UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Using machine learning to manage applied behavior analysis objectives in individualized education plans for children with autism spectrum disorder

Bruno Teixeira Botelho

Trabalho de Conclusão de Curso

MBA em Inteligência Artificial e Big Data

ICMC USP

SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Using machine learning to manage applied behavior analysis objectives in individualized education plans for children with autism spectrum disorder

Bruno Teixeira Botelho

USP - São Carlos

2025

Bruno Teixeira Botelho

Using machine learning to manage applied behavior analysis objectives in individualized education plans for children with autism spectrum disorder

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial.

Orientadora: Profa. Dra. Mariana Curi.

USP - São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

```
        Botelho, Bruno Teixeira
B748u      Using machine learning to manage applied behavior
        analysis objectives in individualized education
        plans for children with autism spectrum disorder /
        Bruno Teixeira Botelho; orientadora Mariana Curi. --
        São Carlos, 2025.
           100 p.


           Trabalho de conclusão de curso (MBA em
        Inteligência Artificial e Big Data) -- Instituto de
        Ciências Matemáticas e de Computação, Universidade
        de São Paulo, 2025.


           1. INTELIGÊNCIA ARTIFICIAL. 2. APRENDIZADO
        COMPUTACIONAL. 3. TRANSTORNO AUTÍSTICO. 4. ANÁLISE
        DO COMPORTAMENTO. 5. ANÁLISE DE SÉRIES TEMPORAIS. I.
        Curi, Mariana, orient. II. Título.
```

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATION

Ao meu pai, que não se encontra mais neste plano, mas que, de onde quer que esteja, certamente vibra por mais esta conquista. A ele, que tanto acreditava na educação como fator de mudança, crescimento e oportunidade de mobilidade social.

To my father, who is no longer with us, but who I know is cheering for this achievement from wherever he is. He so deeply believed in education as a force for change, growth, and social mobility.

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my family for their unconditional love, encouragement, and support, whether near or far. Your support has been a constant foundation throughout my academic journey. I dedicate a special thought to my father, who remains an inspiration for instilling in me the values of hard work and the importance of education.

My sincere thanks go to my advisor for her invaluable guidance and insightful feedback.

I also wish to thank the Universidade de São Paulo (USP) for again believing in me and providing the scholarship that made this MBA possible. To return to this esteemed institution has been a privilege and a great honor.

This research would not have been possible without the collaboration of the partner company. I am deeply grateful for the access to their data and, most importantly, for the time and support from their entire team. The people I had the pleasure to interact with were incredibly supportive, and their willingness to share their knowledge was essential to this project.

I also appreciate the dedicated psychologists and therapists whose diligent work conducting the ABA sessions generated the invaluable data forming this thesis's basis. Most importantly, I wish to acknowledge the children with autism and their families who participated in these therapy sessions. Their experiences are the heart of this research, and their journeys are the motivation behind this work. It is for them that this work is ultimately intended, and I hope that it contributes, in some small way, to improving their outcomes.

Finally, to my friends and all those with whom I could not spend as much time as I would have liked, thank you for your patience and understanding as I pursued this MBA alongside another Master's degree. Your support, even from a distance, was deeply appreciated.

EPIGRAPH

If you've met one person with autism, you've met one person with autism.

Stephen Shore (2003)

ABSTRACT

BOTELHO, B. T. **Using machine learning to manage applied behavior analysis objectives in individualized education plans for children with autism spectrum disorder**. 2025. 99 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

This study addresses the clinical challenge of managing and adapting Individualized Education Plans (IEPs) for children with Autism Spectrum Disorder (ASD), a process reliant on time-intensive manual review. The objective was to investigate the feasibility of applying machine learning (ML) to predict the status progression ('Validated', 'Completed', or 'Rejected') of Applied Behavior Analysis (ABA) learning objectives, using longitudinal therapy data to create a foundation for an effective clinical decision-support tool. The methodology involved a longitudinal clinical dataset from a Brazilian healthcare provider, comprising 3,344 learning objectives from 438 children. A feature engineering process transformed the raw time series of progress scores into a set of 24 descriptive features, capturing the temporal dynamics of learning. A pre-trained Transformer-based model, Tabular Prior-Data Fitted Network (TabPFN), and a gradient-boosting model, XGBoost, were implemented and benchmarked against baseline algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and a sequential Long Short-Term Memory (LSTM) network. The results demonstrated that TabPFN and XGBoost achieved superior performance with high overall accuracy (0.82 and 0.81, respectively), a weighted F1-score of 0.81, and weighted One-vs-Rest AUC-ROC scores of approximately 0.90. Both models effectively classified the 'Validated' (F1-score: 0.85) and 'Completed' (F1-score: ~0.80) classes, but struggled with the highly imbalanced 'Rejected' class. Feature importance analysis consistently identified the duration of the intervention and engineered features quantifying recent performance as the most influential predictors. The study concludes that ML models can effectively support the dynamic management of ABA IEPs by learning clinically relevant patterns from therapy data. This work provides a proof-of-concept for a clinical decision-support tool designed to augment, not replace, clinical expertise, paving the way for more scalable, responsive, and personalized interventions in ASD care and allowing therapists to focus on more complex clinical reasoning.

Keywords: Machine Learning. Autism Spectrum Disorder. Applied Behavior Analysis. Individualized Education Plan. Time-Series Classification. Clinical Decision Support.

RESUMO

BOTELHO, B. T. **Uso de aprendizagem de máquina para gerenciar objetivos de análise do comportamento aplicada em planos de ensino individualizados para crianças com transtorno do espectro autista**. 2025. 99 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Este estudo aborda o desafio clínico do gerenciamento e adaptação de Planos de Ensino Individualizados (PEIs) para crianças com Transtorno do Espectro Autista (TEA), um processo que depende de revisão manual intensiva. O objetivo foi investigar a viabilidade da aplicação de aprendizagem de máquina (ML) para prever a progressão de status ('Validado', 'Concluído' ou 'Rejeitado') dos objetivos de aprendizagem da Análise do Comportamento Aplicada (ABA), utilizando dados longitudinais de terapia para criar a base para uma ferramenta eficaz de apoio à decisão clínica. A metodologia envolveu um conjunto de dados clínicos longitudinais de um provedor de saúde brasileiro, compreendendo 3.344 objetivos de aprendizagem de 438 crianças. Um processo de engenharia de características transformou as séries temporais brutas de pontuações de progresso em um conjunto de 24 características descritivas, capturando a dinâmica temporal da aprendizagem. Um modelo pré-treinado baseado em Transformer, a Tabular Prior-Data Fitted Network (TabPFN), e um modelo de gradient-boosting, XGBoost, foram implementados e comparados com algoritmos de base, incluindo Máquinas de Vetores de Suporte (SVM), K-Nearest Neighbors (KNN) e uma rede Long Short-Term Memory (LSTM). Os resultados demonstraram que o TabPFN e o XGBoost alcançaram desempenho superior com alta acurácia geral (0,82 e 0,81, respectivamente), F1-score ponderado de 0,81 e pontuações AUC-ROC One-vs-Rest ponderadas de aproximadamente 0,90. Ambos os modelos classificaram eficazmente as classes 'Validado' (F1-score: 0,85) e 'Concluído' (F1-score: ~0,80), mas tiveram dificuldades com a classe 'Rejeitado', que era altamente desbalanceada. A análise de importância das características identificou consistentemente a duração da intervenção e as características de engenharia que quantificam o desempenho recente como os preditores mais influentes. O estudo conclui que os modelos de ML podem apoiar eficazmente o gerenciamento dinâmico dos PEIs de ABA, aprendendo padrões clinicamente relevantes a partir de dados de terapia. Este trabalho fornece uma prova de conceito para uma ferramenta de apoio à decisão clínica projetada para aumentar, e não substituir, a perícia clínica, abrindo caminho para intervenções mais escaláveis, responsivas e personalizadas no cuidado de pessoas com TEA e permitindo que os terapeutas se concentrem em raciocínios clínicos mais complexos.

Palavras-chave: Aprendizagem de Máquina. Transtorno do Espectro Autista. Análise do Comportamento Aplicada. Plano de Ensino Individualizado. Classificação de Séries Temporais. Apoio à Decisão Clínica.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABA	-	Applied Behavior Analysis
ABC	-	Adaptive Behavior Composite
ADASYN	-	Adaptive Synthetic Sampling
AI	-	Artificial Intelligence
ASD	-	Autism Spectrum Disorder
AUC	-	Area Under the Curve
AUC-ROC	-	Area Under the Receiver Operating Characteristic Curve
DTW	-	Dynamic Time Warping
ENN	-	Edited Nearest Neighbors
FN	-	False Negative
FP	-	False Positive
ICL	-	In-Context Learning
IEP	-	Individualized Education Plan
LSTM	-	Long Short-Term Memory
ML	-	Machine Learning
OvO	-	One-vs-One
OvR	-	One-vs-Rest
RF	-	Random Forest
RNN	-	Recurrent Neural Network
ROC	-	Receiver Operating Characteristic
SHAP	-	SHapley Additive exPlanations
SMOTE	-	Synthetic Minority Over-sampling Technique
SVM	-	Support Vector Machines
TabPFN	-	Tabular Prior-Data Fitted Network
TN	-	True Negative
TP	-	True Positive
TSC	-	Time Series Classification
Vineland-3	-	Vineland Adaptive Behavior Scales Third Edition
XGBoost	-	Extreme Gradient Boosting

INDEX

1 INTRODUCTION

1.1 Background

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by deficits in social communication and interaction alongside restricted, repetitive patterns of behavior, interests, or activities (American Psychiatric Association, 2013). The prevalence of ASD represents a significant public health consideration globally. In the United States, recent estimates suggest that approximately 1 in 36 children are diagnosed with ASD (Maenner, 2023). In Brazil, estimates suggest a substantial population living with ASD, highlighting the need for adequate support systems (Paiva Junior, 2023). The presentation of ASD varies significantly among individuals, encompassing a broad spectrum of abilities and support needs (F. Volkmar et al., 2014). Early (Dawson et al., 2010; F. Volkmar et al., 2014) and individualized intervention is crucial for improving the long-term outcomes and quality of life of individuals with ASD (Council of Autism Service Providers, 2024; Roane et al., 2016).

Interventions for children with ASD are often intensive and multidisciplinary, frequently involving Applied Behavior Analysis (ABA), speech therapy, and occupational therapy (Ghafghazi et al., 2021; F. Volkmar et al., 2014). ABA is a commonly implemented, evidence-based approach focused on understanding and changing behavior through different forms of reinforcement that can help eliminate or encourage certain behaviors (Duarte, 2018; Morris et al., 2005; Myers et al., 2007). A central component of managing these interventions is the Individualized Education Plan (IEP). The IEP outlines specific, measurable goals tailored to the child's unique needs and serves as a roadmap for therapeutic intervention and progress monitoring. Effective management and timely adaptation of these IEPs are essential for maximizing therapeutic gains (National Research Council, 2001).

This research was conducted with a Brazilian healthcare provider specializing in caring for children with ASD. This partner company employs a multidisciplinary approach, utilizing IEPs to guide therapy, and collects data systematically during treatment sessions, including progress towards specific objectives (items) outlined in the Vineland Adaptive Behavior Scales, an individually administered instrument used for measuring adaptive behavior (Sparrow et al., 2016) and other assessments. This context provides an opportunity to explore data-driven approaches to enhance IEP management.

1.2 Problem statement

Frequent adjustments to the IEP, guided by ongoing assessment and a deep understanding of the individual child and their family, are essential for maximizing positive outcomes (National Research Council, 2001). However, the effective management and adaptation of IEPs for children with ASD present significant challenges. Each child exhibits unique strengths and needs that evolve, requiring frequent adjustments to their intervention plan (National Research Council, 2001). The partner company's current process for reviewing and adapting IEPs relies heavily on the manual analysis of session data, therapist notes, assessment results (such as the Vineland), and clinical judgment by senior therapists.

While clinically grounded, this manual process is inherently time-consuming and resource-intensive, as interviews with senior therapists confirmed. It can limit the frequency of IEP reviews and adaptations, potentially delaying necessary adjustments for optimizing a child's progress. Furthermore, the scalability of this manual approach is constrained, particularly as the number of children receiving services grows. Relying solely on human analysis may also risk overlooking subtle patterns in a child's progress data or failing to identify predictive factors that could inform more precise and timely interventions (Ghafghazi et al., 2021). A specific challenge within this process is determining when a learning objective has been mastered (status changes to 'Completed'), when it requires continued focus (status remains 'Valid'), or when it should be discontinued for other reasons, such as the child not showing progress in the objective (status changes to 'Rejected'). This three-way decision is critical for efficient progression through the IEP, but often involves subjective clinical judgment. The lack of a standardized approach to guide these status updates can lead to inefficiencies and potentially suboptimal intervention pacing.

1.3 Justification and motivation

The challenges associated with the manual review and adaptation of IEPs underscore the need for more efficient, scalable, and data-driven solutions. The increasing prevalence of ASD necessitates approaches that can support clinicians in delivering high-quality, personalized care to a growing number of children (Grosvenor et al., 2024). Artificial intelligence (AI) and machine learning (ML) offer potential tools to address these challenges (Rêgo & Araújo-Filho, 2024; Thabtah, 2019).

By leveraging ML techniques to analyze longitudinal therapy data, it may be possible to identify patterns associated with objective mastery and predict future status progression. Such a system could serve as a decision-support tool for senior therapists, augmenting their clinical expertise with data-driven insights. This could lead to several benefits:

- Increased efficiency: Automating parts of the data analysis could free up valuable therapist time, allowing them to focus more on direct clinical care and complex decision-making (Garikipati et al., 2023; Ghafghazi et al., 2021).
- Enhanced personalization: Identifying predictive patterns could enable more timely and precise adjustments to IEPs, better tailoring interventions to the individual child's learning trajectory (Schwartz et al., 2021).
- Improved scalability: An AI-assisted system could help manage the complexities of IEP adaptation for many children, supporting the scalability of intervention services.
- Potential for improved outcomes: By facilitating more responsive and individualized intervention planning, ML tools may contribute to better developmental outcomes for children with ASD (Kohli et al., 2022).

This research is motivated by the potential for ML to significantly enhance the management of interventions for children with ASD within the partner company's framework and potentially beyond. It represents an opportunity to apply ML techniques to a meaningful real-world problem, contributing to more effective and efficient healthcare delivery for a vulnerable population. The availability of structured, longitudinal data from the partner company provides a foundation for developing and evaluating such an ML-based approach.

1.4 Research question and objectives

This project seeks to answer the following research question: How can ML be applied to assist in managing ABA IEPs for children with ASD, specifically in predicting the status progression ('Valid', 'Completed', or 'Rejected') of learning objectives based on longitudinal therapy data?

To address this question, the general objective of this project is to develop and evaluate ML models capable of predicting the status progression of objectives within the IEPs for children with ASD, using longitudinal therapy data provided by a partner healthcare company.

The specific objectives are:

1. Preprocess and structure the longitudinal therapy data, including objective progress scores and relevant contextual features, into a format suitable for ML analysis.
2. Train one of the most recent and well-performing ML algorithms (TabPFN) to classify the status progression of IEP objectives over time.
3. Explore and train other ML models using well-known algorithms (SVM, KNN, XGBoost, LSTM) to be used as a baseline for comparison.
4. Evaluate and compare the performance of these models (Accuracy, Precision, Recall, F1-score, AUC-ROC).
5. Evaluate the features' importance (SelectKBest, Random Forest, XGBoost feature importance, SHAP values).

1.5 Structure of the project

This project is organized into six chapters. Chapter 1 introduces the background context of ASD intervention and IEP management, outlines the problem statement and justification, and presents the research question and objectives. Chapter 2 reviews the relevant literature on ASD, ABA, IEPs, Vineland, the state-of-the-art usage of ML for ASD management, ML techniques for Time-Series Classification, feature selection methods, strategies for handling imbalanced data, and metrics for model evaluation. Chapter 3 details the methodology employed, including data acquisition from the partner company, data preprocessing steps, feature engineering, the specific ML models implemented, and the evaluation metrics used. Chapter 4 presents the results of the data analysis and model evaluations. Chapter 5 discusses interpreting these results and their implications for clinical practice, academic and practical contributions, the study's limitations, and potential avenues for future research. Finally, Chapter 6 concludes the project by summarizing the key findings, contributions, implications, limitations, and future research.

2 LITERATURE REVIEW

This chapter provides a review of the literature relevant to the research topic. Section 2.1 will briefly cover Autism Spectrum Disorder (ASD), Applied Behavior Analysis (ABA), and Individualized Education Plan (IEP), including the Vineland assessment. Section 2.2 focuses on applying machine learning (ML) in ASD management. Section 2.3 synthesizes the literature reviewed in previous sections and identifies the research gap this work aims to address. Section 2.4 details relevant ML techniques, and subsequent sections cover feature selection, data imbalance, and model evaluation metrics.

2.1 Autism Spectrum Disorder

ASD is a neurodevelopmental disorder characterized by persistent deficits in social communication and social interaction across multiple contexts alongside restricted, repetitive behavior patterns, interests, or activities. These core characteristics manifest in difficulties with social-emotional reciprocity, challenges in developing, maintaining, and understanding relationships, and deficits in nonverbal communicative behaviors used for social interaction. Additionally, individuals with ASD exhibit restricted, repetitive patterns of behavior, interests, or activities, which may include stereotyped motor movements or speech, insistence on sameness or inflexible adherence to routines, highly restricted, fixated interests that are abnormal in intensity or focus, or hyper- or hyporeactivity to sensory input (American Psychiatric Association, 2013; F. Volkmar et al., 2014).

ASD typically manifests in the early developmental period (usually before age 3). However, symptoms may not become fully apparent until social demands exceed limited capacities, or learned strategies in later life may mask them (American Psychiatric Association, 2013). The term 'spectrum' underscores the significant heterogeneity observed in symptom presentation, severity, and associated features, including intellectual ability and language skills, which range from nonverbal individuals to those with fluent speech (Lord et al., 2000). This variability extends to cognitive profiles, with some individuals exhibiting intellectual disability while others possess average or above-average intelligence (National Research Council, 2001).

Because ASD begins early in life and affects many parts of development, it is important to identify and intervene early. Evidence consistently indicates timely, appropriate, and often intensive interventions can significantly improve developmental trajectories, adaptive functioning, and long-term outcomes for individuals with ASD (Dawson et al., 2010; F. Volkmar et al., 2014). The inherent variability and developmental nature of ASD necessitate highly individualized approaches to treatment and support. Intervention plans must be dynamic and adaptable to accommodate each child's unique and evolving strengths, challenges, and learning styles (National Research Council, 2001).

2.1.1 Applied Behavior Analysis

ABA is a scientific field that uses learning principles to change behavior that is important to society. (Cooper et al., 2019). ABA is widely recognized as an evidence-based intervention for individuals with ASD and is considered a gold-standard treatment approach by many practitioners (Garikipati et al., 2023; Howard et al., 2005). ABA interventions aim to increase helpful behaviors and decrease behaviors that are harmful or affect learning by breaking down complex skills into smaller, teachable steps and utilizing reinforcement strategies (Haring & Kennedy, 1988; National Research Council, 2001). Interventions are typically individualized, data-driven, and focus on observable behaviors and environmental factors (Howard et al., 2005; Myers et al., 2007; National Research Council, 2001).

2.1.2 Individualized Education Plan

IEPs are fundamental tools in structuring interventions for children with developmental disabilities, including ASD. An IEP is a written document outlining objectives tailored to the individual child's needs. These objectives should be observable, measurable behaviors and skills, accomplished within 1 year, and expected to affect a child’s participation in education, community, and family life goals. It is a collaborative plan developed by therapists, educators, and parents, guiding the therapeutic process and providing a framework for monitoring progress. Key components typically include current levels of performance, annual goals, specific intervention strategies, necessary accommodations, and methods for evaluating progress. The dynamic nature of ASD necessitates that IEPs are not static documents; they require regular review and adaptation based on the child's evolving skills and challenges to remain effective. Determining when objectives are met or need modification is crucial for treatment efficacy (National Research Council, 2001).

2.1.3 Assessment in ASD: The Vineland Adaptive Behavior Scales

Assessment is integral to developing and adapting IEPs. Various tools are used to evaluate different domains of functioning in children with ASD. The Vineland Adaptive Behavior Scales Third Edition (Vineland-3) is a widely used standardized assessment tool that measures adaptive behavior – the practical, everyday skills an individual uses to function and meet environmental demands in daily life, rather than just their potential ability (Sparrow et al., 2016).

The Vineland-3 assesses adaptive behavior across core domains: Communication, Daily Living Skills, and Socialization. A Motor Skills domain is also included for younger individuals (typically under age 9). Additionally, it often contains scales assessing Maladaptive Behavior (Sparrow et al., 2016).

