate of beneficiaries at risk who are not being correctly identified (low recall). This behavior may be related to data imbalance, the decision threshold, or the characteristics of the variables used in training.

The models that performed best were IF-FS-T (Isolation Forest with Feature Selection and Tuning) and SVM-FS-T (One-Class SVM with Feature Selection and Tuning), highlighting that the combination of these techniques can improve results. IF-FS-T achieved an accuracy of 85.90%, while ONE-SVM-FS-T achieved 83.07%, promising results for the One-Class approach, demonstrating its potential for future studies.

In the application of CDD (Class Distribution Discrepancy), IF-FS-T stood out in most metrics, except for precision, where models without tuning and feature selection performed better.

The IF-FS-T model was selected for interpretive analysis via SHAP. The most representative variable was "psychologist," indicating whether the beneficiary requested a consultation with a psychologist. As observed in the binary classification models, this variable was one of the most relevant, corroborating the importance of psychologists in the formal ASD diagnosis process.

The application of machine learning models, combined with feature selection techniques and hyperparameter tuning, enabled an in-depth analysis of the potential of these approaches for the early screening of beneficiaries at risk of ASD and the identification of beneficiaries with patterns for requesting special therapy. Moreover, the interpretation of the results was carried out in collaboration with healthcare professionals, which significantly contributed to the validation of the models.

This process involved a physician specializing in special therapies and autism, an occupational therapist with experience in the same area, and an analyst responsible for reviewing requests for special therapies. All three professionals work directly with these topics in their daily routines, providing technically grounded and context-aware evaluations. Their collaboration strengthened the applicability of the developed models. It highlighted the benefits that early identification of both ASD and the need for special therapies can bring, such as greater agility in the diagnostic process, better resource allocation, and, ultimately, improved quality of life for the beneficiaries.

6.1 FUTURE WORKS

The application of AI and ML models in critical sectors such as healthcare has enormous transformative potential. However, to ensure the effectiveness of these models, it is essential to adopt interpretable approaches and integrate specialists into the process. This work presented promising results using ML models with data that are not directly related to clinical diagnoses.

For future research, an essential aspect is expanding the database. Many variables considered important by specialists were discarded due to lack of representativeness or due to impossibility of collection at the time of database creation. A future update could include data from new beneficiaries who joined the plan, broadening the study's scope. Furthermore, new experiments could be carried out, such as categorizing numerical variables, which could mitigate the impact of missing values and allow the application of this database to other types of models.

Regarding binary classification models, future analyzes could explore the performance of more robust models, including deep neural networks (Deep Learning), to assess whether more complex architectures result in better pattern detection in the data. The same can be done with One-Class Classification models, which in this study were tested only in basic configurations. Exploring more advanced versions of these models could bring significant benefits, especially in identifying beneficiaries at risk of ASD.

Additionally, to address data imbalance issues, techniques such as class reweighting, data augmentation for the minority class, or even the use of models specifically designed to handle imbalance could be tested to improve the overall performance of the models.

A potential extension of this work would be the implementation of a hierarchical approach combining One-Class Classification and Binary Classification models. The One-Class model could be used to identify beneficiaries with usage patterns consistent with the risk of requesting therapies for ASD, while the binary model would refine this analysis by determining which of these beneficiaries have a high risk of ASD. These analyzes should be validated on local tests before possible implementation in the healthcare provider's system.

Another relevant advancement lies in the use of SHAP for model interpretability. In this study, SHAP was applied to consolidate the models developed and identify the most relevant variables at the global level. However, a future improvement would be the application of local interpretability, which would allow the identification of the most important variables for each specific prediction. In the operational context of the company, this feature would be highly valuable.

BIBLIOGRAPHY

AKIBA, T. et al. Optuna: A next-generation hyperparameter optimization framework. In: **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2019.

ALZUBI, J.; NAYYAR, A.; KUMAR, A. Machine learning from theory to algorithms: an overview. In: IOP PUBLISHING. **Journal of physics: conference series**. Bangalore, India, 2018. v. 1142, p. 012012.

ARAUJO, L. A. d. et al. **Manual de Orientação do Departamento Científico de Pediatria do Desenvolvimento e Comportamento**. Rio de Janeiro, 2019.