These broad domains are further broken down into specific subdomains. For instance:

- The Communication domain includes subdomains like Receptive (understanding), Expressive (speaking), and Written.
- Daily Living Skills covers Personal (self-care), Domestic (household tasks), and Community (functioning in public) skills.
- Socialization is assessed through subdomains such as Interpersonal Relationships, Play and Leisure, and Coping Skills.
- Motor Skills differentiate between Gross and Fine motor abilities.

Measurement is typically conducted via a semi-structured interview or questionnaire completed by a parent, caregiver, or teacher familiar with the individual. Specific behavioral items within each subdomain are rated based on the frequency with which the individual performs the behavior, often using a scale like 2 (usually performs), 1 (sometimes performs), or 0 (never performs). For instance, the Communication Domain includes items like "Says at least 50 words" or "Use phrases with a noun and a verb (for example, "Mommy stay," "Give a ball")" (Sparrow et al., 2016).

Raw scores from these items' ratings are then converted into norm-referenced scores. The primary scores include standard scores for each domain and an overall Adaptive Behavior Composite (ABC) score. These standard scores typically have a mean of 100 and a standard deviation of 15, allowing comparison to a normative sample of same-aged peers. Subdomains are often reported using v-scale scores, which typically have a mean of 15 and a standard deviation of 3 (Sparrow et al., 2016).

The Vineland-3 offers valuable information for identifying specific strengths and weaknesses by providing detailed scores across domains and subdomains. This aids in setting appropriate IEP goals, planning interventions, and quantitatively monitoring a child's progress in adaptive functioning over time (F. R. Volkmar et al., 1993). Data from tools like the Vineland, collected longitudinally, can provide quantitative insights into a child's development throughout an intervention.

2.2 Machine learning in ASD management

Technology has increasingly been explored as a means to support individuals with ASD. Applications range from diagnostic aids and screening tools to therapeutic interventions and communication supports (Abbas et al., 2020; Du et al., 2019; Hyde et al., 2019). Robots, virtual reality, and mobile applications have been developed to teach social skills, manage anxiety, and provide structured learning environments (Du et al., 2019; Nie et al., 2021; Rêgo & Araújo-Filho, 2024; Santos et al., 2021).

ML has shown promise in analyzing complex ASD-related data and is increasingly being applied to improve diagnosis, treatment, and support for individuals with ASD. The complexity and heterogeneity of ASD and the need for highly individualized and intensive interventions present significant challenges for diagnosis, treatment planning, and ongoing management (Rêgo & Araújo-Filho, 2024). This heterogeneity of ASD, the complexity of behavioral data, and the need for highly personalized interventions make it a suitable area for ML applications, which excel at identifying patterns in large and complex datasets (Shatte et al., 2019). In recent years, ML and artificial intelligence (AI) techniques have emerged as promising tools to address these challenges, offering potential improvements in efficiency, accuracy, personalization, and care accessibility (Pandya et al., 2024). This section reviews key application areas of ML in ASD management.

Early and accurate diagnosis is crucial for timely intervention in ASD (Dawson & Bernier, 2013). Traditional diagnostic processes can be lengthy and rely heavily on behavioral observation (Abbas et al., 2020). ML offers methods to automate, expedite, and improve the objectivity of screening and diagnosis (Kumar & Das, 2022; Rajagopalan et al., 2024; Shinde & Patil, 2023), to identify biomarkers (Zhou et al., 2014) or to understand etiology using genetic data (Kou et al., 2012). Researchers have applied various ML algorithms, including Support Vector Machines (SVM), Random Forests (RF), K-Nearest Neighbors (KNN), and deep learning models (e.g., CNNs, RNNs), to diverse data types. The data sources include behavioral questionnaires, clinical records, neuroimaging data (fMRI, DTI), electroencephalography (EEG) signals, and eye-tracking data. (Hyde et al., 2019; Shatte et al., 2019)

A growing body of work is exploring ML for treatment and its management. Researchers have recognized the potential for ML to improve the efficiency and effectiveness of interventions like ABA. Several studies have demonstrated using ML to analyze behavioral data relevant to ASD management. For instance, ML models have been developed to suggest optimal treatment and predict treatment outcomes or responsiveness. ML models have been used to predict treatment response (Linstead et al., 2015, 2017). Schwartz et al. (2021) integrated ML and statistical algorithms to recommend optimal therapy types (Cognitive Behavioral Therapy vs. Psychodynamic Therapy) based on patient characteristics, showcasing the potential for ML-driven treatment selection. Closer to the ABA context, Maharjan et al. (2023) developed an ML model to classify the appropriate ABA treatment plan type (comprehensive vs. focused) based on patient intake data, aiming to standardize the initial decision-making process.

Furthermore, ML has been applied to personalize treatment goals within ABA frameworks. Kohli et al. (2022) used patient similarity and collaborative filtering methods to recommend specific treatment goals (domains and targets) for individuals with ASD undergoing ABA, demonstrating that ML can potentially augment clinical judgment in tailoring interventions. Other research has focused on using AI and sensor data to monitor behaviors relevant to ASD, such as stereotyped movements or emotional states, which could feed into adaptive intervention systems (Ghafghazi et al., 2021).

2.2.1 Key predictors in machine learning models for ASD treatment

A goal of applying ML in ASD research is to identify the most influential features for diagnosis, prognosis, and treatment response. The literature reveals various predictive factors, ranging from treatment parameters to child-specific characteristics.

Treatment-related features, particularly the intensity and duration of the intervention, are consistently identified as strong predictors of outcomes. Studies using linear models and neural networks have found a robust relationship between the number of therapy hours and the mastery of learning objectives (Linstead et al., 2015, 2017). This indicates that a higher treatment dosage is generally associated with greater progress. Maharjan et al. (2023)

reinforced this finding, identifying the hours per week of past ABA treatment as one of the most important features for determining the required intensity of a future treatment plan.

Child-specific characteristics are also important. Age at the start of the intervention is a frequently cited factor, with earlier intervention generally leading to better outcomes (Dawson & Bernier, 2013). Baseline skill levels, including pretreatment IQ, adaptive behavior, and language skills, have also been shown to predict gains in adaptive behavior and treatment response (American Psychiatric Association, 2013; National Research Council, 2001). Furthermore, specific skills, such as bathing ability, have been identified as highly predictive features for determining the necessary intensity of ABA therapy, likely as a proxy for a child's broader adaptive functioning levels (Maharjan et al., 2023).

In addition to these clinical and demographic variables, ML models have successfully utilized data from various sources. Features extracted from neuroimaging data, such as functional connectivity from fMRI (functional magnetic resonance imaging), have been used to classify individuals and predict ASD (Zhou et al., 2014). Similarly, behavioral data from eye-tracking, analysis of facial expressions, and motor skills have been identified as valuable markers (Pandya et al., 2024). While the features vary depending on the specific research question, the duration of intervention and the child's baseline functional abilities consistently emerge as key predictors of progress in ABA therapy.

2.3 Synthesis and research gap

The literature highlights the importance of individualized, data-driven approaches in ASD intervention (F. Volkmar et al., 2014) and the growing potential of AI and ML in healthcare, including ASD (Hyde et al., 2019; Rêgo & Araújo-Filho, 2024; Thabtah, 2019). The applications highlight the potential for ML to extract meaningful patterns from clinical data and support more personalized and data-driven decision-making in ASD care. However, challenges remain in the application of ML to ASD research, including issues with diagnostic coding, feature selection, and data imbalances (Thabtah, 2019). Moreover, despite these advancements, the application of ML to the dynamic management and adaptation of IEPs based on longitudinal therapy progress data remains a relatively underexplored area. Most existing studies focus on initial diagnosis, overall outcome prediction, or recommending initial treatment paths rather than the ongoing, iterative process of evaluating and adjusting specific learning objectives within an IEP based on mastery progression. The challenge lies in analyzing the time-series nature of therapy data (e.g., daily progress scores on specific objectives) to identify when a skill is mastered or requires modification, an essential task for efficient ABA delivery but often reliant on time-consuming manual review and subjective clinical judgment (Ghafghazi et al., 2021).

This research aims to address this specific gap. Techniques for Time-Series Classification (TSC) and analysis of tabular data, such as RNNs, XGBoost, and TabPFN, provide potential tools for analyzing longitudinal progress data. By applying these techniques to longitudinal data collected during ABA sessions, specifically focusing on the progression of objectives (related to items from Vineland) derived from assessments like the Vineland (Sparrow et al., 2016), this project explores the feasibility of classifying objective status ('Valid', 'Completed', or 'Rejected'). This moves beyond the final prediction or initial recommendation toward using ML to support the continuous adaptation and management inherent in effective ABA intervention and IEP implementation. The following sections will review the specific ML techniques applicable to this TSC problem.

2.4 Machine learning for Time Series Classification

The core task in this project involves analyzing longitudinal data to predict the status of IEP objectives over time. The core task of defining the current status ('Valid', 'Completed', or 'Rejected') of an IEP objective based on its historical progress data is a form of TSC, where the input is a sequence of progress scores and contextual information. The output is a categorical status ('Valid', 'Completed', or 'Rejected').

2.4.1 Time Series Classification

TSC involves assigning a categorical label to a sequence of ordered data points (Bagnall et al., 2017). Unlike traditional classification tasks, TSC must account for the sequential nature of the data, which often contains temporal dependencies and correlations that influence predictive accuracy (Ismail Fawaz et al., 2019). In the context of this project, the sequence represents the progress of a specific IEP objective over successive therapy sessions. The goal is to predict that objective's current status based on its historical trajectory and potentially other static or dynamic features. The categorical classification involves discrete states like 'Valid' (continue working on the objective) or 'Completed' (objective mastered).

Various ML algorithms can be adapted or are specifically designed to tackle TSC problems. These methods range from traditional statistical models adapted for sequences to complex deep-learning architectures designed to capture temporal dependencies. The choice of algorithm often depends on the nature of the time series (e.g., length, dimensionality, presence of noise) and the specific prediction task (Esling & Agon, 2012). This section reviews specific algorithms relevant to this project, considering the structured, longitudinal data the partner healthcare company provided.

2.4.2 Support Vector Machine

SVMs are supervised learning models primarily known for classification tasks. The core idea is to find an optimal hyperplane that best separates data points from different classes in a high-dimensional feature space. SVMs aim to maximize the margin (distance) between the separating hyperplane and the nearest data points (support vectors) of any class, contributing to good generalization. Kernel functions (e.g., linear, polynomial, radial basis function) allow SVMs to model non-linear relationships. Strengths include effectiveness in high-dimensional spaces and robustness due to margin maximization. However, performance can be sensitive to the choice of kernel and parameters, and computational cost can be high for large datasets. (Bishop & Nasrabadi, 2006). While initially designed for static data, SVMs can be applied to TSC, often requiring significant feature engineering (Kampouraki et al., 2009). Adapting SVMs effectively for time series usually depends heavily on the quality of the engineered features (Hyde et al., 2019).

2.4.3 K-Nearest Neighbors

KNN is a non-parametric, instance-based learning algorithm for classification and regression tasks. In the context of classification, KNN assigns a class label to a new data point based on the majority class of its 'k' nearest neighbors in the feature space. The 'nearness' is typically determined using a metric like Euclidean distance (Kramer, 2013).

For TSC, standard KNN can be applied by extracting relevant features from the time series data and transforming the sequence into a fixed-length feature vector. Alternatively, specialized distance measures designed for sequences, such as Dynamic Time Warping (DTW), can be used directly within the KNN framework to find the nearest neighboring time series. KNN with DTW has been a strong baseline in many TSC benchmarks (Bagnall et al., 2017).

The main advantages of KNN are its simplicity and the lack of assumptions about the underlying data distribution. However, its performance can degrade with high-dimensional data (the 'curse of dimensionality'), it can be computationally expensive during inference as it requires calculating distances to all training points, and its effectiveness is sensitive to the choice of 'k' and the distance metric used (Kramer, 2013).

2.4.4 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is known for its high accuracy, speed, and feature-importance capabilities. It is popular for tabular and structured data, including time series transformed into a tabular format. XGBoost is a powerful ensemble technique based on gradient-boosted decision trees. XGBoost builds trees sequentially, with each new tree attempting to correct the errors made by the previous ones, resulting in a robust final model. It incorporates regularization techniques to prevent overfitting and can handle missing data implicitly (Chen & Guestrin, 2016). While not inherently sequential models like Recurrent Neural Networks (RNN), it is highly effective for tabular data classification and regression tasks and often achieves state-of-the-art performance (Gorishniy et al., 2021). To apply XGBoost to time series data, the sequence typically needs to be transformed into a fixed-length feature vector, for instance, by extracting statistical features (mean, variance, slope) or using lagged values from the time series as input features (Sharma, 2024).

Traditional ML models often struggle with time series data because they do not inherently capture the temporal dependencies. For instance, simple classifiers may treat each time stamp independently, losing crucial information about the series' evolution over time (Bishop & Nasrabadi, 2006). While traditional tree-based models have dominated the tabular domain, deep learning models, including those adapted for sequential data, are increasingly used for classification tasks, which inherently involve assigning categorical labels (Breejen et al., 2025).

2.4.5 Recurrent Neural Network

RNNs are a class of neural networks specifically designed to handle sequential data, making them naturally suited for time series analysis (Hüsken & Stagge, 2003). They possess internal memory loops that allow information from previous steps in the sequence to persist and influence the processing of current steps. This makes them naturally suited for tasks with temporal dependencies, such as analyzing progress trends over time (Mehdiyev et al., 2017). Variants like Long Short-Term Memory (LSTM) were developed to better handle long-range dependencies often in time series data (Hochreiter & Schmidhuber, 1997; Malhotra et al., 2016).

For TSC, RNNs can directly process the raw time series data point by point, updating their internal state at each step. The final state, or output derived from the sequence of states, can be used for classification (Hüsken & Stagge, 2003; Ismail Fawaz et al., 2019). RNNs excel at capturing temporal patterns but can be computationally expensive to train, especially for long sequences (Mehdiyev et al., 2017).

2.4.6 Transformer

Initially introduced for natural language processing tasks, the Transformer architecture has shown significant promise in various sequence modeling problems, including time series analysis (L. Yang et al., 2025). Its core component is the self-attention mechanism, which allows the model to weigh the importance of different elements in the input sequence when processing a particular element, regardless of their distance. This enables Transformers to capture long-range dependencies more effectively than traditional RNNs. Unlike RNNs, which process data sequentially, Transformers can process sequence elements in parallel, potentially leading to faster training times on suitable hardware (Vaswani et al., 2017). However, standard Transformers require large amounts of training data to perform well and can be computationally intensive. Their application to relatively short or simple time series, such as the progress data for a single IEP objective, might be overly complex or prone to overfitting without sufficient data or regularization (Zerveas et al., 2021).

2.4.7 Tabular Prior-Data Fitted Network

Initially developed for natural language processing, transformers have recently been adapted for tabular data. A recent development in ML for tabular data classification is the

Tabular Prior-Data Fitted Network (TabPFN) introduced by Hollmann et al. (2023). TabPFN is a transformer-based model that has been explicitly developed for tabular data. Unlike traditional models that require training from scratch or fine-tuning on specific downstream datasets, TabPFN is pre-trained on many synthetically generated tabular datasets. This pre-training process aims to learn a 'prior' over various possible tabular data structures and relationships. It uses an attention mechanism and In-Context Learning (ICL) (Brown et al., 2020), meaning it can predict a new data point by considering other data points (the context) provided alongside it in a single forward pass, without requiring task-specific fine-tuning. This makes it particularly strong for small datasets where traditional deep-learning models might struggle (Hollmann et al., 2023, 2025). This algorithm represents a novel direction, adapting the Transformer architecture to the specific challenges of tabular data, including its application to time-series derived features. For TSC, the sequence data must be appropriately formatted as tabular input for TabPFN, potentially including lagged values or time-based features alongside static patient information (Hoo et al., 2025).

The core idea behind TabPFN is its ability to perform zero-shot classification at inference time. Given a new classification task, TabPFN takes the labeled training examples (support set) and the unlabeled test examples (query set) as input context within a single sequence fed into the Transformer. It then directly predicts the labels for the test examples in a single forward pass, without requiring any updates to its model weights (Hollmann et al., 2023).

The first version of the TabPFN was recommended for datasets of up to 1,000 training data points and up to 100 purely numerical features without missing values (Hollmann et al., 2023; Ma et al., 2024). But the latest official version, TabPFN v2, "can be applied to any small-to-moderate-sized dataset and yields dominant performance for datasets with up to 10,000 samples and 500 features" and inherently handles heterogeneous features (numerical, categorical) and missing values (Hollmann et al., 2025). In classification and regression tasks, TabPFN has shown strong performance on tabular datasets of up to 10,000 samples, outperforming tuned and default configurations of XGBoost, CatBoost, and RF using less training time (McElfresh et al., 2023). "In 2.8 s, TabPFN outperforms an ensemble of the strongest baselines tuned for 4 h in a classification setting" (Hollmann et al., 2025).

Given its unique ICL approach and strength on smaller datasets, TabPFN represents a novel direction in tabular data modeling. For IEP management, where datasets related to individual children or specific objectives might be relatively small, TabPFN could offer a fast and effective classification tool. Its ability to function without task-specific hyperparameter tuning is also appealing in clinical settings where extensive model optimization may not be feasible. However, its scalability limitations must be considered for broader applications across large patient cohorts or highly complex feature sets.

2.5 Feature selection and importance techniques

In ML, particularly when dealing with complex, high-dimensional datasets like longitudinal therapy data, identifying the most relevant features (variables) is appropriate. Feature selection is selecting a subset of the original features that are most informative for building a predictive model (Awad & Khanna, 2015). This process offers several advantages: it can reduce the computational cost of training models, mitigate the risk of overfitting by removing irrelevant or redundant features, improve model performance, enhance model interpretability by focusing on the most influential factors, and potentially improve generalization by mitigating the curse of dimensionality (Guyon & Elisseeff, 2003).

Feature selection methodologies are commonly categorized into three main approaches: filter, wrapper, and embedded methods (Guyon & Elisseeff, 2003). In the context of managing IEPs for children with ASD, the available data can be high-dimensional. This data includes longitudinal progress scores for various objectives, assessment results (e.g., Vineland scores per domain), patient demographic information (e.g., age), and treatment history (e.g., number of sessions). Feature selection techniques are crucial for identifying which aspects of this data most predict an objective's future status (e.g., 'Completed' or 'Valid'). This section reviews three relevant techniques: Variance Threshold, SelectKBest, Backward Elimination, Forward Selection, and SHapley Additive exPlanations (SHAP) values.

2.5.1 Filter methods

Filter methods evaluate features based on intrinsic properties, often using statistical metrics (e.g., Chi-Square), and select those exceeding a predefined threshold, independent of any ML model. They are computationally efficient and provide a quick way to screen variables. A straightforward filter is the Variance Threshold, which removes all features with zero or low variance, as these constant or near-constant features provide little to no predictive information (Guyon & Elisseeff, 2003). More sophisticated filter methods use univariate statistical tests to score the relationship between each feature and the target variable. The

SelectKBest strategy, for instance, can be paired with statistical tests like the ANOVA F-test (f_classif), which measures the linear relationship between numerical features and a categorical target (Cabral et al., 2024), or mutual information, which can capture any kind of statistical dependency, including non-linear relationships (François et al., 2007).

2.5.2 Wrapper methods

In contrast, wrapper methods assess the utility of feature subsets by iteratively evaluating their impact on the performance (e.g., accuracy) of a specific predictive model (Kohavi & John, 1997). While potentially yielding model-specific optimal feature sets, this iterative evaluation renders wrapper methods computationally intensive and resource-demanding. Notably, wrapper approaches have been frequently employed in studies applying ML to ASD (Thabtah, 2019). Two classic examples are Backward Elimination and Forward Selection.

Backward Elimination is a type of wrapper method for feature selection. It starts with a model containing all available features. In an iterative process, it removes the feature whose elimination causes the smallest decrease in model performance (or the largest increase if performance improves upon removal). The significance of a feature is typically evaluated using statistical tests (like p-values in regression) or model performance metrics (like accuracy or AUC using cross-validation). This process continues until removing any further feature would lead to a statistically significant drop in performance. While relatively simple to understand, backward elimination can be computationally expensive because it requires retraining the model multiple times (once for each feature removal considered at each step). It is also a greedy approach, meaning it might not find the globally optimal feature subset, as removing a feature early on might prevent finding a better combination later (Guyon & Elisseeff, 2003; Kohavi & John, 1997).

Forward Selection is another wrapper method that works in the opposite direction of backward elimination. It begins with an empty set of features. The first step evaluates all models built with a single feature and selects the feature that yields the best performance according to the chosen evaluation metric. In subsequent iterations, it adds one feature at a time from the remaining available features – specifically, the feature that provides the greatest performance improvement when added to the currently selected set of features. The process terminates when adding more features does not significantly improve model performance beyond a certain threshold or when a predefined number of features has been selected.

Forward selection is often computationally less expensive than backward elimination, if the final number of selected features is small relative to the total number of features. Like backward elimination, it is a greedy algorithm and does not guarantee finding the globally optimal subset of features (Guyon & Elisseeff, 2003; Kamalov et al., 2024; Kohavi & John, 1997).

2.5.3 Embedded methods

Embedded methods perform feature selection as an integral part of the model training process, combining the advantages of both filter and wrapper methods. They are typically more computationally efficient than wrapper methods and are model-specific. A prominent example is the feature importance derived from tree-based ensembles like RF and XGBoost (Theng & Bhoyar, 2024). During the training process, the model can calculate how much each feature contributes to decreasing impurity (e.g., Gini impurity) across all the decision trees in the forest. Features that contribute more to reducing impurity are considered more important (Breiman, 2001). XGBoost can calculate several importance metrics, such as 'gain,' which represents the average improvement in accuracy brought by a feature when it is used in a split, and 'weight,' which is the number of times a feature is used across all trees (*Python API Reference — Xgboost 3.0.4 Documentation*, n.d.). Features that contribute more to the model's decision-making process receive higher importance scores and are considered more significant (Chen & Guestrin, 2016).

2.5.4 SHapley Additive exPlanations

SHAP represents a more recent, unified approach to explaining the output of ML models based on Shapley values from cooperative game theory. Instead of directly selecting a subset of features, SHAP assigns an importance value to each feature for every individual prediction. It explains a prediction by computing the contribution of each feature to pushing the model's output from a base value (e.g., the average prediction over the training set) to the final prediction. This contribution (the SHAP value) is calculated by averaging the feature's marginal contribution across all possible permutations or coalitions of features (Lundberg & Lee, 2017; Molnar, 2020).

A significant advantage of SHAP is its foundation in theory, providing properties like local accuracy (the sum of feature contributions equals the difference between the prediction and the baseline), missingness (features with no impact have zero contribution), and consistency (a change in the model that increases a feature's reliance should not decrease its contribution) (Lundberg & Lee, 2017). While primarily an explainability technique, SHAP values can be leveraged for feature selection. One can measure global feature importance by calculating the average absolute SHAP value for each feature across all samples in a dataset. Features can then be ranked based on this global importance, and a subset can be selected using a threshold or a predefined number. This model-agnostic approach can be applied to various ML algorithms and provides insights into the important features and how they influence predictions locally and globally. This interpretability is particularly valuable in applications like healthcare, where understanding the model's decision-making process is important, and for understanding complex, 'black-box' models like gradient boosting machines (e.g., XGBoost) or neural networks, providing insights beyond simple coefficient or tree-based importance measures (Rajagopalan et al., 2024; Zhu et al., 2024).

2.6 Handling data imbalance

Data imbalance is a common challenge in ML classification tasks, particularly prevalent in real-world datasets like those found in healthcare (Kaur et al., 2019; Khushi et al., 2021). This issue arises when the distribution of classes within the dataset is highly skewed, meaning one class (the majority class) significantly outnumbers the other class(es) (the minority class) (Fernández et al., 2018). In managing IEPs, it is plausible that the dataset derived from therapy sessions will exhibit an imbalance. For instance, an objective might remain in a 'Valid' status for many observations before transitioning to 'Completed,' making 'Completed' a minority class for specific objectives.

Standard ML algorithms are often designed with the implicit assumption of balanced class distributions (Peng & Park, 2022). When applied to imbalanced datasets, these algorithms tend to exhibit bias towards the majority class, as classifying instances into the larger group often minimizes the overall error rate during training. This bias can lead to poor predictive performance for the minority class, which might be the class of primary interest (Gnip et al., 2021). For example, a model might achieve high overall accuracy by simply predicting most objectives as 'Valid' while failing to accurately predict the transition to 'Completed.' Relying on overall accuracy as an evaluation metric in such scenarios can be misleading (Thabtah et al., 2020).

Several techniques have been developed to address the challenges posed by imbalanced data. These techniques can be broadly categorized into data-level approaches and algorithm-level approaches. Algorithm-level methods involve modifying existing learning algorithms to make them more sensitive to the minority class, often by incorporating cost functions that penalize the misclassification of minority instances more heavily. On the other hand, data-level approaches focus on modifying the training dataset to achieve a more balanced distribution before applying a standard ML algorithm (Johnson & Khoshgoftaar, 2019).

Data-level techniques primarily involve resampling the original dataset, either through undersampling or oversampling (Gnip et al., 2021). Undersampling aims to balance the dataset by removing instances from the majority class. While simple to implement, random undersampling carries the risk of discarding potentially valuable information contained within the removed majority instances (Liu et al., 2009). Oversampling focuses on increasing the representation of the minority class. The simplest form, random oversampling, involves duplicating existing minority instances. However, this can lead to overfitting, as the model may become too specialized to the specific minority examples it has seen multiple times (Amin et al., 2016). To mitigate the overfitting risk of simple oversampling, Chawla et al. (2002) introduced the Synthetic Minority Over-sampling Technique (SMOTE). This widely used technique generates new, synthetic minority instances by interpolating between existing minority instances and their nearest neighbors in the feature space. This creates a larger, more diverse minority class representation without duplicating data points. Building upon SMOTE, the Adaptive Synthetic Sampling (ADASYN) approach adaptively generates more synthetic samples for minority instances that are harder to learn based on their density distribution (He et al., 2008). It focuses on synthetic sample generation near the decision boundary, aiming to push the boundary towards the difficult-to-classify minority examples.

Recognizing that oversampling can sometimes introduce noise or blur class boundaries, hybrid methods combine oversampling and undersampling techniques. The goal is to first increase the representation of the minority class using oversampling (like SMOTE) and then apply a targeted undersampling method to clean the data, removing potentially noisy or overlapping instances introduced or exacerbated by the synthetic generation process (Khushi et al., 2021). SMOTE-Tomek method combines SMOTE with Tomek Links undersampling. A Tomek link consists of a pair of instances from different classes that are each other's nearest neighbors. In cleaning after SMOTE, the majority class instance from any identified Tomek link is removed. This helps clarify the class boundary by removing the majority of instances that are potentially too close to minority instances (Batista et al., 2004). The SMOTE-ENN approach combines SMOTE with Edited Nearest Neighbors (ENN) undersampling. After SMOTE generates synthetic minority samples, ENN is applied to remove instances (from either class) whose class label differs from the class of most of its k-nearest neighbors. This acts as a cleaning mechanism, removing instances considered noisy or potentially mislabeled, thereby improving the definition between classes (Batista et al., 2004).

Given the potential for imbalanced classes in the IEP objective status data (e.g., 'Valid' vs. 'Completed'), understanding and potentially applying these data resampling techniques is crucial. Addressing data imbalance helps ensure that the resulting ML model is sensitive to clinically meaningful changes, even if they are relatively infrequent in the dataset.

2.7 Model evaluation

Evaluating the performance of ML models is an important step in the development process. Objective metrics are necessary to quantify how well a model performs on a given task and to compare different models or approaches. In the context of this project, the task is a classification problem: predicting whether the status of an IEP objective should be 'Valid', 'Completed', or 'Rejected.' Several standard metrics are commonly used for evaluating classification models.

2.7.1 Accuracy, Precision, Recall, Specificity

Accuracy is a common metric, representing the proportion of correct predictions out of the total predictions made. While intuitive, accuracy can be misleading, especially when dealing with imbalanced datasets (Kaur et al., 2019). For example, if 95% of objectives remain 'Valid' and only 5% are 'Completed,' a model that always predicts 'Valid' would achieve 95% accuracy but would be useless for identifying completed objectives.

Precision and Recall are often used to get a more nuanced view of performance, especially with imbalanced classes (Kaur et al., 2019). These metrics are derived from the confusion matrix, which categorizes predictions into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

- Precision: Measures the proportion of correctly predicted positive instances among all instances predicted as positive (TP / (TP + FP)). In our context (considering

'Completed' as the positive class), high precision means that when the model predicts an objective is 'Completed,' it is likely correct. This is important if stopping to work on an objective prematurely has significant implications (e.g., an objective is classified as 'Completed,' but it was still not mastered).

- Recall (Sensitivity): Measures the proportion of actual positive instances that were correctly identified by the model (TP / (TP + FN)). High recall means the model successfully identifies most of the objectives that are truly 'Completed.' This is important if failing to recognize a completed objective and unnecessarily continuing intervention is costly or inefficient.
- Specificity: Measures the proportion of actual negative instances correctly identified by the model (TN / (TN + FP)). High specificity indicates that when an objective is still 'Valid' (not completed), the model is likely to correctly identify it as such. This is important to avoid prematurely stopping work on an objective still requiring attention.

There is often an inherent trade-off between Precision and Recall; improving one may negatively impact the other. The relative importance depends on the specific clinical implications of false positives versus false negatives (Khushi et al., 2021).

2.7.2 F1-score

The F1-score provides a single metric that balances Precision and Recall. It is calculated as the harmonic mean of Precision and Recall: F1 = 2 * (Precision * Recall) / (Precision + Recall).

The F1-score is useful when the class distribution is imbalanced and when both false positives and false negatives are essential to minimize. A high F1-score indicates the model has high precision and recall (Kohli et al., 2022).

2.7.3 Area Under the Receiver Operating Characteristic Curve

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a graphical plot illustrating the diagnostic ability of a binary classifier system as its discrimination threshold is varied. A Receiver Operating Characteristic (ROC) curve is generated by plotting the TP Rate (Recall) against the FP Rate (1 - Specificity) at various threshold settings (Swets, 1988). The Area Under the Curve (AUC) represents the area under this ROC curve. It provides a single scalar value summarizing the overall performance of a classifier across all possible threshold values (Bradley, 1997).

The AUC quantifies the overall ability of the model to discriminate between the positive and negative classes across all possible thresholds. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 indicates performance no better than random guessing (Fawcett, 2006). AUC is a valuable metric because it is generally insensitive to class distribution (data imbalance) and provides a comprehensive, threshold-independent measure of classification performance (Chawla et al., 2002; Farshidvard et al., 2023).

While the standard AUC is defined for binary classification, its application to multi-class problems requires a strategy to decompose the task into one or more binary problems. The two most common strategies are One-vs-Rest (OvR) (Rifkin & Klautau, 2004) and One-vs-One (OvO) (Hastie & Tibshirani, 1997). The OvR approach trains a single binary classifier for each class against all other classes combined. The final AUC is then typically an average of the AUC scores from each classifier. The OvO approach, in contrast, trains a separate binary classifier for every unique pair of classes. While OvO can be effective, particularly for algorithms that do not scale well with the number of samples, the OvR approach is often preferred for its computational efficiency with many classes and its high degree of interpretability, as each classifier's performance directly measures a single class's separability from the rest (Y. Yang et al., 2024).

The choice of evaluation metrics is crucial for understanding a model's strengths and weaknesses in the specific context of IEP objectives status prediction. Depending on the clinical goals, whether it is more critical to avoid missing completed objectives (favoring Recall) or to be highly confident when predicting completion (favoring Precision), different metrics or a combination of them, like F1-score or AUC, should guide model selection and interpretation (Kaur et al., 2019).

3 METHODOLOGY

This chapter details the quantitative methodology employed to develop and evaluate machine learning (ML) models for predicting the status progression of Applied Behavior Analysis (ABA) Individualized Education Plan (IEP) objectives. The research utilizes a longitudinal dataset from a partner healthcare company specializing in Autism Spectrum Disorder (ASD) intervention in children. The methodology follows a structured pipeline encompassing data acquisition, preprocessing, feature engineering, model training, and evaluation. A central task was transforming time-series progress data into a tabular format, a necessary step for applying models designed for structured data (Gorishniy et al., 2021). Subsequently, feature engineering techniques were used to extract temporal patterns into a set of predictive variables. The study primarily evaluates the performance of a pre-trained Transformer-based model, Tabular Prior-Data Fitted Network (TabPFN), which has shown strong results on tabular data (Hollmann et al., 2023, 2025). Its performance is benchmarked against several well-established ML algorithms to provide a comparative context. Model performance was assessed using standard classification metrics to handle the potential for class imbalance inherent in clinical data (Kaur et al., 2019).

3.1 Data acquisition and description

The data for this research were acquired through a partnership with a Brazilian healthcare provider specializing in delivering multidisciplinary care, including ABA therapy, to children with ASD. The dataset comprises de-identified, longitudinal clinical records collected systematically during routine therapeutic sessions. The partner company granted access to this anonymized dataset. To ensure the research was grounded in the IEP management process's clinical reality, semi-structured interviews were conducted with key personnel, including product managers, software developers, and senior psychologists. These consultations were essential for several aspects of the study's design:

- Understanding the data: The interviews helped clarify the meaning of key variables, the context of data collection during therapy sessions, and the process by which therapists assign progress scores.
- Validating assumptions: Discussions with senior psychologists validated the fundamental assumption of a consistent clinical logic behind status changes ('Validated', 'Completed', 'Rejected'), essential for framing the ML problem.

- Defining the scope: The insights gained helped to determine the scope of the problem, such as the decision to focus on the primary Vineland-3 domains and later to cap the time-series data at 52 checks, which was confirmed as a clinically relevant timeframe.

Raw data was extracted from the provider's production database using a series of SQL queries, resulting in a comprehensive collection of session notes and progress evaluations. The initial raw dataset contained 63,808 rows (objective progress checks), which refers to 3,967 unique learning objectives, described by 17 initial variables.

The fundamental unit of analysis is a single learning objective within a child's IEP. The dataset captures the progression of these objectives over time. The raw dataset is structured in a long format, where each row represents a single progress check for a specific objective at a particular point in time. Key variables can be categorized into identifiers, static features, and the time-series component.

The primary outcome variable for this study is the status_after_check, a categorical label assigned by a senior therapist that indicates the current status of an objective. This target variable defines the classification task for the ML models and has three possible classes:

- Validated: The objective remains active and in progress. The child continues to work on this skill.
- Completed: The objective has been mastered by the child. The intervention for this specific skill is considered successful.
- Rejected: The objective has been discontinued for reasons other than mastery, typically due to a sustained lack of progress.

Predicting this variable is a multi-class classification problem that mirrors the core decision-making process of senior therapists when reviewing an IEP.

3.1.1 Variable selection and rationale

The variables included in the initial dataset were selected to provide a comprehensive view of the child, the therapeutic context, and the learning trajectory of each objective. The rationale for including these key variables is grounded in their clinical relevance to ABA therapy, documented importance in the literature, and insights from clinical experts interviewed at the partner company. Identifiers include unique codes for each child (case_id) and the specific objective (objective_id). The static features, which represent variables that do not change throughout a single objective's timeline, include:

- Demographic and child-specific features, such as age and speaking_ability_answer, were included as they are known to influence treatment outcomes. Age at the start of an intervention is a consistently cited predictor of progress in ASD, with earlier intervention generally associated with more favorable outcomes (Dawson & Bernier, 2013). The child's speaking ability provides a baseline measure of communication skills, a core deficit in ASD, and a primary target for intervention (American Psychiatric Association, 2013; Sparrow et al., 2016).
- Treatment-related features, specifically number_of_sessions and minutes_aba_last_recomended, were selected to quantify the intensity and duration of the intervention. The literature consistently demonstrates a strong relationship between the therapy volume and mastery of learning objectives (Linstead et al., 2017; Maharjan et al., 2023). These variables represent the therapeutic 'dose' and are considered essential predictors of progress. This was corroborated in interviews, where clinicians noted that the recommended hours of therapy are a key component of the treatment plan.
- Clinical baseline features were represented by the initial scores in the four core Vineland-3 domains: communication_domain, socialization_domain, daily_living_domain, and motor_domain. These scores provide a standardized, multi-dimensional assessment of the child's adaptive functioning at the start of the objective's tracking period (Sparrow et al., 2016). They serve as a baseline to measure progress and contextualize the learning process. To maintain a focused scope and reduce interdisciplinary variability, the initial dataset was filtered to include only objectives managed by psychologists and associated with the Vineland protocol.
- Objective-specific features, such as protocol_item_id and item_domain, describe the nature of the learning objective itself. While protocol_item_id was later removed due to high cardinality, its corresponding item_domain was retained. This categorical feature allows the model to consider whether the learning trajectory differs across skill domains (e.g., 'communication' vs. 'motor skills'), which is a clinically relevant factor.
- The raw dataset also included identifier and metadata variables essential for data integrity and clinical context, even if not all were used as final model predictors. The assessed_by_id column, representing the unique identifier for the supervising psychologist, was initially considered as it could potentially capture inter-therapist variability. The date_update_check timestamp was crucial during data extraction to ensure the correct temporal sequence of records. As detailed in the data reduction

section, these variables were ultimately excluded from the final feature set for technical reasons, specifically to avoid overfitting from high-cardinality identifiers and to remove non-predictive metadata.

These features provide a comprehensive snapshot of the child and the therapeutic context at the start and during the intervention for a specific objective.

Finally, the raw data captures the time-series component in a long format. For each learning objective, multiple entries can exist, with each representing a progress check-in. The day_check column indicates the temporal sequence of these checks, while the objective_progress column records the therapist-assigned score at each point. This longitudinal structure provides the raw data for tracking the learning trajectory, which is then transformed into a wide format for the analysis, as detailed in Chapter 3.2, and from which dynamic features were later engineered, as detailed in Chapter 3.3.

This study operates on fundamental assumptions regarding the data generation process. It is assumed that a consistent clinical logic is applied by senior psychologists when evaluating and updating the status of an objective. This implies that the decision-making process for changing a status is reasonably standardized across different therapists and cases. Furthermore, the data structure suggests a consistent temporal sequence where progress checks are recorded at regular intervals, and the status of an objective is determined immediately following each recorded progress check. These assumptions are critical for treating the longitudinal data as a coherent time series for each objective.

3.2 Data preprocessing and preparation

Following data acquisition, a multistep preprocessing pipeline was implemented to transform the raw, longitudinal data into a structured format suitable for ML analysis. This process was essential for cleaning the data, handling its temporal nature, and ensuring compatibility with the selected modeling algorithms.

3.2.1 Initial data reduction

The first preprocessing step was the removal of columns that were not intended for use as predictive features. This initial reduction streamlined the dataset to focus exclusively on variables containing potentially predictive demographic, clinical, and temporal information.

These fell into three categories: unique identifier columns, metadata columns, and high-cardinality categorical identifiers:

- Unique identifier columns: Variables such as case_id and objective_id were dropped. While essential for data management and joining tables during the acquisition phase, they serve only as unique labels for each instance and hold no generalizable predictive value for the model.
- Metadata columns: The date_update_check column, a timestamp for when the record was last updated, was removed. This type of metadata pertains to the data logging process rather than the child's clinical progression. It is not a direct feature of the therapeutic process being modeled.
- High-cardinality categorical identifiers: Columns assessed_by_id (unique identifier for the psychologist) and protocol_item_id (unique identifier for the specific ABA curriculum item) were also excluded from the final feature set. These variables are characterized by high cardinality (many unique values). Including them via one-hot encoding would dramatically increase the feature space with hundreds of sparse, binary columns, heightening the risk of overfitting. To validate this theoretical concern, an experiment was conducted where these identifiers were one-hot encoded and included in the model training (when keeping the one-hot encoded variables assessed_by_id and protocol_item_id, it resulted in 489 columns). The result was a slight degradation in predictive performance on the test set, confirming that the model struggled to generalize from these highly sparse features. Based on this empirical evidence, these high-cardinality identifiers were definitively removed from the final feature set to create a more robust and generalizable model.

3.2.2 Data filtering and structuring

The refined dataset was then filtered to create a focused analytical sample. A key decision was made regarding handling the three possible objective statuses: 'Validated', 'Completed', and 'Rejected'. Although the 'Rejected' status constitutes a small minority of the instances, presenting a significant class imbalance challenge, it was retained in the dataset. This decision ensured the predictive model would be trained on the full spectrum of clinical outcomes, as discontinuing an objective is a valid and essential therapeutic event. The problem was therefore framed as a multi-class classification task. To further refine the scope and reduce variability, the dataset was filtered to include only objectives corresponding to the primary ABA domains within the Vineland-3 framework: Communication, Socialization, Daily Living Skills, and Motor Skills. This filtering step reduced the number of objectives for analysis to 3,344.

A central preprocessing task transformed the data from a long format, where each row represented a single progress check, to a wide format. In this new structure, each row corresponds to a unique learning objective. The time-series component was unstacked so that the sequence of progress scores was represented by a series of columns, each corresponding to a specific progress check (e.g., '1', '2', '3', ...). This pivoting is a prerequisite for applying tabular ML models that expect a fixed number of input features per instance.

The maximum length of these sequences was capped at 52 progress checks. This limit was established based on both clinical practice and data distribution. Clinically, the Vineland assessment, which provides the framework for these objectives, is re-evaluated semiannually. It is uncommon for a single objective to remain active without a change beyond 3 months without significant review or modification. The objective check is performed around weekly, but not necessarily, because different therapists may do a check in the same week. Being conservative, it was considered around 1 year of check. A dataset analysis confirmed that 95% of the objectives had fewer than 52 checks, and within this 52-check timeframe, 42% of the objectives in the dataset reached a terminal status ('Completed' or 'Rejected'). Therefore, this cutoff effectively captures the most predictive phase of an objective's lifecycle. This truncation standardizes the input feature space for all models. It mitigates the influence of outlier objectives that might remain active for exceptionally long periods, potentially due to inconsistent data tracking or shifts in therapeutic focus.