BALAKRISHNA, N. et al. Autism spectrum disorder detection using machine learning. In: IEEE. **2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)**. Greater Noida, India, 2023. p. 1645–1650.

BIAU, G.; SCORNET, E. A random forest guided tour. **Test**, Springer, v. 25, p. 197–227, 2016.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. New York, NY: Springer, 2006. v. 4.

BOSE, S.; SETH, P. Screening of autism spectrum disorder using machine learning approach in accordance with dsm-5. In: IEEE. **2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)**. Kolkata, India, 2023. p. 1–6.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

CHAPMAN, P. Crisp-dm 1.0: Step-by-step data mining guide. In: **CRISP-DM: The CRISP-DM process model and methodology**. Copenhagen, Denmark; Germany; USA; The Netherlands: NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc., OHRA Verzekeringen en Bank Groep B.V., 2000.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322.

CHEN, Y.; ZHOU, X. S.; HUANG, T. One-class svm for learning in image retrieval. In: **Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)**. Thessaloniki, Greece: IEEE., 2001. v. 1, p. 34–37 vol.1.

DAWSON, G. et al. Randomized, controlled trial of an intervention for toddlers with autism: the early start denver model. **Pediatrics**, American Academy of Pediatrics, v. 125, n. 1, p. e17–e23, 2010.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **The Journal of Machine learning research**, JMLR. org, v. 7, p. 1–30, 2006.

DHOLAKIA, N. H. et al. Autism detection using artificial neural networks: A comparative study. In: IEEE. **2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)**. India, 2024. p. 1–6.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: gradient boosting with categorical features support. **arXiv preprint arXiv:1810.11363**, 2018.

ESTES, A. et al. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. **Journal of the American Academy of Child & Adolescent Psychiatry**, Elsevier, v. 54, n. 7, p. 580–587, 2015.

FRAŇO, M. **Web Scraping as a Data Source for Machine Learning Models and the Importance of Preprocessing Web Scraped Data**. Dissertação (B.S. thesis) — University of Twente, 2024.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

GILL, K. S. et al. Utilization of naive bayes classifier for autism risk assessment using machine learning. In: IEEE. **2024 3rd International Conference for Innovation in Technology (INOCON)**. Bangalore, India, 2024. p. 1–5.

GURALNICK, M. J. Early intervention for children with intellectual disabilities: An update. **Journal of Applied Research in Intellectual Disabilities**, Wiley Online Library, v. 30, n. 2, p. 211–229, 2017.

HALLADAY, A. K. et al. Sex and gender differences in autism spectrum disorder: summarizing evidence gaps and identifying emerging areas of priority. **Molecular autism**, Springer, v. 6, p. 1–5, 2015.

HASAN, S. M. et al. A machine learning framework for early-stage detection of autism spectrum disorders. **IEEE Access**, IEEE, v. 11, p. 15038–15057, 2022.

HODGES, H.; FEALKO, C.; SOARES, N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. **Translational pediatrics**, AME Publications, v. 9, n. Suppl 1, p. S55, 2020.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International journal of data mining & knowledge management process**, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

HYDE, K. K. et al. Applications of supervised machine learning in autism spectrum disorder research: a review. **Review Journal of Autism and Developmental Disorders**, Springer, v. 6, p. 128–146, 2019.

KAVITHA, V.; SIVA, R. Classification of toddler, child, adolescent and adult for autism spectrum disorder using machine learning algorithm. In: IEEE. **2023 9th International conference on advanced computing and communication systems (ICACCS)**. Coimbatore, India, 2023. v. 1, p. 2444–2449.

KHAN, S. S.; MADDEN, M. G. One-class classification: taxonomy of study and review of techniques. **The Knowledge Engineering Review**, v. 29, n. 3, p. 345–374, 2014.

KOLYSHKINA, I.; SIMOFF, S. Interpretability of machine learning solutions in industrial decision engineering. In: SPRINGER. **Australasian Conference on Data Mining**. Singapore, 2019. p. 156–170.