3.2.3 Data cleaning and encoding

Data cleaning and type conversion were performed to ensure data integrity. Some numerical variables were initially stored as string objects, including the progress scores and baseline Vineland domain scores. These were converted to a numeric float format. Categorical variables, such as item_domain, were transformed using one-hot encoding to create new binary columns for each category, preventing the model from assuming an ordinal relationship between them. Following the removal of columns, filtering, structuring, and encoding steps, the dataset, ready for feature engineering, consisted of 3,344 rows (unique objectives) and 65 columns. Finally, the categorical target variable, status_after_check, was numerically encoded to 'Validated': 0, 'Completed': 1, and 'Rejected': 2, to serve as the dependent variable for the classification models.

3.3 Feature engineering

While the wide-format representation of progress scores provides a complete temporal history, the raw sequence data may not explicitly highlight the most predictive patterns for ML models. To address this, a feature engineering process was undertaken to extract the dynamics of each objective's learning trajectory into a set of quantitative, descriptive features. This transformation is vital for enabling tabular models, which do not inherently process sequential data, to interpret the temporal characteristics of the progress data. For each unique objective (i.e., each row in the dataset), the following features were systematically generated from its corresponding sequence of up to 52 progress scores. The engineered features were designed to capture distinct aspects of an objective's history and were grouped into several categories.

- Basic statistical summaries: This set of features was calculated to summarize the overall performance distribution for each objective. These included the mean (ts_mean), median (ts_median), standard deviation (ts_std_dev), minimum (ts_min), and maximum (ts_max) of all recorded progress scores for the sequence. These metrics provide a global overview, representing the average performance, typical performance, consistency of progress, and the range of scores achieved, respectively.
- Sequence characteristics: These features were created to describe the sequence characteristics. The ts_length was calculated as the count of non-missing progress checks, representing the total duration of active work on the objective. The ts_last_progress captured the sequence's most recent valid progress score, indicating the child's current performance level.
- Trend feature: It was engineered to quantify the overall learning trajectory. A linear regression was fitted to the sequence of non-missing progress scores against their corresponding time steps. The resulting slope (ts_slope_overall) was stored as a feature. A positive slope indicates a general trend of improvement over time, a negative slope suggests regression, and a slope near zero indicates stagnation.
- Recency and stability features: This group focused on the most recent therapeutic activity, assuming that recent performance is highly predictive of the next status. These features were calculated using the last five valid data points in each sequence. They included these recent scores' mean (mean_last_n) and standard deviation (std_last_n) to capture short-term performance and stability. To measure the consistency of peak performance, time_at_max_last_n was engineered to count how

many times the objective's overall maximum score appeared within this recent five-point window. Additionally, days_since_improve_last_n was calculated to measure the number of sessions that have passed since the last recorded improvement, directly quantifying momentum.

This feature engineering process transformed the dataset from a simple sequence of scores into a meaningful feature space. Each objective was described by its raw progress and overall statistical profile, trend, and recent dynamics, providing a more informative input for the subsequent feature selection and model training stages. Specifically, the 52 raw time-series columns were replaced by the 12 engineered features. The final feature matrix for model training consisted of 3,344 samples (rows) and 24 features (columns).

3.4 Feature selection and data imbalance

After the feature engineering process culminated in a final set of 24 predictive features, the next step was to investigate whether a smaller subset of these variables could yield an equally or more effective model. This section details the experiments conducted for feature selection and the approach to handling the inherent class imbalance in the target variable.

3.4.1 Feature selection and importance

While a feature set of 24 variables is not excessively large, feature selection remains an important step in an ML pipeline. The primary goals were to identify the most influential predictors of an objective's status, reduce potential noise from less relevant features, ensure the final model was as simple as possible without sacrificing predictive power (Guyon & Elisseeff, 2003), and to understand the impact of each feature on the final result of the models. Therefore, a systematic feature selection process was employed to evaluate the contribution of the engineered and static features.

Several standard feature selection methodologies were considered. Wrapper methods, such as forward selection and backward elimination, were initially explored. However, due to their iterative nature, which requires training the model multiple times, they were found to be computationally prohibitive within the project's environment used for this project and were not pursued further. The investigation, therefore, focused on more computationally efficient filter and embedded methods, and model explainability methods.

- Filter methods: As a preliminary step, a VarianceThreshold filter was applied to remove any zero-variance features, confirming none were present. The filter method tested was SelectKBest using the mutual_info_classif score to quantify the statistical dependency between each feature and the target variable.
- Embedded method: Feature importance derived from a trained Random Forest model was used to rank features based on their contribution to the ensemble. Additionally, to further explore feature relevance and provide a basis for potential comparison, feature importance was extracted from a trained XGBoost model, using 'gain' as importance type.
- Explainability-based method: SHAP, a model-agnostic method, generated feature importance scores. Due to the high computational cost of calculating SHAP values on the full dataset, the analysis was performed on a balanced subset of the test data. This subset was created by randomly sampling up to 20 instances from each of the two majority target classes ('Validated', 'Completed') and 17 instances from the minority target classes ('Rejected'), ensuring a more representative analysis across both majority and minority classes. While this approach approximates global feature importance rather than an exact calculation, it offers valuable model-agnostic insights into which features most influenced the classifier's predictions.
- A correlation analysis was also conducted on the engineered numerical features to identify and remove multicollinearity. Highly correlated features (Pearson correlation > 0.9) were identified, such as ts_mean and ts_median. An experiment was run in which the feature with the lower individual correlation to the target variable (ts_median) was removed. However, this led to a slight degradation in model performance, suggesting that both features, while correlated, captured unique nuances valuable to the model.

The results of these experiments consistently showed that applying feature selection did not improve, and in some cases slightly degraded, the predictive performance of the primary model, TabPFN. This outcome is likely due to TabPFN's underlying Transformer architecture, which contains internal self-attention mechanisms that act as a dynamic, sample-specific form of feature selection (Hollmann et al., 2023, 2025). Given that an external selection step offered no benefit, the decision was made to retain the complete set of demographic and child-specific, treatment-related, clinical baseline, and objective-specific (item_domain), and engineered features for the final analysis (total of 24 features), allowing the model to leverage all available information.

3.4.2 Handling data imbalance

The dataset exhibited a significant class imbalance, a common challenge in clinical datasets. Specifically, of the 3,344 objectives in the final dataset, 1,953 (58.4%) were labeled 'Validated', 1,307 (39.1%) were 'Completed', and only 84 (2.5%) were 'Rejected'. This distribution shows that the 'Rejected' class is a distinct minority. Standard ML algorithms can develop a bias towards the majority class in such scenarios, leading to poor predictive performance for minority classes, often of high clinical interest (Kaur et al., 2019).

To address this, data-level resampling techniques were explored. Two methods were tested: ADASYN and SMOTE-ENN. Experiments showed that applying both degraded the TabPFN's predictive performance on the test set. Given these results, a decision was made to proceed without applying data resampling techniques. Instead, the challenge of class imbalance was addressed during the evaluation phase by utilizing metrics that provide a more nuanced assessment than overall accuracy, such as the weighted F1-score, class-specific precision and recall, and AUC. This approach ensures that the model is trained on the natural distribution of the clinical data.

3.5 Model implementation and training

This study implemented and evaluated a state-of-the-art ML model to address the research question and benchmarked its performance against several well-established baseline algorithms. The primary goal was to assess the efficacy of a pre-trained, transformer-based model for this specific clinical prediction task. At the same time, the baselines provide a comparative context rooted in standard ML practices.

3.5.1 Primary model: TabPFN

The primary model selected for this project is the TabPFN. TabPFN is a recent advancement in tabular data classification that leverages a Transformer-based architecture pre-trained on a vast array of synthetically generated datasets (Hollmann et al., 2023, 2025). This initial training helps the model understand general patterns in table-like data, enabling it to make accurate predictions on new, small-to-moderate-sized datasets in a zero-shot manner, meaning it does not require task-specific hyperparameter tuning.

TabPFN was chosen for several reasons. First, its performance on datasets with up to 10,000 samples makes it well-suited for this study's clinical dataset scale. Second, its efficiency is a significant advantage; the model is used directly for inference after a single fit call, which is computationally less demanding than the extensive tuning required by traditional models. This is particularly relevant for potential clinical deployment, where computational resources and time may be limited. The implementation utilized the official TabPFNClassifier library ("PriorLabs/TabPFN", 2025), which was trained on the fully preprocessed and feature-engineered dataset.

3.5.2 Baseline models for comparison

A suite of well-established classification algorithms was also implemented to provide a benchmark for TabPFN's performance. These models were selected to represent various learning paradigms, from traditional statistical models to deep learning architectures. It is important to note that these baseline models were trained using their default or common parameters without extensive hyperparameter tuning. This approach establishes a fair comparison against the zero-shot nature of TabPFN.

The baseline models included:

- XGBoost: A widely used gradient boosting algorithm known for its high performance on structured, tabular data. It builds an ensemble of decision trees sequentially, with each tree correcting the errors of its predecessor (Chen & Guestrin, 2016). The objective function was set to multi:softprob to handle the multi-class nature of the problem, enabling the model to output a probability distribution across the three target classes. The evaluation metric used during training was the multiclass classification error rate, or merror. A learning rate of 0.1 was selected to control the contribution of each tree to the final ensemble. To manage model complexity and prevent overfitting, the maximum depth of each tree was limited to 10. Additionally, stochastic gradient boosting was employed by setting both the subsample and colsample_bytree parameters to 0.8, meaning each tree was built using a random subsample of 80% of the training instances and 80% of the features, respectively.

- SVM: A classic classification algorithm that finds an optimal hyperplane to separate classes in a high-dimensional space. The implementation used a radial basis function (RBF) kernel to handle non-linear relationships.
- KNN: A non-parametric, instance-based learning algorithm that classifies a data point based on the majority class of its 'k' nearest neighbors in the feature space. In this case, k=10 was used.
- LSTM: An LSTM-RNN was implemented to leverage the data's sequential nature. Unlike the other models that relied on the feature-engineered tabular data, the LSTM was trained on a dataset structured as sequences. The architecture consisted of two LSTM layers with a hidden size of 64, followed by a fully connected layer that integrated the final hidden state with the static features (e.g., age, baseline scores) to produce the final classification. This allowed for a direct comparison between the feature engineering and end-to-end sequential modeling approaches.

3.6 Experimental design and evaluation

An experimental design was implemented to ensure a reproducible assessment of the ML models. This section details the complete workflow, from data preparation to the metrics used for performance evaluation, ensuring that the comparison between the primary TabPFN model and the baselines was conducted under fair and consistent conditions.

3.6.1 Experimental pipeline

The evaluation of all models followed a structured pipeline to ensure consistency. The process was executed in the following sequence:

1. Data acquisition and preprocessing: Raw data were loaded, filtered, and transformed from a long to a wide format as described in Section 3.2.
2. Feature finalization: The final feature set was constructed by combining the pre-existing static and categorical features with the 12 features engineered from the time-series data, resulting in a total of 24 predictive variables as detailed in Section 3.3
3. Data splitting: The dataset was partitioned into training and testing sets.
4. Imputation and scaling: Missing values were handled, and features were scaled. These steps were performed after the data split to prevent data leakage from the test set into the training process.

5. Model training: Each model was trained exclusively on the preprocessed training dataset.
6. Model evaluation: The trained models were used to make predictions on the unseen test dataset, and their performance was quantified using a standard set of classification metrics.

3.6.2 Data splitting, imputation, and scaling

The complete dataset, consisting of 3,344 unique objectives and their 24 corresponding features, was divided into a training set (80% of the data) and a testing set (20%). This partitioning was performed using a stratified split to ensure that the distribution of the target classes ('Validated', 'Completed', 'Rejected') was preserved in both sets. A fixed random_state was used to guarantee the reproducibility of the split across all experiments.

Following the split, preprocessing steps were carefully applied to avoid data leakage. The pipeline included a SimpleImputer with a 'mean' strategy as a measure to handle any potential missing values across all numerical features, both the static variables from the original dataset and the newly engineered ones. However, data analysis after the train-test split confirmed no missing values were present in the training or testing sets. Therefore, while the Imputer was fitted on the training data as a procedural safeguard, it ultimately did not alter any values. Subsequently, all numerical features were standardized using the StandardScaler, which scales data with a mean of 0 and a standard deviation of 1. Consistent with best practices, the scaler was fitted exclusively on the training data and applied to both sets.

3.6.3 Model evaluation metrics

Given the dataset's multi-class nature and significant class imbalance, relying solely on accuracy can be misleading. A model that always predicts the majority class ('Validated') could achieve high accuracy while being clinically useless. Therefore, a set of metrics, derived from the confusion matrix, was used to evaluate performance:

- Precision: Measures the proportion of correct positive predictions (True Positives / (True Positives + False Positives)). In this context, high precision for the 'Completed' class is critical, as it indicates that when the model suggests an objective is mastered, it is highly likely to be correct, thus minimizing the risk of prematurely ending intervention on a skill.

- Recall (Sensitivity): Measures the proportion of actual positive instances that were correctly identified by the model (True Positives / (True Positives + False Negatives)). High recall for the 'Completed' class is essential for therapeutic efficiency, ensuring that mastered objectives are promptly identified so clinical focus can shift to new goals.
- F1-score: The harmonic mean of Precision and Recall, providing a single score that balances both metrics. It is beneficial for evaluating performance on imbalanced classes where both false positives and false negatives are important.

These metrics were calculated for each class individually. Weighted and macro averages for Precision, Recall, and F1-score were reported to summarize overall performance. The weighted average accounts for class imbalance by weighting the score of each class by its proportion in the dataset, while the macro average treats all classes equally, providing a better indication of performance on the infrequent minority classes.

In addition to these metrics, the AUC-ROC was calculated. The AUC-ROC score represents a model's ability to discriminate between classes across all possible thresholds. For this multi-class problem, the AUC-ROC was computed using the One-vs-Rest (OvR) strategy. This approach was primarily chosen over the One-vs-One (OvO) alternative for its clear interpretability. The OvR method calculates an AUC score for each class by treating it as the positive class and all other classes as the negative class. This directly measures the model's ability to answer clinically relevant questions, such as "How well can the model distinguish a 'Completed' objective from all non-completed objectives?" A weighted average of the individual class AUCs accounted for class imbalance in the final aggregate score.

3.7 Computational environment

The entire data analysis and ML pipeline was implemented using the Python programming language within the Google Colaboratory cloud environment. Data acquisition from the Google BigQuery database was performed using standard SQL queries. Subsequently, data manipulation, cleaning, and structuring relied on the pandas library for DataFrame operations and NumPy for numerical computations. The implementation of the ML models was accomplished using specialized libraries. The scikit-learn library provided the framework for baseline models such as Support Vector Machines and K-Nearest Neighbors, and essential preprocessing tasks including data splitting, scaling, and evaluation metric calculation (Pedregosa et al., 2011). The primary model was implemented using the TabPFN library (“PriorLabs/TabPFN”, 2025). The XGBoost model was built using the xgboost library (“XGBoost Documentation — xgboost 3.0.4 documentation”, 2025), and the LSTM was constructed and trained using the PyTorch deep learning framework (Paszke et al., 2019). Experiments with data resampling were conducted using the imbalanced-learn library, and data visualizations were generated with Matplotlib. Model training and computation were accelerated using an NVIDIA GPU provided within the Google Colab environment, accessed through the PyTorch CUDA interface.

4 RESULTS

This chapter presents the empirical findings from the machine learning (ML) experiments. The structure of this chapter is organized to first provide a descriptive overview of the analytical dataset. It then details the comparative performance of the implemented models, followed by a focused analysis of the most effective models and an investigation into the features that drive their predictions.

4.1 Descriptive statistics of the analytical dataset

Following the data preprocessing and feature engineering pipeline detailed in Chapter 3, the final dataset for analysis was constructed. The dataset consists of 3,344 unique learning objectives, derived from the records of 438 unique children. Each objective is represented by 24 predictive features in total, including static demographic and clinical variables, and 12 features engineered to capture the temporal dynamics of the learning progress scores.

The distribution of the target variable, status_after_check, confirms the presence of a significant class imbalance, a common characteristic of clinical data (Johnson & Khoshgoftaar, 2019). The 'Validated' class, representing objectives that remain in progress, is the majority class, with 1,953 instances, which accounts for 58.4% of the dataset. For objectives that have been mastered, the' Completed' class is the next most frequent, comprising 1,307 instances (39.1%). The 'Rejected' class, representing objectives discontinued for reasons other than mastery, is a distinct minority, containing only 84 instances (2.5%). This skewed distribution, especially the rarity of the 'Rejected' class, informed the selection of evaluation metrics designed to provide a balanced assessment of model performance across all classes.

The descriptive statistics for the continuous numerical features are summarized in Table 1. The data represent a pediatric population with a mean age of 5.37 years. The intervention intensity shows considerable variability, with the number of sessions ranging from 2 to 397. The analysis of the engineered time-series features reveals that, on average, an objective has approximately 17 progress checks (ts_length). The average progress score (ts_mean) is 0.63, and the general trend, as indicated by the mean slope (ts_slope_overall), is slightly positive at 0.02. The distribution of the primary categorical feature, item_domain, shows that most objectives fall under 'Communication' (52.9%) and 'Socialization' (38.6%), followed by 'Motor skills' (4.8%) and 'Daily Living Skills' (3.7%).

Table 1 - Descriptive statistics of the analytical dataset

Feature	Mean	Std. Dev.	Min	Median	Max
Static features					
age	5.37	1.39	2.00	5.00	9.00
number_of_sessions	100.32	77.43	2.00	76.00	397.00
minutes_aba_last_recomended	344.59	151.97	0.00	360.00	720.00
speaking_ability_answer	1.07	0.63	0.00	1.00	2.00
daily_living_domain	72.07	12.67	20.00	72.00	135.00
communication_domain	61.17	20.96	20.00	64.00	115.00
socialization_domain	71.06	14.67	32.00	71.00	116.00
motor_domain	82.94	14.25	20.00	82.00	140.00
Engineered features					
ts_mean	0.63	0.18	0.00	0.65	1.00
ts_median	0.65	0.22	0.00	0.70	1.00
ts_std_dev	0.23	0.09	0.00	0.23	0.71
ts_min	0.26	0.25	0.00	0.25	1.00
ts_max	0.91	0.18	0.00	1.00	1.00
ts_length	16.82	14.08	2.00	12.00	52.00
ts_last_progress	0.72	0.28	0.00	0.75	1.00
ts_slope_overall	0.02	0.10	-0.75	0.01	1.00
mean_last_n	0.68	0.20	0.00	0.70	1.00
std_last_n	0.19	0.11	0.00	0.19	0.71
time_at_max_last_n	1.90	1.37	0.00	2.00	5.00
days_since_improve_last_n	1.44	1.51	0.00	1.00	5.00

4.2 Model performance comparison

To evaluate the feasibility of predicting the status progression of IEP objectives, the primary model, TabPFN, was benchmarked against four baseline algorithms: XGBoost, SVM, KNN, and an LSTM network. Each model was trained on the preprocessed training set, comprising 2,675 unique objectives, and evaluated on the unseen test set, comprising 669 unique objectives. The overall performance was assessed using metrics suited for multi-class classification with imbalanced data.

The comparative results are summarized in Table 2. The findings indicate a clear performance distinction between the models. The TabPFN and XGBoost models demonstrated superior predictive capabilities across all key metrics. TabPFN achieved the highest accuracy at 0.82 and the highest weighted OvR AUC-ROC score of 0.900. XGBoost performed nearly identically, with an accuracy of 0.81 and a weighted AUC-ROC of 0.897. Both models attained a weighted F1-score of 0.81, indicating a strong ability to classify the majority class objectives ('Validated' and 'Completed') while accounting for the class imbalance. However, their handling of the minority classes differed. The difference is in the macro-averaged F1-scores, where XGBoost (0.61) outperformed TabPFN (0.55), suggesting a better, though still limited, ability to classify the underrepresented 'Rejected' class.

Table 2 - Comparative performance of all models on the Test Set

Model	**Accuracy**	**Weighted avg F1-score**	**Macro avg F1-score**	**Weighted OvR AUC-ROC**
TabPFN	**0.82**	**0.81**	0.55	**0.900**
XGBoost	0.81	**0.81**	**0.61**	0.897
SVM	0.73	0.72	0.48	0.826
LSTM	0.72	0.71	0.52	0.810
KNN	0.71	0.70	0.47	0.793

Note: Bold values indicate the best performance for each metric.

In contrast, the other baseline models, SVM, KNN, and LSTM, exhibited lower performance. Their accuracy scores ranged from 0.71 to 0.73, and their weighted F1-scores were between 0.70 and 0.72 (see Appendix A for detailed, class-specific results). The macro-averaged F1-scores, which treat all classes equally, were consistently low across all models, ranging from 0.47 to 0.61. This suggests that while the models performed well on the majority classes, all faced challenges in correctly identifying the minority 'Rejected' class. Based on their superior overall performance, TabPFN and XGBoost were identified as the most effective models for this classification task.

4.3 In-depth analysis of the best-performing models

While overall metrics provide a high-level summary, a class-specific analysis is necessary to understand the clinical utility of the best-performing models, TabPFN and XGBoost. This section examines their performance in predicting each of the three possible objective statuses: 'Validated', 'Completed', and 'Rejected'.

The detailed classification reports for both models are presented in Table 3. Both TabPFN and XGBoost demonstrated good performance on the two majority classes. For the 'Validated' class, both models achieved a high F1-score of 0.85, indicating an ability to correctly identify objectives that require ongoing intervention. Performance on the

'Completed' class was also strong, with TabPFN achieving an F1-score of 0.80 and XGBoost scoring 0.79. These results suggest that both models effectively recognize when a learning objective has been mastered, which is a primary requirement for a clinical decision-support tool.

Table 3 - Class-specific performance of TabPFN and XGBoost on the Test Set

Model	Class	Precision	Recall	F1-score	Support
TabPFN	Validated	0.84	0.86	0.85	391
	Completed	0.79	0.81	0.80	261
	Rejected	0.00	0.00	0.00	17
XGBoost	Validated	0.83	0.86	0.85	391
	Completed	0.79	0.79	0.79	261
	Rejected	0.67	0.12	0.20	17

The primary challenge for both models was predicting the 'Rejected' class, which was severely underrepresented in the dataset. The TabPFN model, despite its high overall performance, failed to correctly identify any instances of the 'Rejected' class, resulting in precision, recall, and F1-scores of 0.00. The XGBoost model showed a marginal ability to identify this class, achieving a recall of 0.12 and a precision of 0.67, leading to a low F1-score of 0.20. This indicates that while XGBoost correctly identified a small fraction of 'Rejected' cases, its predictions for this class were infrequent.

The discriminative ability of the models for each class is further illustrated by the ROC curves, shown in Figure 1 for TabPFN and Figure 2 for XGBoost. The AUC measures a model's ability to distinguish between classes across all classification thresholds. Both models produced high AUC values for the' Validated' and' Completed' classes, ranging from 0.89 to 0.91, confirming their excellent discriminative power for these classes. Notably, even for the 'Rejected' class, where the F1-scores were poor, the AUC values were relatively high (TabPFN: 0.84, XGBoost: 0.80). This suggests that the models' underlying probability scores contain a meaningful signal for distinguishing 'Rejected' cases, even if the default decision threshold does not lead to correct classifications.

Figure 1 - ROC curves for the TabPFN model

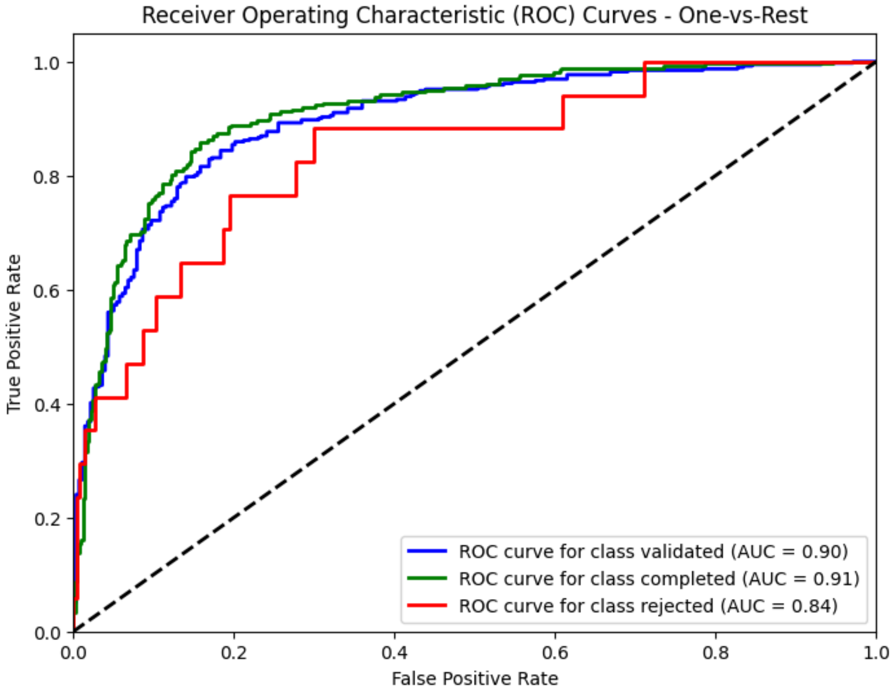
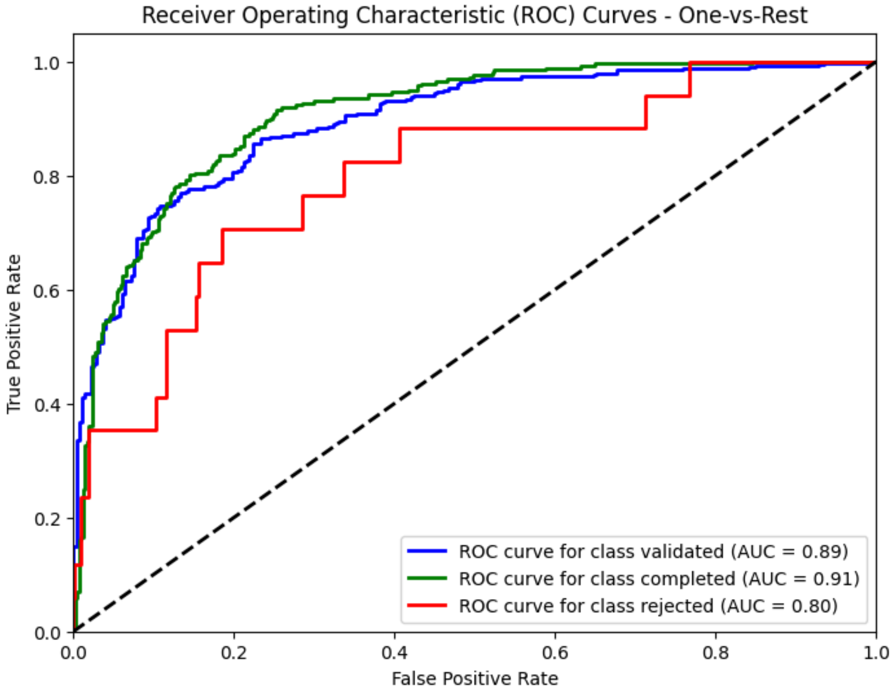


Figure 2 - ROC curves for the XGBoost model

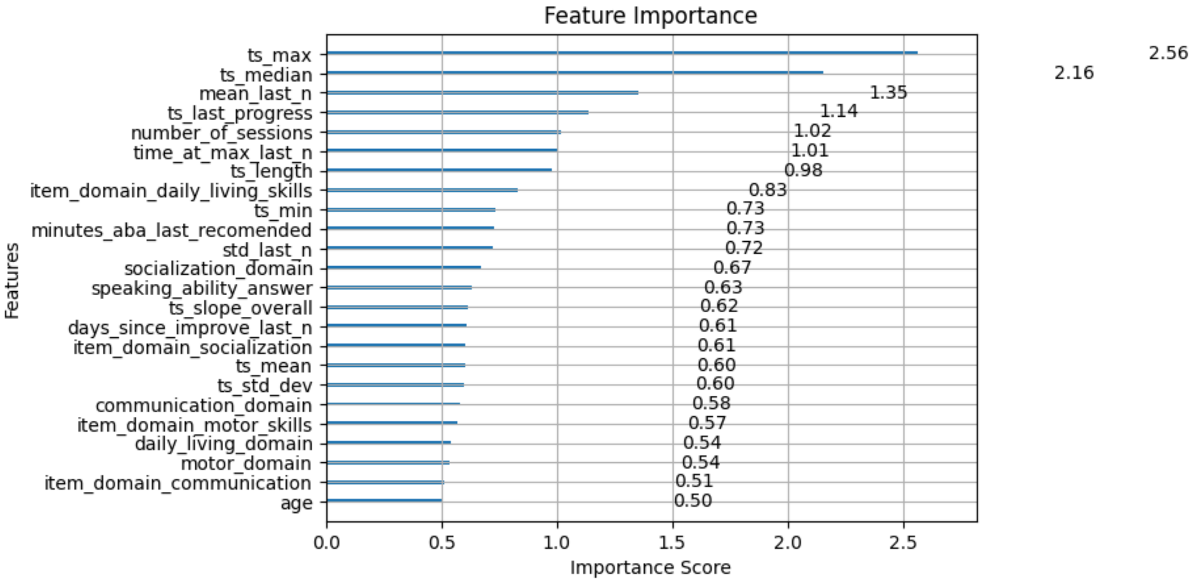


4.4 Feature importance analysis

A feature importance analysis was conducted to understand the key factors driving the predictions of the top-performing models. This analysis provides insight into which aspects of a child's therapeutic progress are most predictive of an objective's future status. In addition to the primary methods detailed below, feature importance was assessed using a standalone Random Forest model and the SelectKBest filter method during the feature selection phase. These preliminary analyses corroborated the main findings, consistently identifying the duration of intervention and engineered time-series features as the most predictive variables (see Appendix B for full results). As the insights were redundant with the more detailed analyses of the primary models, this section focuses on the importance of the feature derived from the XGBoost model and the SHAP method for the TabPFN model.

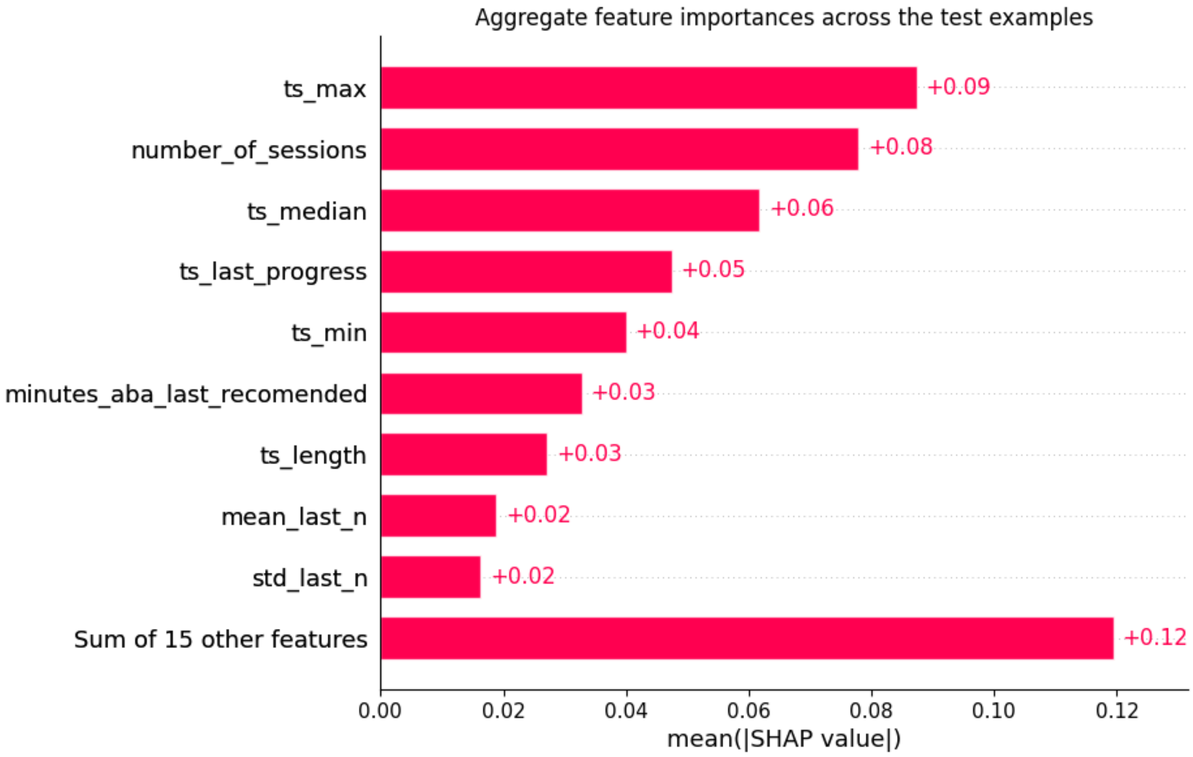
The feature importance ranking from the XGBoost model, using 'gain' as the importance metric, is presented in Figure 3. The analysis reveals that ts_max (the maximum progress score achieved) was the most influential feature. This indicates that, for the XGBoost model, the peak performance reached during an objective's lifecycle is the single most important factor for prediction. This is followed by other engineered time-series features, including ts_median, mean_last_n, and ts_last_progress, as well as the number_of_sessions. The high ranking of these features collectively suggests that a combination of peak performance, recent performance trends, and the overall duration of the intervention drives the model's predictions.

Figure 3 - Feature importance from the XGBoost model



To provide a more detailed, model-agnostic view, a SHAP analysis was conducted on the TabPFN model. Figure 4 presents an aggregate feature importance plot, which summarizes the overall impact of each feature. This plot is generated by taking the mean absolute SHAP value for each feature across all predictions in the test set. It provides a straightforward ranking of which features were most influential in the model's decisions, regardless of the prediction's direction.

Figure 4 - Aggregate feature importance from the TabPFN model (SHAP analysis)



The aggregate analysis in Figure 4 confirms that ts_max is the most impactful feature, followed by number_of_sessions and ts_median. This high-level view reinforces the importance of peak performance and intervention duration.

While this provides a general overview of feature importance, a more detailed understanding is provided by the class-specific SHAP analysis, which explains the impact of each feature on the prediction for each class. The SHAP summary plots for the ‘Validated’, ‘Completed’, and ‘Rejected’ classes are shown in Figures 5, 6, and 7, respectively.

Figure 5 - SHAP summary plot for the 'Validated' class (TabPFN)

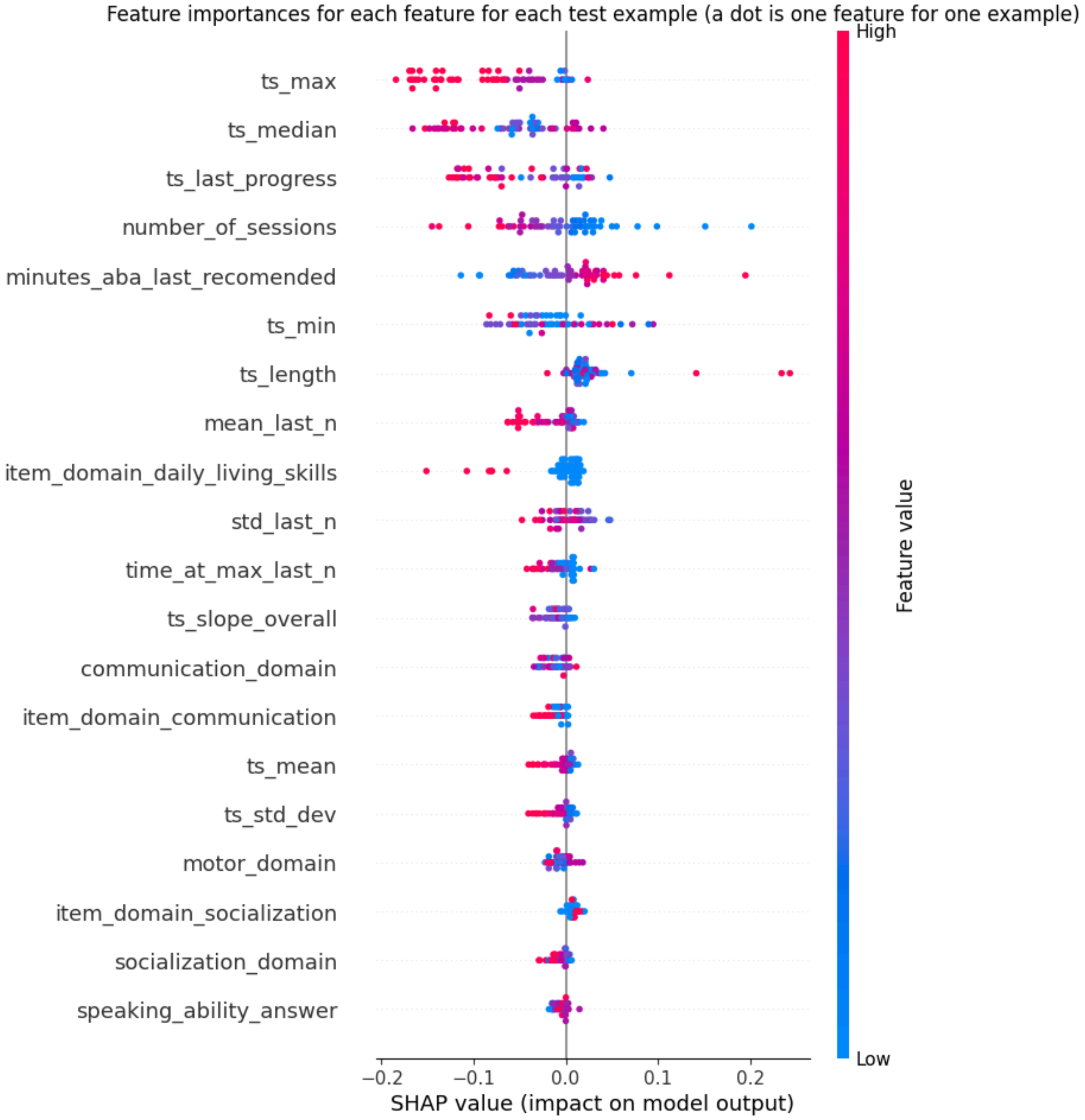


Figure 6 - SHAP summary plot for the 'Completed' class (TabPFN)