KOLYSHKINA, I.; SIMOFF, S. Interpretability of machine learning solutions in public healthcare: The crisp-ml approach. **Frontiers in big data**, Frontiers Media SA, v. 4, p. 660206, 2021.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: **2008 Eighth IEEE International Conference on Data Mining**. Pisa, Italy: IEE, 2008. p. 413–422.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation-based anomaly detection. **ACM Trans. Knowl. Discov. Data**, Association for Computing Machinery, New York, NY, USA, v. 6, n. 1, mar. 2012. ISSN 1556-4681. Disponível em: <https://doi.org/10.1145/2133360.2133363>.

LORD, C. et al. Autism spectrum disorder. **Nature reviews Disease primers**, Nature Publishing Group, v. 6, n. 1, p. 1–23, 2020.

LORD, C. et al. Autism spectrum disorder. **The lancet**, Elsevier, v. 392, n. 10146, p. 508–520, 2018.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems 30**. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

MANOJ, M.; PRAVEEN, J. I. A hybrid approach to support the detection of autism spectrum disorder (asd) through machine learning and deep learning techniques. In: IEEE. **2023 12th International Conference on Advanced Computing (ICoAC)**. India, 2023. p. 1–7.

MITTAL, K. et al. Leveraging machine learning with adaboost classification to assess autism spectrum disorder (asd) probability. In: IEEE. **2024 Asia Pacific Conference on Innovation in Technology (APCIT)**. India, 2024. p. 1–4.

MOHER, D. et al. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. **International journal of surgery**, Elsevier, v. 8, n. 5, p. 336–341, 2010.

MUKHERJEE, S. B. Autism spectrum disorders—diagnosis and management. **The Indian Journal of Pediatrics**, Springer, v. 84, n. 4, p. 307–314, 2017.

NAIK, S. K. R. et al. Determination and diagnosis of autism spectrum disorder using efficient machine learning algorithm. In: IEEE. **2023 3rd International Conference on Intelligent Technologies (CONIT)**. India, 2023. p. 1–5.

NASTESKI, V. An overview of the supervised machine learning methods. **Horizons. b**, v. 4, n. 51-62, p. 56, 2017.

ONZI, F. Z.; GOMES, R. de F. Transtorno do espectro autista: a importância do diagnóstico e reabilitação. **Caderno pedagógico**, v. 12, n. 3, 2015.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. **Advances in neural information processing systems**, v. 31, 2018.

RAJ, S.; MASOOD, S. Analysis and detection of autism spectrum disorder using machine learning techniques. **Procedia Computer Science**, Elsevier, v. 167, p. 994–1004, 2020.

RAJAGOPALAN, S. S. et al. Machine learning prediction of autism spectrum disorder from a minimal set of medical and background information. **JAMA Network Open**, American Medical Association, v. 7, n. 8, p. e2429229–e2429229, 2024.

ROMERO-GARCÍA, R. et al. Q-chat-nao: A robotic approach to autism screening in toddlers. **Journal of Biomedical Informatics**, Elsevier, v. 118, p. 103797, 2021.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach (4th Edition)**. Pearson, 2020. ISBN 9780134610993. Disponível em: <http://aima.cs.berkeley.edu/>.

SARANYA, A.; ANANDAN, R. Autism spectrum prognosis using worm optimized extreme learning machine (woem) technique. In: **2021 international conference on advance computing and innovative technologies in engineering (icacite)**. India: IEEE, 2021. p. 636–641.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. **SN computer science**, Springer, v. 2, n. 3, p. 160, 2021.

SCHÖLKOPF, B. et al. Estimating the support of a high-dimensional distribution. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 13, n. 7, p. 1443–1471, 2001.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying crisp-dm process model. **Procedia Computer Science**, Elsevier, v. 181, p. 526–534, 2021.

SHARMA, S. R.; GONDA, X.; TARAZI, F. I. Autism spectrum disorder: classification, diagnosis and therapy. **Pharmacology & therapeutics**, Elsevier, v. 190, p. 91–104, 2018.

STIGLIC, G. et al. Interpretability of machine learning-based prediction models in healthcare. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 10, n. 5, p. e1379, 2020.

SUKIENNIK, R.; MARCHEZAN, J.; SCORNAVACCA, F. Challenges on diagnoses and assessments related to autism spectrum disorder in brazil: a systematic review. **Frontiers in Neurology**, Frontiers, v. 12, p. 598073, 2022.