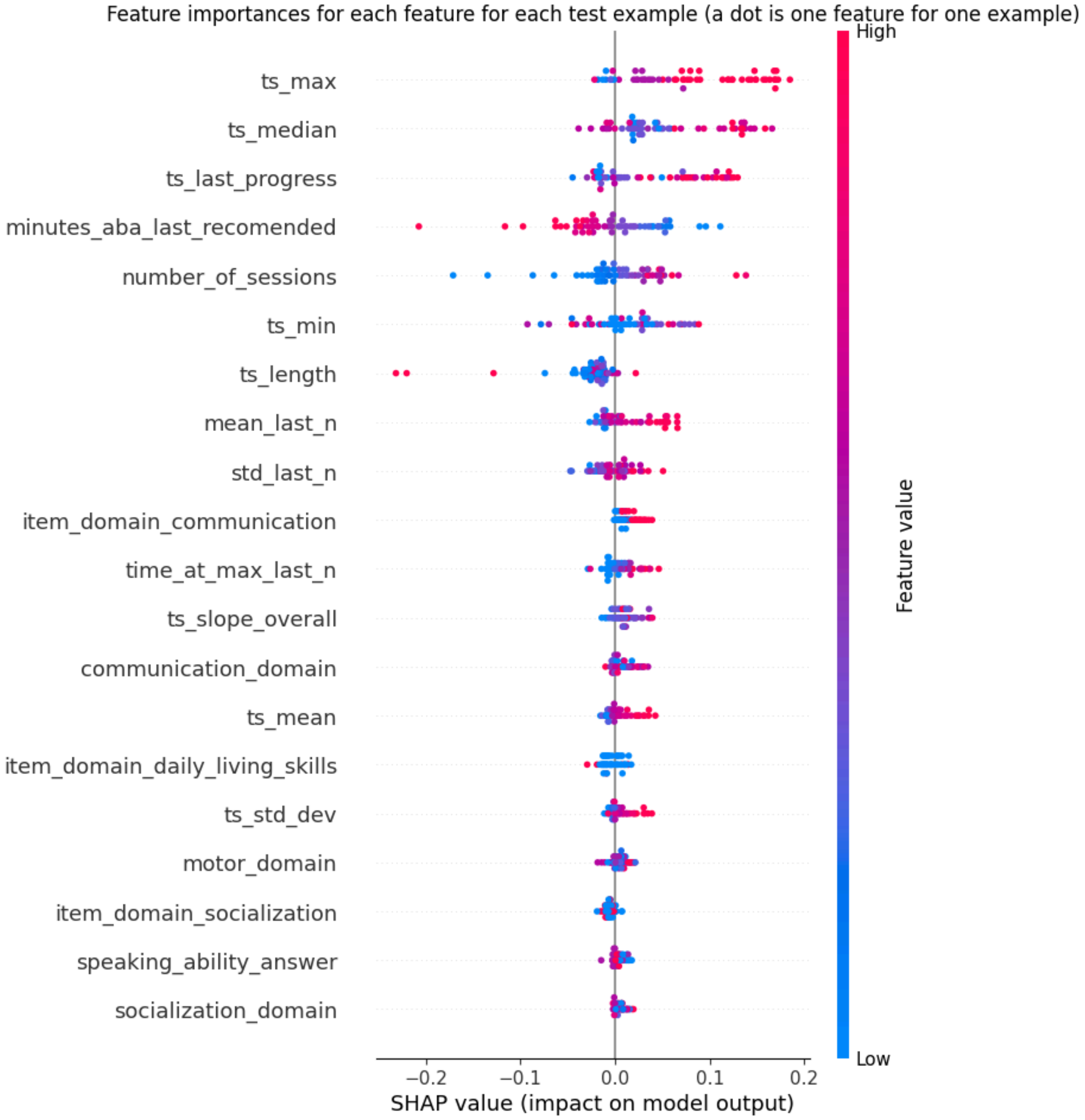
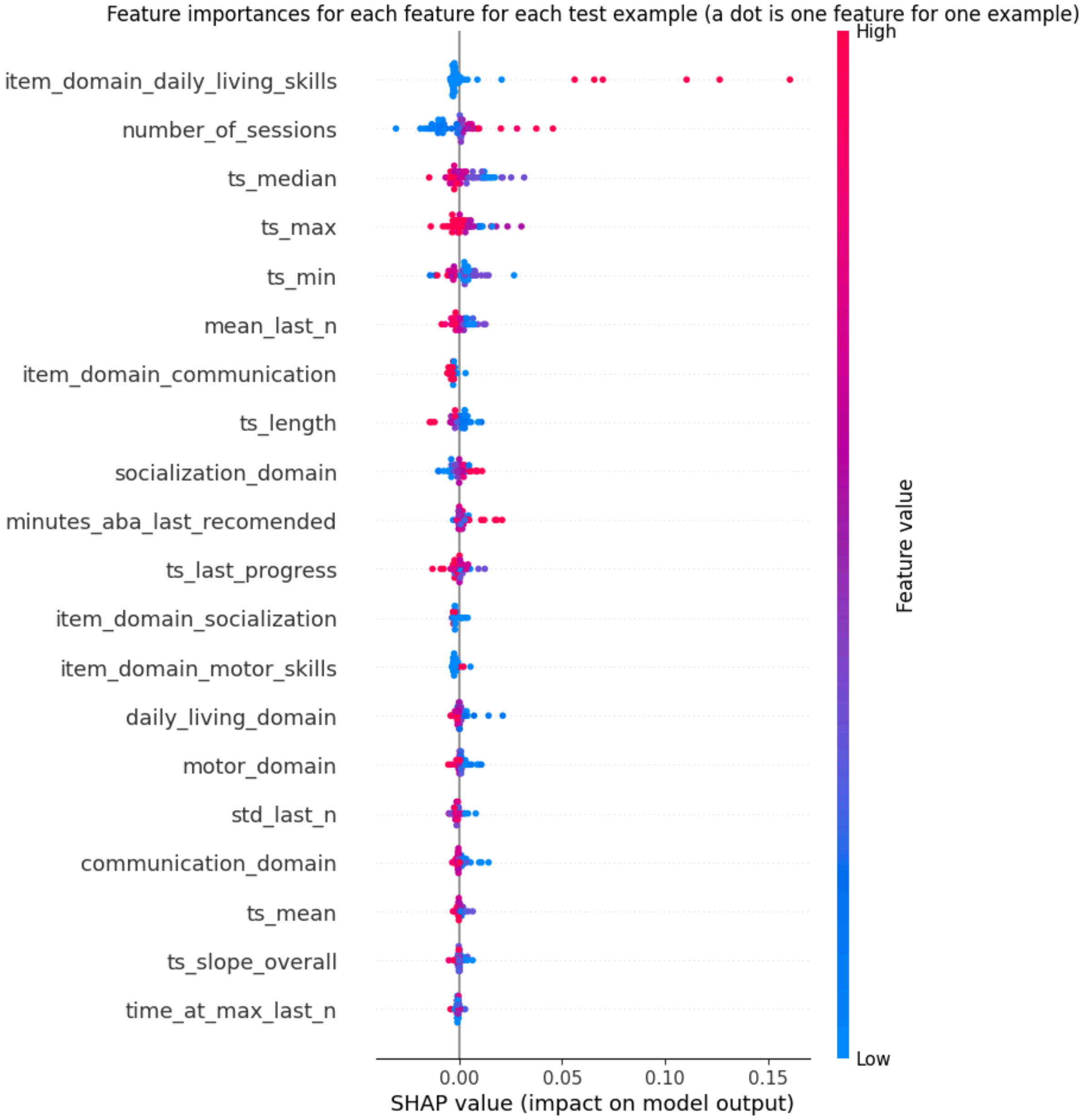


Figure 7 - SHAP summary plot for the 'Rejected' class (TabPFN)



For predicting both 'Validated' and 'Completed' statuses, the most important features were consistent: ts_max (the maximum progress score ever achieved), ts_median (the median score), ts_last_progress (the most recent score), number_of_sessions, and minutes_aba_last_recomended. The direction of their impact aligns with clinical intuition. As seen in Figure 5, low values for these features (e.g., a low maximum score) push the prediction towards 'Validated', suggesting the objective is still in progress. Conversely, as shown in Figure 6, high values for these features strongly push the prediction towards 'Completed', indicating mastery. An exception is the minutes_aba_last_recomended feature, where a higher value also pushes the prediction towards 'Validated', indicating that a greater recommended therapy intensity is associated with objectives that are still actively being worked on.

The feature importance for the 'Rejected' class differed (Figure 7). The most dominant predictor was item_domain_daily_living_skills, where high values for this feature (indicating the objective is in this domain) increased the likelihood of a ‘Rejected’ prediction. Number_of_sessions followed this. A high value for this feature, representing the total duration of a child's therapy, increased the likelihood of a 'Rejected' prediction. This suggests the model learned a clinically relevant pattern: objectives for children with a longer history of intervention are more likely to be discontinued, the same as for objectives within the Daily Living Skills domain. Other features, such as the specific item_domain, also played a role. Collectively, these analyses show that the models based their predictions on logical and interpretable patterns in the therapy data, primarily related to the duration of the intervention and the level of performance achieved over time.

To synthesize these findings, Table 4 provides a consolidated view of the top-ranked features across all four evaluation methods explored in this project. The table reveals a certain degree of consensus among the different techniques.

Table 4 - Comparison of top 10 feature importance rankings across models

Feature	SelectKBest (F-score)	Random Forest (Importance)	XGBoost (Importance)	TabPFN (Mean | SHAP Value)
ts_max	146 (6)	—	2.56 (1)	0.08 (1)
ts_median	214 (3)	0.08 (3)	2.16 (2)	0.07 (2)
ts_last_progress	175 (4)	0.06 (6)	1.14 (4)	0.05 (3)
number_of_sessions	82 (7)	0.10 (1)	1.02 (5)	0.04 (4)
minutes_aba_last_recomended	56 (8)	—	0.73 (10)	0.04 (5)
ts_min	—	—	0.73 (9)	0.04 (6)
ts_length	—	0.07 (4)	0.98 (7)	0.03 (7)
mean_last_n	271 (1)	0.09 (2)	1.35 (3)	0.02 (8)
item_domain_daily_living_skills	31 (10)	—	0.83 (8)	0.02 (9)
std_last_n	—	0.05 (9)	—	0.01 (10)
communication_domain	33 (9)	—	—	—
ts_mean	222 (2)	0.07 (5)	—	—
ts_slope_overall	—	0.05 (8)	—	—
socialization_domain	—	0.04 (10)	—	—
time_at_max_last_n	169 (5)	—	1.01 (6)	—
ts_std_dev	—	0.05 (7)	—	—
motor_domain	—	—	—	—

Note: The table displays the top 10 features ranked by each method. The primary value (F-score, Gini importance, XGBoost importance score, or mean absolute SHAP value) is shown, with the feature's rank for that specific method in parentheses. The features are ordered by rank in the TabPFN (SHAP) analysis, as it is the primary model. A dash (—) indicates that the feature was not in the top 10 for that method.

Features such as ts_median, ts_last_progress, number_of_sessions, and mean_last_n appear in the top 10 ranking for all the methods, regardless of whether the evaluation is model-agnostic (SHAP), model-specific (XGBoost, Random Forest), or purely statistical (SelectKBest). This agreement shows the robustness of these features as predictors. While there is a consensus on this core set of features, the different methods assign varying priorities. For instance, the XGBoost and TabPFN models had ts_max in the top position, emphasizing peak performance. In contrast, the Random Forest model prioritizes the overall duration through number_of_sessions. Despite these differences, the consistent appearance of these variables reinforces the primary conclusion of this analysis: the most critical factors for predicting an objective's status are the duration of the intervention and key statistical indicators of the child's learning trajectory.

5 DISCUSSION

This chapter interprets the results presented in Chapter 4, connecting the empirical findings to the broader clinical context and existing literature. The discussion is structured to summarize the principal findings, then interpret their clinical significance, compare them with previous research, and outline the study's implications, limitations, and directions for future work.

5.1 Summary of principal findings

This research aimed to develop and evaluate machine learning (ML) models for predicting the status progression of learning objectives within ABA Individualized Education Plans for children with ASD, classifying them as 'Validated', 'Completed', or 'Rejected' based on longitudinal therapy data. The principal finding is that ML can effectively predict the status of these objectives using a feature set grounded in clinical practice. The pre-trained transformer model, TabPFN, and the gradient boosting model, XGBoost, demonstrated superior performance compared to baseline models, such as SVM, KNN, and LSTM. Both top-performing models achieved high overall accuracy (0.82 and 0.81, respectively), weighted F1-scores (0.81), and weighted OvR AUC-ROC scores of approximately 0.90, indicating a strong ability to discriminate between the different objective statuses.

A detailed analysis revealed that TabPFN and XGBoost effectively classified the two majority classes, achieving F1-scores of 0.85 for 'Validated' objectives and 0.80 and 0.79 for 'Completed' objectives, respectively. This highlights their utility for the most common clinical decisions: determining if an objective requires continued work or has been mastered. The primary challenge for all models was predicting the 'Rejected' class, representing a distinct minority of the data (2.5%). The TabPFN model failed to correctly identify any instances, and the XGBoost model performed only marginally, with an F1-score of 0.20. Despite these low classification scores, the ROC analysis revealed that both models generated a meaningful discriminative signal for the 'Rejected' class, with AUC scores of 0.84 and 0.80 for TabPFN and XGBoost, respectively.

The feature importance analysis provided clear insights into the predictive drivers of the models. It highlighted the duration of the intervention and the engineered time-series features as the most influential predictors. Across multiple evaluation techniques, the duration of the intervention (number_of_sessions) and features engineered from the time-series data, such as ts_max, ts_median, ts_last_progress, and mean_last_n, consistently ranked as the most influential variables. These findings suggest that ML can effectively model the progression toward objective mastery based on clinically relevant patterns related to the duration and trajectory of a child's therapeutic progress. Still, the prediction of rare outcomes like discontinuation remains a challenge.

5.2 Interpretation of findings in the clinical context

The quantitative results of this study gain significance when interpreted within the practical context of ABA therapy and IEP management. The models' performance, particularly that of TabPFN and XGBoost, reflects an ability to learn and replicate key aspects of the clinical decision-making process from longitudinal data. The high performance on the classification of 'Validated' and 'Completed' objectives, with F1-scores of 0.85 and approximately 0.80, respectively, aligns with the most frequent and fundamental decisions made by supervising therapists. Whether to continue working on a skill or confirm its mastery, this binary decision forms the core of the IEP review cycle, relying heavily on the clinical judgment of senior therapists (Ghafghazi et al., 2021). Clinicians described this process as time-consuming, which involves analyzing weekly progress checks and in-session observations to evaluate a child's learning trajectory. This aligns with the motivation of this research to address the time-consuming nature of manual IEP review. The models effectively automate a part of this analytical process, demonstrating that the engineered features successfully captured the signals clinicians use to make these judgments. Automating a preliminary assessment of these high-volume statuses could allow clinicians to focus on more complex cases and higher-level planning.

The feature importance analysis further grounds the models' logic in clinical reality. The consistent prominence of features such as ts_max, ts_median, ts_last_progress, number_of_sessions, and mean_last_n confirms the initial hypothesis that the model's predictions would be based on clinically intuitive patterns. The duration of the intervention (number_of_sessions, ts_length) and the trajectory of recent performance (ts_last_progress, mean_last_n) are critical factors for clinicians. As one supervising psychologist noted, an objective that remains active for over three months without mastery is a signal for review. The model learned this pattern, identifying intervention and objective durations as predictors. The SHAP analysis confirmed that consistent, high performance pushes predictions toward 'Completed', mirroring a clinician's judgment.

Interpreting the model's performance on the 'Rejected' class requires a more detailed clinical perspective. While the low F1-scores indicate a failure in direct classification at standard thresholds, the high AUC scores (0.84 for TabPFN) are clinically meaningful. This discrepancy suggests that although the model hesitates to assign the 'Rejected' label due to its rarity, its underlying probability scores effectively discriminate these cases from others. A practical implementation of this model in a clinical setting would not be to automate the classification of 'Rejected' objectives, but to use the model's probabilistic output to flag objectives for review. For instance, a rule could be set to alert a senior therapist whenever the predicted probability for the 'Rejected' class for a given objective exceeds a certain threshold (e.g., 0.3), even if it is not the most likely class. This would draw attention to objectives showing signs of stagnation, a task that supervisors perform manually and can be time-consuming across a large caseload. The fact that number_of_sessions was the second most dominant predictor for this class (Figure 6) confirms that the model learned a clinical heuristic. Therapists noted that objectives active for a prolonged period without significant progress are often discontinued. The model's reliance on number_of_sessions, a proxy for the total duration of a child's therapy, reflects this pattern: the longer a child is in treatment, the higher the chance that a challenging, long-standing objective will be rejected. Furthermore, the influence of the item_domain_daily_living_skills feature suggests the model identified a pattern where objectives in this domain are more likely to be discontinued than others, a factor clinicians might consider implicitly.

Ultimately, the findings suggest the model's utility as a decision-support tool rather than an autonomous decision-maker. It automates the quantitative analysis of progress data, a task clinicians describe as laborious, involving synthesizing data from multiple therapists, session notes, and parental reports. By providing data-driven predictions on objective status and flagging at-risk objectives, such a system could increase the efficiency and consistency of the IEP review process. This would allow senior therapists to dedicate more time to high-value clinical activities that the model cannot perform, such as direct observation, mentoring junior therapists, and complex clinical reasoning.

5.3 Comparison with the literature

The findings of this project align with the broader literature that highlights the potential of ML in managing ASD (Pandya et al., 2024; Rêgo & Araújo-Filho, 2024). However, this research addresses a specific gap by focusing on the dynamic, objective-level management of IEPs. Most existing research in this domain has concentrated on initial diagnosis and screening (Abbas et al., 2020; Shinde & Patil, 2023), on making upfront decisions about treatment, such as classifying the appropriate ABA plan type (Maharjan et al., 2023), or recommending initial treatment goals (Kohli et al., 2022). In contrast, this study addresses the ongoing, iterative process of managing learning objectives. While Linstead et al. (2017) also used ML to predict learning outcomes, their work focused on the relationship between overall treatment intensity and mastery. They provide a model that confirms that intensive therapy works, while this project offers a model designed to help manage how it works, one objective at a time. This work extends theirs by moving from a general outcome score to a specific, actionable status classification for individual learning targets, which is more directly aligned with the clinical workflow of IEP management.

The challenges encountered in this study are also consistent with those documented in the literature. The difficulty in predicting the minority 'Rejected' class is a practical example of the class imbalance problem, which is frequently cited as a significant challenge in applying ML to healthcare and ASD-related data (Thabtah, 2019; Y. Yang et al., 2024). The finding that this class imbalance hindered the performance of all models, including the best-performing ones, reinforces the idea that specialized techniques are often necessary to handle the skewed data distributions common in clinical datasets (Kaur et al., 2019).

From a methodological standpoint, transforming the longitudinal data into a tabular format with engineered features aligns with established practices for applying powerful tree-based models like XGBoost to time-series data (Gorishniy et al., 2021). The performance of this approach confirms its viability for this type of clinical data. Furthermore, the application of TabPFN (Hollmann et al., 2023, 2025), a recent transformer-based model, to this clinical problem is novel. It demonstrates the potential of state-of-the-art tabular data models to effectively analyze complex, real-world healthcare data without extensive hyperparameter tuning. This is a significant advantage in clinical settings where computational resources may be limited. This study, therefore, contributes to the literature by demonstrating a feasible and effective data-driven methodology for a specific, practical, and previously unaddressed problem in the day-to-day management of ABA therapy.

Another methodological finding of this study was that applying external feature selection techniques did not improve the predictive performance of the TabPFN model. While counterintuitive compared to traditional ML workflows, this result is consistent with the literature describing the model's architecture.

The resilience of TabPFN to a high-dimensional feature space, and the lack of improvement from feature selection, is attributable to its self-attention mechanism, which allows the model to dynamically weigh the importance of different features for each prediction (Hollmann et al., 2023, 2025). It is plausible that the pre-selection of features using univariate statistical tests removed variables that, while weak individually, provided valuable contextual information when combined with other features. Since TabPFN is pre-trained to be robust to noisy and irrelevant data, it can effectively ignore non-informative features, rendering an external feature selection step unnecessary and potentially detrimental due to information loss.

5.4 Implications and contributions of the study

The findings of this study have practical and methodological implications for the management of ABA interventions for children with ASD. The primary practical implication is the demonstrated feasibility of using ML as a clinical decision-support tool. By automating the analysis of longitudinal progress data, a task that currently consumes valuable clinician time, the developed model could significantly increase the efficiency of the IEP review process. This would allow senior therapists to allocate more time to direct clinical care, complex decision-making, and mentoring junior staff, which require human expertise (Ghafghazi et al., 2021). Furthermore, by providing an objective, data-driven assessment, the model could help standardize the process of reviewing IEPs, potentially reducing variability in clinical judgments across different therapists and supporting the scalability of high-quality intervention services to a growing population (Grosvenor et al., 2024).

From a methodological perspective, this research provides a validated framework for applying tabular ML models to longitudinal clinical data. The feature engineering approach successfully transformed raw time series of progress scores into descriptive features that enabled high-performing models like TabPFN and XGBoost to learn clinically relevant patterns. This methodological approach can guide future studies that want to analyze similar data in this or other medical fields. The strong performance of the pre-trained TabPFN model, in particular, suggests that modern transformer-based architectures offer a promising avenue for developing practical clinical tools without the need for extensive, task-specific hyperparameter tuning.

Ultimately, implementing such a decision-support tool must be guided by ethical principles of responsible AI in healthcare. It is crucial to ensure that the tool is used to augment, not replace, the clinical expertise of therapists, and that its recommendations are transparent and explainable to clinicians and the families of the children receiving care. The system can prompt timely and focused clinical review by flagging objectives nearing mastery or showing signs of stagnation. This supports a more proactive and data-informed approach to IEP management, aligning with the principles of developing responsible and effective AI-based prediction models for healthcare (de Hond et al., 2022).

5.5 Limitations and future research

This study has limitations that should be considered when interpreting the findings and that pave the way for future research. First, the model development and evaluation data were sourced from a single Brazilian healthcare provider. While this provided a consistent and high-quality dataset, it may limit the generalizability of the findings. Clinical practices, patient demographics, and data recording standards can vary across institutions and geographical regions. Therefore, the model's performance may not be directly transferable to other clinical settings without further validation.

Second, the primary challenge encountered was the model's difficulty in classifying the 'Rejected' class due to extreme class imbalance. The F1-scores for this minority class were poor, indicating that the model is unreliable for autonomously identifying objectives that should be discontinued. Preliminary experiments with standard resampling techniques like ADASYN and SMOTE-ENN did not yield performance improvements on the test set; therefore, the problem of handling imbalanced data remains a critical area for improvement.

Third, the study is retrospective, relying on historically collected clinical data. This design means that the model was developed based on patterns in past interventions, and it cannot account for unobserved variables that may have influenced clinical decisions, such as changes in a child's family situation or a therapist's specific clinical judgment. The feature set was limited to the data available in the clinical records. It did not include potentially influential factors such as therapist experience, parental engagement levels, or the specific materials used in therapy, which were highlighted during interviews as important contextual variables.

These limitations inform several directions for future research. The immediate next step should be to validate the current model on external datasets from different providers to assess its robustness and generalizability. To address the class imbalance problem, future work could explore more advanced methods designed explicitly for imbalanced datasets, such as cost-sensitive learning or hybrid resampling techniques (Y. Yang et al., 2024).

Furthermore, future research should enrich the feature set by incorporating more detailed contextual data, such as variables related to therapist expertise and parental involvement. This would likely require a prospective study designed to collect this information. Such a study represents a step in validating the model, as it would allow for an evaluation of its real-world impact on clinical workflows, the efficiency of decision-making, and ultimately, patient outcomes. This approach is consistent with the recommended 'impact assessment phase' for the responsible implementation of AI in healthcare (de Hond et al., 2022). Finally, further research is needed on how to best integrate such a tool into the clinical workflow. This involves focusing on user interface design and determining how to present the model's probabilistic outputs to clinicians in a way that effectively supports, rather than overrides, their expert judgment.

6 CONCLUSION

This chapter summarizes the key findings of the research and discusses their implications for the field. It synthesizes the principal results, interprets their clinical significance, and outlines the study's contributions, limitations, and directions for future work.

6.1 Summary of research and key findings

This study addressed the significant clinical challenge of managing and adapting Individualized Education Plans (IEPs) for children with Autism Spectrum Disorder (ASD), a process traditionally reliant on time-intensive manual review and subjective clinical judgment. The central objective was to investigate the feasibility of applying machine learning (ML) to predict the status progression of Applied Behavior Analysis (ABA) learning objectives, 'Validated', 'Completed', or 'Rejected', based on longitudinal therapy data, thereby creating a foundation for an effective clinical decision-support tool.

The principal finding of this research is a clear demonstration that ML models can effectively automate and support this clinical decision-making process by learning from a feature set chosen for its clinical relevance. Through a comprehensive methodology involving data preprocessing, feature engineering, and comparative model evaluation on a longitudinal clinical dataset, two advanced models, the pre-trained transformer TabPFN and the gradient-boosting model XGBoost, emerged as particularly effective. These models achieved high overall predictive accuracy and demonstrated a robust ability to distinguish between objectives that require ongoing intervention ('Validated') and those that have been mastered ('Completed').

A key finding was the challenge posed by the severe class imbalance of 'Rejected' objectives. While direct classification of this minority class proved difficult for all models, the high Area Under the Curve (AUC) scores indicated that the models successfully learned a meaningful discriminative signal. This suggests their utility as a flagging mechanism for objectives at risk of stagnation. Furthermore, the feature importance analysis consistently revealed that the most influential predictors were the duration of the intervention (number_of_sessions) and engineered features quantifying the objective's recent performance trajectory (ts_median, ts_last_progress, mean_last_n). This alignment with clinical heuristics confirms that the models learned clinically relevant patterns from the therapy data.

6.2 Principal contributions and implications

The principal contribution of this project is the development and validation of a novel ML framework for a critical aspect of ABA therapy: the objective-level management of ABA IEPs, a critical and previously unaddressed challenge in the application of ML to ASD intervention. This work moves beyond the more common applications of ML in ASD research, which typically focus on initial diagnosis or static outcome prediction, to address the dynamic, iterative process of clinical intervention management. By demonstrating the effectiveness of a pre-trained transformer model (TabPFN) in this context, this work provides a practical and scalable solution that can enhance the efficiency and consistency of clinical decision-making.

This research provides a proof-of-concept for a decision-support tool designed to automate the analysis of objective progression. This directly addresses the significant efficiency and scalability challenges inherent in manual review, offering a pathway to standardize clinical judgments and enable senior therapists to focus on higher-value tasks such as complex case analysis and mentorship (Ghafghazi et al., 2021).

Methodologically, the project contributes a validated framework that bridges clinical domain knowledge with advanced ML. The successful application of the pre-trained TabPFN model is a novel contribution to this specific clinical domain. It demonstrates that modern transformer-based architectures can achieve high performance without the extensive hyperparameter tuning typically required, an advantage for practical deployment (de Hond et al., 2022). By bridging a specific clinical need with an advanced methodological solution, this study provides an actionable tool concept designed to augment, not replace, clinical expertise, thereby supporting a more proactive, consistent, and data-informed approach to IEP management.

6.3 Limitations and future directions

While this research provides valuable insights, limitations must be acknowledged, which also highlight possible future work. The study's reliance on data from a single Brazilian healthcare provider may limit the generalizability of the findings to different clinical and cultural contexts. A primary methodological challenge was the extreme class imbalance of the 'Rejected' class, which hindered the models' predictive performance for this rare but clinically significant outcome. Furthermore, the study's retrospective nature meant the analysis was constrained to the data available in clinical records, excluding potentially influential contextual factors such as therapist experience or parental engagement.

These limitations directly inform next steps. The priority is to validate the current models on external datasets from diverse providers to assess their robustness and generalizability. Future work should also explore more advanced methods designed explicitly for imbalanced data, such as cost-sensitive learning or hybrid resampling techniques (Y. Yang et al., 2024), to improve the prediction of minority classes. Ultimately, a prospective study is necessary to collect richer contextual data and to conduct an 'impact assessment phase' (de Hond et al., 2022). Such a study would evaluate the model's real-world influence on clinical workflows, decision-making efficiency, and patient outcomes. Further research is also needed on the optimal integration of such a tool into the clinical workflow, focusing on user interface design and determining how to present the model's probabilistic outputs to clinicians in a way that effectively supports, rather than overrides, their expert judgment.

6.4 Concluding remarks

This research successfully demonstrated the viability of applying advanced ML models to the dynamic management of ABA IEPs for children with ASD. The study established that models like TabPFN and XGBoost can accurately predict the progression of learning objectives by learning clinically relevant patterns from longitudinal therapy data. This work provides a proof-of-concept for a clinical decision-support tool intended to augment, not replace, the therapists' expertise. Increasing the efficiency and consistency of the IEP review process, this data-driven approach paves the way for more scalable, responsive, and personalized interventions. Ultimately, this study represents a step toward improving therapeutic outcomes for children with ASD.

REFERENCES

ABBAS, Halim *et al.* Multi-modular AI Approach to Streamline Autism Diagnosis in Young Children. **Scientific Reports**, v. 10, n. 1, p. 5014, 19 mar. 2020.

AMERICAN PSYCHIATRIC ASSOCIATION. **Diagnostic and Statistical Manual of Mental Disorders: Dsm-5**. 5th edition ed. Washington: American Psychiatric Publishing, 2013.

AMIN, Adnan *et al.* Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. **IEEE Access**, v. 4, p. 7940–7957, 2016.

AWAD, Mariette; KHANNA, Rahul. Support Vector Machines for Classification. *In*: AWAD, Mariette; KHANNA, Rahul (Orgs.). **Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers**. Berkeley, CA: Apress, 2015. p. 39–66.

BAGNALL, Anthony *et al.* The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. **Data Mining and Knowledge Discovery**, v. 31, n. 3, p. 606–660, 1 maio 2017.

BATISTA, Gustavo E. A. P. A.; PRATI, Ronaldo C.; MONARD, Maria Carolina. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explor. Newsl.**, v. 6, n. 1, p. 20–29, 1 jun. 2004.

BISHOP, Christopher M.; NASRABADI, Nasser M. **Pattern recognition and machine learning**. *[S.l.]*: Springer, 2006. v. 4

BRADLEY, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**, v. 30, n. 7, p. 1145–1159, 1 jul. 1997.

BREEJEN, Felix den *et al.* **Fine-tuned In-Context Learning Transformers are Excellent Tabular Data Classifiers**. arXiv, , 23 jan. 2025. Disponível em: <http://arxiv.org/abs/2405.13396>. Acesso em: 1 abr. 2025

BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 1 out. 2001.

BROWN, Tom *et al.* Language Models are Few-Shot Learners. *In*: Curran Associates, Inc., 2020. Disponível em: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>. Acesso em: 11 abr. 2025

CABRAL, Thales W. *et al.* Analysis of variance combined with optimized gradient boosting machines for enhanced load recognition in home energy management systems. **Sensors**, v. 24, n. 15, p. 4965, 2024.

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 1 jun. 2002.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. *In*: : KDD ’16.New York, NY, USA: Association for Computing Machinery, 13 ago. 2016. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 7 abr. 2025

COOPER, John; HERON, Timothy; HEWARD, William. **Applied Behavior Analysis**. 3rd edition ed. Hoboken, New Jersey: Pearson, 2019.

COUNCIL OF AUTISM SERVICE PROVIDERS. **Applied behavior analysis practice guidelines for the treatment of Autism Spectrum Disorder: Guidance for healthcare funders, regulatory bodies, service providers, and consumers**. Disponível em: <https://www.casproviders.org/asd-guidelines>. Acesso em: 2 abr. 2025.

DAWSON, Geraldine *et al.* Randomized, Controlled Trial of an Intervention for Toddlers With Autism: The Early Start Denver Model. **Pediatrics**, v. 125, n. 1, p. e17–e23, 1 jan. 2010.

DAWSON, Geraldine; BERNIER, Raphael. A quarter century of progress on the early detection and treatment of autism spectrum disorder. **Development and Psychopathology**, v. 25, n. 4pt2, p. 1455–1472, nov. 2013.

DE HOND, Anne A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. **npj Digital Medicine**, v. 5, n. 1, p. 1–13, 10 jan. 2022.

DU, Yao *et al.* In-Home Speech and Language Screening for Young Children: A Proof-of-Concept Study Using Interactive Mobile Storytime. **AMIA Summits on Translational Science Proceedings**, v. 2019, p. 722–731, 6 maio 2019.

DUARTE, Cintia Perez. **Estratégias da Análise do Comportamento Aplicada Para Pessoas com Transtornos do Espectro do Autismo**. 1ª edição ed. *[S.l.]*: APGIQ, 2018.

ESLING, Philippe; AGON, Carlos. Time-series data mining. **ACM Comput. Surv.**, v. 45, n. 1, p. 12:1-12:34, 7 dez. 2012.

FARSHIDVARD, A.; HOOSHMAND, F.; MIRHASSANI, S. A. A novel two-phase clustering-based under-sampling method for imbalanced classification problems. **Expert Systems with Applications**, v. 213, p. 119003, 1 mar. 2023.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern Recognition Letters**, ROC Analysis in Pattern Recognition. v. 27, n. 8, p. 861–874, 1 jun. 2006.

FERNÁNDEZ, Alberto *et al.* **Learning from Imbalanced Data Sets**. Cham: Springer International Publishing, 2018.

FRANÇOIS, Damien *et al.* Resampling methods for parameter-free and robust feature selection with mutual information. **Neurocomputing**, v. 70, n. 7–9, p. 1276–1288, mar. 2007.

GARIKIPATI, Anurag *et al.* Clinical Outcomes of a Hybrid Model Approach to Applied Behavioral Analysis Treatment. **Cureus**, v. 15, n. 3, p. e36727, mar. 2023.

GHAFGHAZI, Shadi *et al.* AI-Augmented Behavior Analysis for Children With Developmental Disabilities: Building Toward Precision Treatment. **IEEE Systems, Man, and Cybernetics Magazine**, v. 7, n. 4, p. 4–12, out. 2021.

GNIP, Peter; VOKOROKOS, Liberios; DROTÁR, Peter. Selective oversampling approach for strongly imbalanced data. **PeerJ Computer Science**, v. 7, p. e604, 18 jun. 2021.

GORISHNIY, Yury *et al.* Revisiting Deep Learning Models for Tabular Data. *In*: Curran Associates, Inc., 2021. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html>. Acesso em: 8 abr. 2025

GROSVENOR, Luke P. *et al.* Autism Diagnosis Among US Children and Adults, 2011-2022. **JAMA Network Open**, v. 7, n. 10, p. e2442218, 30 out. 2024.

GUYON, Isabelle; ELISSEEFF, André. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research**, v. 3, n. Mar, p. 1157–1182, 2003.

HARING, Thomas G.; KENNEDY, Craig H. Units of Analysis in Task-Analytic Research. **Journal of Applied Behavior Analysis**, v. 21, n. 2, p. 207–215, 1988.

HASTIE, Trevor; TIBSHIRANI, Robert. Classification by pairwise coupling. **Advances in neural information processing systems**, v. 10, 1997.

HE, Haibo *et al.* ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In*: 2008 IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE). **2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)**. jun. 2008. Disponível em: <https://ieeexplore.ieee.org/abstract/document/4633969>. Acesso em: 13 abr. 2025

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, nov. 1997.

HOLLMANN, Noah *et al.* **TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second**. arXiv, , 16 set. 2023. Disponível em: <http://arxiv.org/abs/2207.01848>. Acesso em: 9 abr. 2025

HOLLMANN, Noah *et al.* Accurate predictions on small data with a tabular foundation model. **Nature**, v. 637, n. 8045, p. 319–326, jan. 2025.

HOO, Shi Bin *et al.* **The Tabular Foundation Model TabPFN Outperforms Specialized Time Series Forecasting Models Based on Simple Features**. arXiv, , 9 jan. 2025. Disponível em: <http://arxiv.org/abs/2501.02945>. Acesso em: 10 abr. 2025

HOWARD, Jane S. *et al.* A comparison of intensive behavior analytic and eclectic treatments for young children with autism. **Research in Developmental Disabilities**, v. 26, n. 4, p. 359–383, 1 jul. 2005.

HÜSKEN, Michael; STAGGE, Peter. Recurrent neural networks for time series classification. **Neurocomputing**, v. 50, p. 223–235, 1 jan. 2003.

HYDE, Kayleigh K. *et al.* Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review. **Review Journal of Autism and Developmental Disorders**, v. 6, n. 2, p. 128–146, 1 jun. 2019.

ISMAIL FAWAZ, Hassan *et al.* Deep learning for time series classification: a review. **Data**

Mining and Knowledge Discovery, v. 33, n. 4, p. 917–963, 1 jul. 2019.

JOHNSON, Justin M.; KHOSHGOFTAAR, Taghi M. Survey on deep learning with class imbalance. **Journal of Big Data**, v. 6, n. 1, p. 27, 19 mar. 2019.

KAMALOV, Firuz *et al.* Forward feature selection: empirical analysis. **Journal of Intelligent Systems and Internet of Things**, n. Issue 1, p. 44–54, 1 jan. 2024.

KAMPOURAKI, Argyro; MANIS, George; NIKOU, Christophoros. Heartbeat Time Series Classification With Support Vector Machines. **IEEE Transactions on Information Technology in Biomedicine**, v. 13, n. 4, p. 512–518, jul. 2009.

KAUR, Harsurinder; PANNU, Husanbir Singh; MALHI, Avleen Kaur. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. **ACM Comput. Surv.**, v. 52, n. 4, p. 79:1-79:36, 30 ago. 2019.

KHUSHI, Matloob *et al.* A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. **IEEE Access**, v. 9, p. 109960–109975, 2021.

KOHAVI, Ron; JOHN, George H. Wrappers for feature subset selection. **Artificial Intelligence**, Relevance. v. 97, n. 1, p. 273–324, 1 dez. 1997.

KOHLI, Manu *et al.* Machine learning-based ABA treatment recommendation and personalization for autism spectrum disorder: an exploratory study. **Brain Informatics**, v. 9, n. 1, p. 16, 25 jul. 2022.

KOU, Yan *et al.* Network- and attribute-based classifiers can prioritize genes and pathways for autism spectrum disorders and intellectual disability. **American Journal of Medical Genetics Part C: Seminars in Medical Genetics**, v. 160C, n. 2, p. 130–142, 2012.

KRAMER, Oliver. K-Nearest Neighbors. *In*: KRAMER, Oliver (Org.). **Dimensionality Reduction with Unsupervised Nearest Neighbors**. Berlin, Heidelberg: Springer, 2013. p. 13–23.

KUMAR, Chandan Jyoti; DAS, Priti Rekha. The diagnosis of ASD using multiple machine learning techniques. **International Journal of Developmental Disabilities**, 2 nov. 2022.

LINSTEAD, Erik *et al.* An Application of Neural Networks to Predicting Mastery of Learning Outcomes in the Treatment of Autism Spectrum Disorder. *In*: 2015 IEEE 14TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS (ICMLA). **2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)**. dez. 2015. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7424348>. Acesso em: 11 abr. 2025

LINSTEAD, Erik *et al.* Intensity and Learning Outcomes in the Treatment of Children With Autism Spectrum Disorder. **Behavior Modification**, v. 41, n. 2, p. 229–252, 1 mar. 2017.

LIU, Xu-Ying; WU, Jianxin; ZHOU, Zhi-Hua. Exploratory Undersampling for Class-Imbalance Learning. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 39, n. 2, p. 539–550, abr. 2009.

LORD, Catherine *et al.* Autism Spectrum Disorders. **Neuron**, v. 28, n. 2, p. 355–363, 1 nov.

2000.

LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. *In*: Curran Associates, Inc., 2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>. Acesso em: 11 abr. 2025

MA, Junwei *et al.* **In-Context Data Distillation with TabPFN**. arXiv, , 10 fev. 2024. Disponível em: <http://arxiv.org/abs/2402.06971>. Acesso em: 1 abr. 2025

MAENNER, Matthew J. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. **MMWR. Surveillance Summaries**, v. 72, 2023.

MAHARJAN, Jenish *et al.* Machine learning determination of applied behavioral analysis treatment plan type. **Brain Informatics**, v. 10, n. 1, p. 7, 2 mar. 2023.

MALHOTRA, Pankaj *et al.* **LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection**. arXiv, , 11 jul. 2016. Disponível em: <http://arxiv.org/abs/1607.00148>. Acesso em: 8 abr. 2025

MCELFRESH, Duncan *et al.* When Do Neural Nets Outperform Boosted Trees on Tabular Data? **Advances in Neural Information Processing Systems**, v. 36, p. 76336–76369, 15 dez. 2023.

MEHDIYEV, Nijat *et al.* Time Series Classification using Deep Learning for Process Planning: A Case from the Process Industry. **Procedia Computer Science**, Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA. v. 114, p. 242–249, 1 jan. 2017.

MOLNAR, Christoph. **Interpretable Machine Learning**. *[S.l.]*: Lulu.com, 2020.

MORRIS, Edward K.; SMITH, Nathaniel G.; ALTUS, Deborah E. B. F. Skinner's contributions to applied behavior analysis. **The Behavior Analyst**, v. 28, n. 2, p. 99–131, 1 out. 2005.

MYERS, Scott M.; JOHNSON, Chris Plauché; THE COUNCIL ON CHILDREN WITH DISABILITIES. Management of Children With Autism Spectrum Disorders. **Pediatrics**, v. 120, n. 5, p. 1162–1182, 1 nov. 2007.

NATIONAL RESEARCH COUNCIL. **Educating Children with Autism**. Washington, D.C.: National Academies Press, 2001.

NIE, Guangtao *et al.* An Immersive Computer-Mediated Caregiver-Child Interaction System for Young Children With Autism Spectrum Disorder. **IEEE Transactions on Neural Systems and Rehabilitation Engineering**, v. 29, p. 884–893, 2021.

PAIVA JUNIOR, Francisco. **Por que o Brasil pode ter 6 milhões de autistas? Canal Autismo**, 1 abr. 2023. Disponível em: <https://www.canalautismo.com.br/artigos/por-que-o-brasil-pode-ter-6-milhoes-de-autistas/>. Acesso em: 9 jan. 2025

PANDYA, Shivani; JAIN, Swati; VERMA, Jaiprakash. A comprehensive analysis towards exploring the promises of AI-related approaches in autism research. **Computers in Biology and Medicine**, v. 168, p. 107801, 1 jan. 2024.

PASZKE, Adam *et al.* Pytorch: An imperative style, high-performance deep learning library. **Advances in neural information processing systems**, v. 32, 2019.

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

PENG, Chun-Yang; PARK, You-Jin. A New Hybrid Under-sampling Approach to Imbalanced Classification Problems. **Applied Artificial Intelligence**, v. 36, n. 1, p. 1975393, 31 dez. 2022.

PriorLabs/TabPFN. Prior Labs, , 20 ago. 2025. Disponível em: <https://github.com/PriorLabs/TabPFN>. Acesso em: 21 ago. 2025

Python API Reference — xgboost 3.0.4 documentation. Disponível em: <https://xgboost.readthedocs.io/en/stable/python/python_api.html#module-xgboost.plotting>. Acesso em: 18 ago. 2025.

RAJAGOPALAN, Shyam Sundar *et al.* Machine Learning Prediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background Information. **JAMA Network Open**, v. 7, n. 8, p. e2429229, 19 ago. 2024.

RÊGO, Amália Cinthia Meneses do; ARAÚJO-FILHO, Irami. Leveraging Artificial Intelligence to enhance the Quality of Life for patients with Autism Spectrum Disorder: A Comprehensive Review. **European Journal of Clinical Medicine**, v. 5, n. 5, p. 28–38, 30 set. 2024.

RIFKIN, Ryan; KLAUTAU, Aldebaro. In defense of one-vs-all classification. **Journal of machine learning research**, v. 5, n. Jan, p. 101–141, 2004.

ROANE, Henry S.; FISHER, Wayne W.; CARR, James E. Applied Behavior Analysis as Treatment for Autism Spectrum Disorder. **The Journal of Pediatrics**, v. 175, p. 27–32, 1 ago. 2016.

SANTOS, Laura *et al.* Design of a Robotic Coach for Motor, Social and Cognitive Skills Training Toward Applications With ASD Children. **IEEE Transactions on Neural Systems and Rehabilitation Engineering**, v. 29, p. 1223–1232, 2021.