THABTAH, F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. **Informatics for Health and Social Care**, Taylor & Francis, v. 44, n. 3, p. 278–297, 2019.

UNIMED SANTA CATARINA. **Manual de Terapias Especiais**. Santa Catarina, 2021.

VAKADKAR, K.; PURKAYASTHA, D.; KRISHNAN, D. Detection of autism spectrum disorder in children using machine learning techniques. **SN computer science**, Springer, v. 2, p. 1–9, 2021.

WASHINGTON, P.; WALL, D. P. A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. **Annual Review of Biomedical Data Science**, Annual Reviews, v. 6, p. 211–228, 2023.

WASHINGTON, P.; WALL, D. P. A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. **Annual Review of Biomedical Data Science**, Annual Reviews, v. 6, p. 211–228, 2023.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. New York: Practical Application Company, 2000. v. 1, p. 29–39.

APPENDIX A – RELEVANT VARIABLES FOR IDENTIFYING ASD LISTED

No.	Variable	Description
1	ASD (Label)	Indication of autism spectrum disorder condition
2	Patient ID	Patient identifier code
3	Sex	Patient's sex
4	Current Age	Current age of the patient
5	Maternal Age	Current age of the patient's mother
6	Maternal Age at Birth	Mother's age at the patient's birth
7	Advanced Maternal Age	Indicates whether the mother's age at birth was over 35
8	Paternal Age	Father's age at the patient's birth
9	Paternal Age at Birth	Current age of the patient's father
10	Advanced Paternal Age	Indicates whether the father's age at birth was over 35
11	Clexane	Indicates whether the mother took Clexane during pregnancy
12	Maternal Autoimmune Disease	Indicates whether the mother requested tests for autoimmune disease
13	Paternal Autoimmune Disease	Indicates whether the father requested tests for autoimmune disease
14	Previous Pregnancies	Indicates whether the mother had previous pregnancies
15	Twin Pregnancy	Indicates whether the patient was part of a multiple pregnancy
16	Premature Birth	Indicates whether the patient was born prematurely
17	Maternal Diabetes	Indicates whether the mother had diabetes during pregnancy
18	Delivery Type	Indicates the type of delivery: Cesarean or natural
19	Physiotherapist	Indicates whether the patient had consultations with a physiotherapist
20	Age at Physiotherapist	Age at the first consultation with a physiotherapist
21	Speech Therapist	Indicates whether the patient had consultations with a speech therapist

22	Age at Speech Therapist	Age at the first consultation with a speech therapist
23	Occupational Therapist	Indicates whether the patient had consultations with an occupational therapist
24	Age at Occupational Therapist	Age at the first consultation with an occupational therapist
25	Psychologist	Indicates whether the patient had consultations with a psychologist
26	Age at Psychologist	Age at the first consultation with a psychologist
27	Neurologist	Indicates whether the patient had consultations with a neurologist
28	Age at Neurologist	Age at the first consultation with a neurologist
29	Pediatrician	Indicates whether the patient had consultations with a pediatrician
30	Age at Pediatrician	Age at the first consultation with a pediatrician
31	Ophthalmologist	Indicates whether the patient had consultations with an ophthalmologist
32	Age at Ophthalmologist	Age at the first consultation with an ophthalmologist
33	Critical Ophthalmologist	Indicates whether the patient was younger than 3 years old at the first ophthalmologist consultation
34	Electroencephalogram	Indicates whether the patient requested an electroencephalogram exam
35	Age at Electroencephalogram	Age at the first electroencephalogram exam request
36	Transfontanelle Ultrasound	Indicates whether the patient requested a transfontanelle ultrasound exam
37	Age at Transfontanelle Ultrasound	Age at the first transfontanelle ultrasound exam request
38	Cranial MRI	Indicates whether the patient requested a cranial MRI exam
39	Age at Cranial MRI	Age at the first cranial MRI exam request
40	Critical Cranial MRI	Indicates whether the patient was younger than 3 years old at the first cranial MRI exam request
41	CT Scan	Indicates whether the patient requested a CT scan exam
42	Age at CT Scan	Age at the first CT scan exam request

43	Critical CT Scan	Indicates whether the patient was younger than 3 years old at the first CT scan exam request
44	BERA	Indicates whether the patient requested a BERA exam
45	Age at BERA	Age at the first BERA exam request
46	Critical BERA	Indicates whether the patient was younger than 3 years old at the first BERA exam request
47	Genetic Testing	Indicates whether the patient requested genetic testing
48	Age at Genetic Testing	Age at the first genetic testing request
49	VDRL	Indicates whether the patient requested a VDRL test
50	Age at VDRL	Age at the first VDRL test request
51	Critical VDRL	Indicates whether the patient was younger than 3 years old at the first VDRL test request
52	Cytomegalovirus	Indicates whether the patient requested a cytomegalovirus test
53	Age at Cytomegalovirus	Age at the first cytomegalovirus test request
54	Critical Cytomegalovirus	Indicates whether the patient was younger than 3 years old at the first cytomegalovirus test request
55	Toxoplasmosis	Indicates whether the patient requested a toxoplasmosis test
56	Age at Toxoplasmosis	Age at the first toxoplasmosis test request
57	Critical Toxoplasmosis	Indicates whether the patient was younger than 3 years old at the first toxoplasmosis test request
58	Special Therapy	Indicates whether the patient requested special therapy consultations
59	Age at Special Therapy	Age at the first special therapy consultation request
60	Maternal Special Therapy	Indicates whether the mother requested special therapy consultations
61	Paternal Special Therapy	Indicates whether the father requested special therapy consultations
62	Autism Level	Classifies the patient’s level of autism spectrum disorder
63	Jaundice	Indicates whether the patient requested a jaundice test
64	Birth Complications	Indicates whether the mother experienced complications during delivery

65	Gestational Obesity	Indicates whether the mother experienced obesity during pregnancy
66	Hypertension During Pregnancy	Indicates whether the mother had hypertension during pregnancy
67	Prenatal Care	Indicates whether the mother received prenatal care
68	Birth Weight	The patient’s birth weight

ANNEX A – M-CHAT SCALE

Table 10 – M-CHAT Questionnaire Variables Source: WHO

No.	Question	Yes	No
1	If you point at an object in the room, does your child look at it? (FOR EXAMPLE, if you point at a toy or animal, does your child look at the toy or animal?)	Yes	No
2	Have you ever wondered if your child might be deaf?	Yes	No
3	Does your child engage in pretend play? (FOR EXAMPLE, pretend to drink from an empty cup, pretend to talk on the phone, or pretend to feed a doll or stuffed animal?)	Yes	No
4	Does your child enjoy climbing on things? (FOR EXAMPLE, furniture, playground equipment, or stairs)	Yes	No
5	Does your child make unusual finger movements near their eyes? (FOR EXAMPLE, wiggling fingers in front of their eyes and looking at them?)	Yes	No
6	Does your child point with their finger to ask for something or to get help? (FOR EXAMPLE, pointing at a cookie or toy out of reach?)	Yes	No
7	Does your child point with their finger to show you something interesting? (FOR EXAMPLE, pointing at an airplane in the sky or a big truck on the street)	Yes	No
8	Is your child interested in other children? (FOR EXAMPLE, does your child look at other children, smile at them, or approach them?)	Yes	No

9	Does your child bring things to show you or hold them up so you can see them—not to get help but just to share? (FOR EXAMPLE, showing a flower, a stuffed animal, or a toy truck)	Yes	No
10	Does your child respond when you call their name? (FOR EXAMPLE, do they look at you, speak or make a sound, or stop what they are doing when you call their name?)	Yes	No
11	When you smile at your child, do they smile back at you?	Yes	No
12	Does your child get very upset by everyday noises? (FOR EXAMPLE, does your child scream or cry when hearing sounds like a blender or loud music?)	Yes	No
13	Does your child walk?	Yes	No
14	Does your child look you in the eyes when you are talking to, playing with, or dressing them?	Yes	No
15	Does your child try to imitate what you do? (FOR EXAMPLE, waving goodbye, clapping hands, or blowing kisses)	Yes	No
16	When you turn your head to look at something, does your child look around to see what you are looking at?	Yes	No
17	Does your child try to get you to look at them? (FOR EXAMPLE, does your child look at you for praise, or say “look, mom!” or “oh, mom!”)	Yes	No
18	Does your child understand when you tell them to do something? (FOR EXAMPLE, if you don’t point, does your child understand when you say, “put the cup on the table” or “turn on the TV”?)	Yes	No