SCHWARTZ, Brian *et al.* Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. **Psychotherapy Research**, v. 31, n. 1, p. 33–51, 2 jan. 2021.

SHARMA, Nitika. **How to Use XGBoost for Time-Series Forecasting? Analytics Vidhya**, 4 jan. 2024. Disponível em: <https://www.analyticsvidhya.com/blog/2024/01/xgboost-for-time-series-forecasting/>. Acesso em: 8 abr. 2025

SHATTE, Adrian B. R.; HUTCHINSON, Delyse M.; TEAGUE, Samantha J. Machine learning in mental health: a scoping review of methods and applications. **Psychological**

Medicine, v. 49, n. 9, p. 1426–1448, jul. 2019.

SHINDE, Anita Vikram; PATIL, Dipti Durgesh. A Multi-Classifier-Based Recommender System for Early Autism Spectrum Disorder Detection using Machine Learning. **Healthcare Analytics**, v. 4, p. 100211, 1 dez. 2023.

SHORE, Stephen M. **Beyond the wall: Personal experiences with autism and Asperger syndrome**. *[S.l.]*: AAPC Publishing, 2003.

SPARROW, Sara S.; BALLA, Rolf; CICCHETTI, Domenic V. **Vineland-3: Vineland Adaptive Behavior Scales. Manual**. *[S.l.]*: American Guidance Service, 2016.

SWETS, John A. Measuring the Accuracy of Diagnostic Systems. **Science**, v. 240, n. 4857, p. 1285–1293, 3 jun. 1988.

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

THABTAH, Fadi *et al.* Data imbalance in classification: Experimental evaluation. **Information Sciences**, v. 513, p. 429–441, 1 mar. 2020.

THENG, Dipti; BHOYAR, Kishor K. Feature selection techniques for machine learning: a survey of more than two decades of research. **Knowledge and Information Systems**, v. 66, n. 3, p. 1575–1637, 1 mar. 2024.

VASWANI, Ashish *et al.* Attention is All you Need. *In*: Curran Associates, Inc., 2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>. Acesso em: 9 abr. 2025

VOLKMAR, Fred *et al.* Practice Parameter for the Assessment and Treatment of Children and Adolescents With Autism Spectrum Disorder. **Journal of the American Academy of Child & Adolescent Psychiatry**, v. 53, n. 2, p. 237–257, 1 fev. 2014.

VOLKMAR, FRED R. *et al.* Quantifying Social Development in Autism. **Journal of the American Academy of Child & Adolescent Psychiatry**, v. 32, n. 3, p. 627–632, 1 maio 1993.

XGBoost Documentation — xgboost 3.0.4 documentation. Disponível em: <https://xgboost.readthedocs.io/en/stable/>. Acesso em: 21 ago. 2025.

YANG, Luoxiao *et al.* **ViTime: A Visual Intelligence-Based Foundation Model for Time Series Forecasting**. arXiv, , 8 fev. 2025. Disponível em: <http://arxiv.org/abs/2407.07311>. Acesso em: 14 abr. 2025

YANG, Yuxuan; KHORSHIDI, Hadi Akbarzadeh; AICKELIN, Uwe. A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. **Frontiers in Digital Health**, v. 6, p. 1430245, 26 jul. 2024.

ZERVEAS, George *et al.* A Transformer-based Framework for Multivariate Time Series Representation Learning. *In*: : KDD '21.New York, NY, USA: Association for Computing

Machinery, 14 ago. 2021. Disponível em: <https://dl.acm.org/doi/10.1145/3447548.3467401>. Acesso em: 9 abr. 2025

ZHOU, Yongxia; YU, Fang; DUONG, Timothy. Multiparametric MRI Characterization and Prediction in Autism Spectrum Disorder Using Graph Theory and Machine Learning. **PLOS ONE**, v. 9, n. 6, p. e90405, 12 jun. 2014.

ZHU, Ziyu *et al.* Integrating Machine Learning and the SHapley Additive exPlanations (SHAP) Framework to Predict Lymph Node Metastasis in Gastric Cancer Patients Based on Inflammation Indices and Peripheral Lymphocyte Subpopulations. **Journal of Inflammation Research**, v. 17, p. 9551–9566, 31 dez. 2024.

APPENDIX A - Detailed performance of baseline models

This appendix provides the detailed, class-specific performance results for the baseline models used for comparison in Chapter 4. The models include the SVM, KNN, and LSTM. While their overall performance was lower than that of the TabPFN and XGBoost models, these results are presented for methodological completeness.

Table A1 consolidates the classification reports for all three models, directly comparing their Precision, Recall, and F1-scores on a class-by-class basis. The corresponding ROC curves for each model are presented in the following figures A1, A2, and A3.

Table A1 - Consolidated classification reports for baseline models on the Test Set

Model	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
SVM	Validated	0.75	0.82	0.78	391
	Completed	0.70	0.65	0.67	261
	Rejected	0.00	0.00	0.00	17
KNN	Validated	0.72	0.84	0.78	391
	Completed	0.70	0.58	0.63	261
	Rejected	0.00	0.00	0.00	17
LSTM	Validated	0.77	0.75	0.76	391
	Completed	0.66	0.72	0.69	261
	Rejected	1.00	0.06	0.11	17

Figure A1: ROC curves for the SVM model

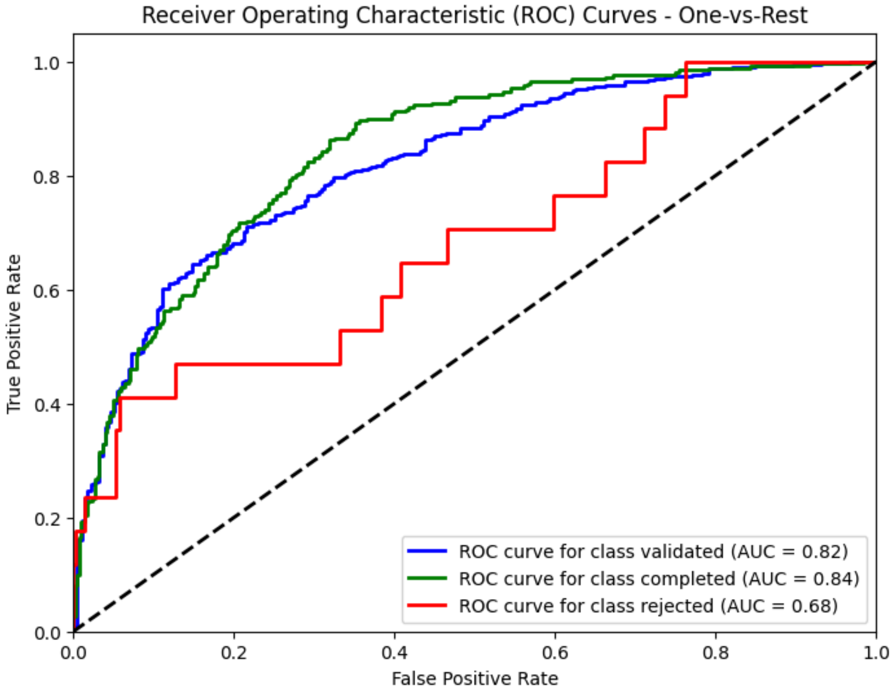


Figure A2: ROC curves for the KNN model

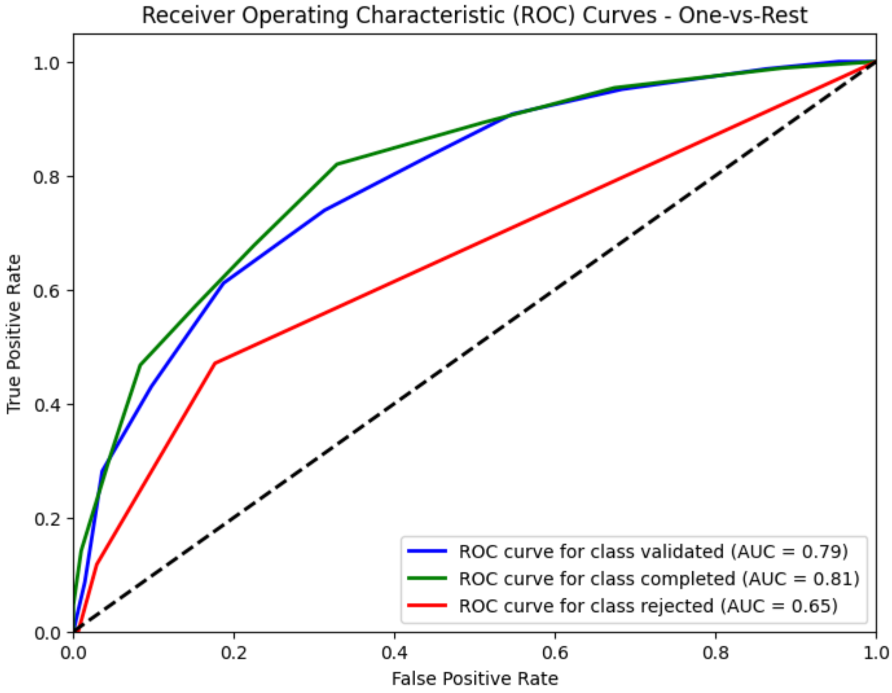
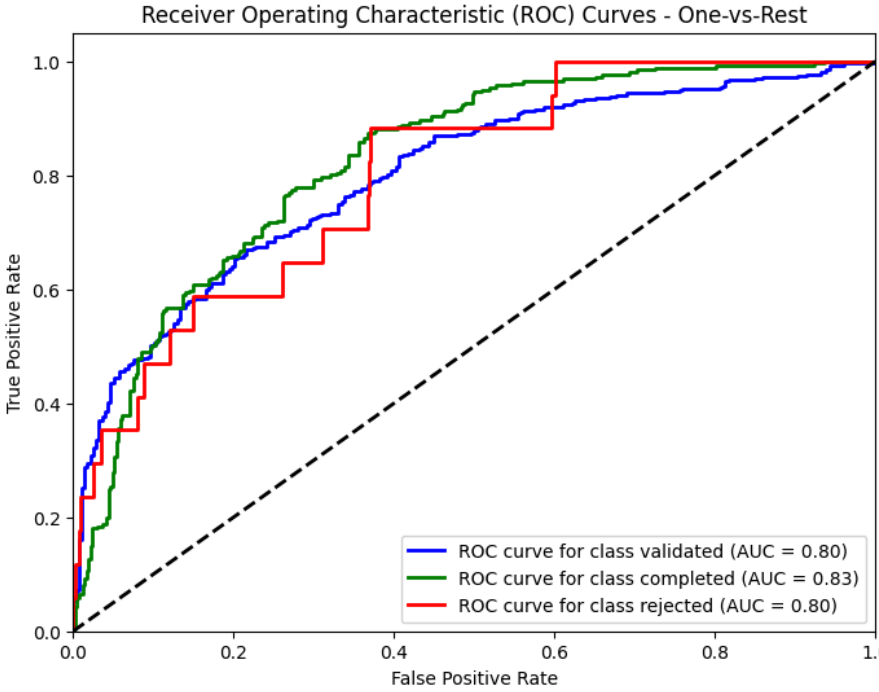


Figure A3: ROC curves for the LSTM model



A.1 SVM

The model demonstrated reasonable performance on the 'Validated' and 'Completed' classes but, like the TabPFN model, failed to correctly identify any instances of the minority 'Rejected' class, resulting in an F1-score of 0.00.

A.2 KNN

The performance was the lowest among all tested models, particularly in its recall for the 'Completed' class (0.58). Like the SVM, the KNN model could not correctly predict any 'Rejected' instances.

A.3 LSTM

The LSTM network was unique in directly processing the raw, sequential time-series data rather than the engineered features. While its overall metrics were modest, the LSTM was the only model besides XGBoost to correctly classify an instance of the 'Rejected' class. It correctly identified one 'Rejected' objective, resulting in a precision of 1.00 but a very low recall of 0.06.

APPENDIX B - Supplementary feature importance analyses

This appendix presents the results from the supplementary feature importance analyses conducted as part of the methodology described in Chapter 3. A filter-based approach (SelectKBest) and an embedded method (Random Forest) were explored to provide a comprehensive view of feature relevance. The findings are presented here for completeness. The results from both methods were consistent with the primary analyses in Section 4.4, identifying time-series dynamics and intervention duration as the most influential predictors.

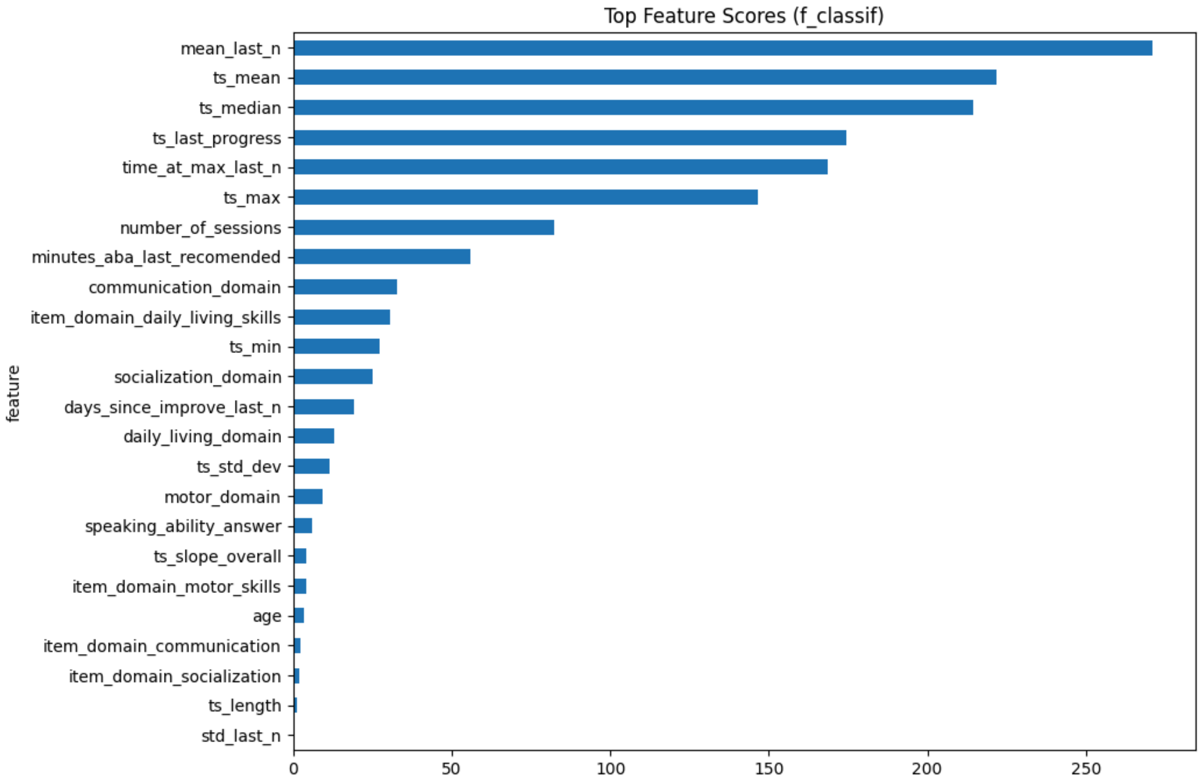
B.1 Feature importance using SelectKBest (ANOVA F-test)

The SelectKBest method, configured with the ANOVA F-test (f_classif), was used to score each feature based on its statistical relationship with the target variable, independent of any predictive model. The resulting F-scores, where a higher value indicates a stronger relationship, are presented in Table B1 and Figure B1. The analysis highlights that features engineered from the time-series data, particularly those describing recent performance (mean_last_n) and overall trends (ts_mean, ts_median), have the strongest statistical association with the objective's final status.

Table B1 - Top 15 features ranked by SelectKBest (ANOVA F-test)

Rank	Feature	F-score
1	mean_last_n	271.00
2	ts_mean	221.87
3	ts_median	214.45
4	ts_last_progress	174.60
5	time_at_max_last_n	168.68
6	ts_max	146.42
7	number_of_sessions	82.18
8	minutes_aba_last_recomended	56.05
9	communication_domain	32.68
10	item_domain_daily_living_skills	30.55
11	ts_min	27.19
12	socialization_domain	25.12
13	days_since_improve_last_n	19.15
14	daily_living_domain	12.96
15	ts_std_dev	11.28

Figure B1 - Feature scores from SelectKBest (f_classif)



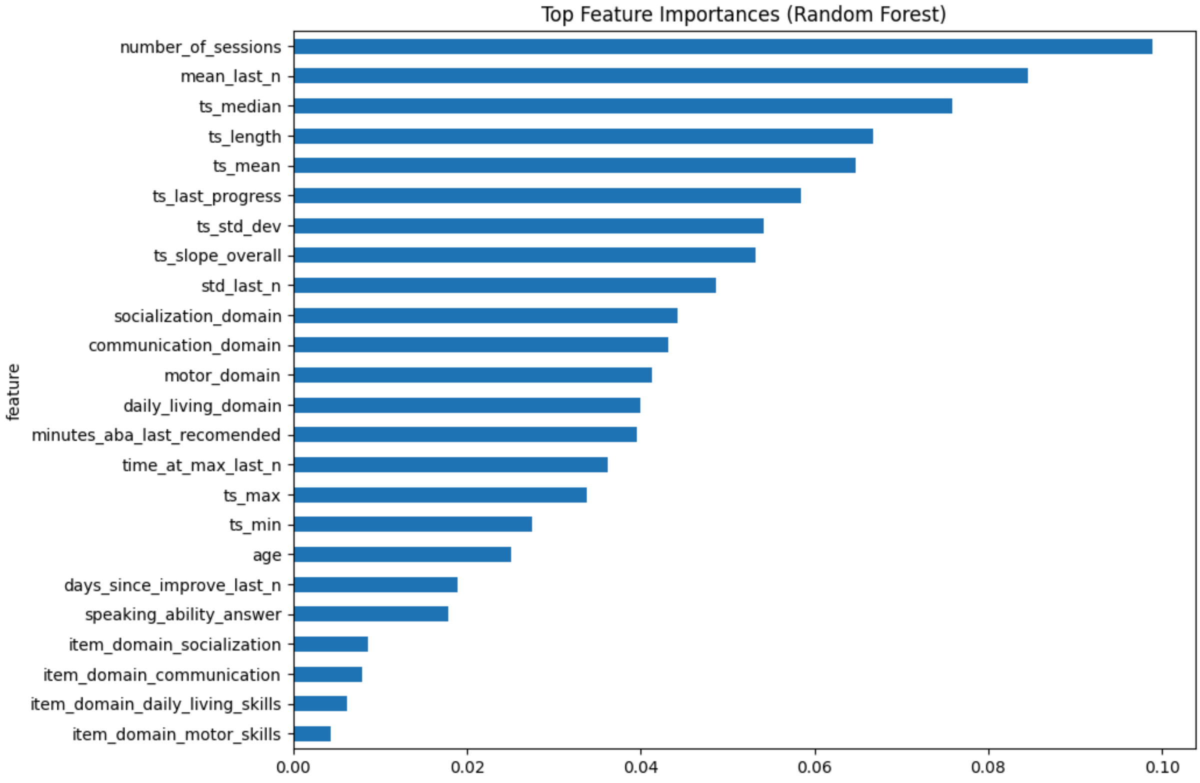
B.2 Feature importance using Random Forest

Feature importance was also derived from a trained Random Forest model. This embedded method calculates importance based on the average decrease in impurity contributed by each feature across all decision trees in the ensemble. The results in Table B2 and Figure B2 provide a model-based perspective on feature relevance. Consistent with the other methods, number_of_sessions and features capturing recent performance (mean_last_n, ts_median, ts_mean) were ranked as the most important predictors.

Table B2 - Feature importance derived from the Random Forest model

Rank	Feature	Importance
1	number_of_sessions	0.0989
2	mean_last_n	0.0845
3	ts_median	0.0759
4	ts_length	0.0668
5	ts_mean	0.0647
6	ts_last_progress	0.0584
7	ts_std_dev	0.0542
8	ts_slope_overall	0.0532
9	std_last_n	0.0487
10	socialization_domain	0.0443
11	communication_domain	0.0432
12	motor_domain	0.0413
13	daily_living_domain	0.0399
14	minutes_aba_last_recomended	0.0396
15	time_at_max_last_n	0.0362

Figure B2 - Feature importance from the Random Forest model


Top Feature Importances (Random Forest)
feature
number_of_sessions
mean_last_n
ts_median
ts_length
ts_mean
ts_last_progress
ts_std_dev
ts_slope_overall
std_last_n
socialization_domain
communication_domain
motor_domain
daily_living_domain
minutes_aba_last_recomended
time_at_max_last_n
ts_max
ts_min
age
days_since_improve_last_n
speaking_ability_answer
item_domain_socialization
item_domain_communication
item_domain_daily_living_skills
item_domain_motor_skills