19	When something new happens, does your child look at your face to see how you feel about what happened? (FOR EXAMPLE, if they hear a strange noise, see something funny, or see a new toy, do they look at your face?)	Yes	No
20	Does your child enjoy movement activities? (FOR EXAMPLE, being swung or bouncing on your knees)	Yes	No

ANNEX B – AQ-10 DATABASE

Table 11 – AQ-10 Database Variables Source: (KAVITHA; SIVA, 2023)

Attr ID	**Features of Q-chat-10-Toddler**
AID1	Age: Age was determined using years
AID2	Gender: Male or Female
AID3	Nationality: Text-based list of all countries
AID4	Family member with PDD: Whether or not any family members had PDD
AID5	Location of residence: List of countries where the user lives
AID6	Born with jaundice: Whether the patient suffered from jaundice by birth
AID7	Screening application: Whether or not a user is using a screening application
AID8	Test completion: Who has completed the experiment
AID9	Screening test type: Emotion, communication, behaviour and social skills
AID10	Screening Score: Grade received based on screening technique scoring mechanism
AID21	Class/ASD: Classifying ASD or Non-ASD
AID11	When you call your child's name, does he/she look at you?
AID12	Does your child make eye contact easily with you?
AID13	Is your child pointing to express a desire? (For instance, an out-of-reach toy)
AID14	Does your child indicate that they share your interests?
AID15	Does your kid act out? (For instance, tend to dolls or use a toy phone)
AID16	Do they follow where you gaze when you look?
AID17	Does your youngster express an interest in comforting others when they see you or another family member clearly upset? (Examples include hugging them and petting their hair)

AID18	How would you describe the initial words of your child?
AID19	What kind of hand motions does your kid make?
AID20	Does your toddler seem to be staring at nothing but nothing?

ANNEX C – SPARK + SSC DATABASE

Table 12 – List of predictors from the SPARK Source: (RAJAGOPALAN et al., 2024)

Item No.	Predictor Variable	Description	Category
1	sex	Sex assigned at birth	basic medical screening
2	gest age	How many weeks (gestational age) was child/dependent when he/she was born?	basic medical screening
3	eating probs	Problems with eating foods - not diagnosed by a professional	basic medical screening
4	feeding dx	Feeding/eating problems	basic medical screening
5	med cond birth	Birth or pregnancy complications	basic medical screening
6	birth oth calc	Self-reported birth complications not represented in the current coding	basic medical screening
7	med cond birth def	Birth defects	basic medical screening
8	med cond growth	Growth conditions (including height, weight, and head size)	basic medical screening
9	growth oth calc	Growth conditions not represented in the current coding	basic medical screening
10	med cond neuro	Neurological conditions	basic medical screening
11	med cond visaud	Vision or hearing conditions	basic medical screening
12	mother highest education	Highest level of education of mother/guardian	background history
13	father highest education	Highest level of education of father/guardian	background history
14	annual household income	Annual household income	background history
15	smiled age mos	Age in months when first smiled	background history

16	sat wo support age mos	Age in months when first sat without support	background history
17	crawled age mos	Age in months when first crawled	background history
18	walked age mos	Age in months when first walked alone	background history
19	fed self spoon age mos	Age in months when first fed self with a spoon	background history
20	used words age mos	Age in months when first used single words	background history
21	combined words age mos	Age in months when first combined words into short phrases or sentences with an action word	background history
22	combined phrases age mos	Age in months when first combined phrases into longer sentences	background history
23	bladder trained age mos	Age in months when first was bladder-trained (daytime)	background history
24	bowel trained age mos	Age in months when first was bowel-trained	background history
25	hand	Which hand does the child/dependent prefer to use?	background history
26	twin mult birth	Is the child/dependent a twin or part of a multiple birth?	background history
27	num asd parents	Do either of the biological parents have a professional diagnosis of ASD?	background history
28	num asd siblings	Do you have any full biological siblings who have a professional diagnosis of ASD? If so, how many?	background history