Predictive model for Autistic Spectrum Disorder Screening Data for Children

Carlos André V. Tinin[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

Abstract. *This work aims to propose a machine learning project focused on the early prediction of Autistic Spectrum Disorder (ASD) in children. Therefore, addressing the need for efficient and accessible screening methods to mitigate the substantial healthcare costs and lengthy diagnosis waiting times associated with the condition. Leveraging a dataset that includes ten behavioral features (AQ-10-Child) and ten individual characteristics publicly available in UCI ML, the project seeks to improve the efficiency in terms of recall of ASD screening. The study explores and optimizes several machine learning algorithms, including kNN with optimized k values, SVM with optimized C parameters, Gaussian Naive Bayes, Decision Trees with optimized* max_depth*, and Random Forests with optimized* $n\_estimators$*. Each model's performance is evaluated using confusion matrices and recall scores on a 3 way hold-out validation, with a particular emphasis on recall as the primary metric for selecting the most effective model. Finally, all steps of the project, from data loading to model evaluation, are encapsulated within a Docker environment, ensuring a consistent and reproducible execution across different computing environments.*

Resumo. *Este trabalho tem como objetivo propor um projeto de aprendizado de máquina focado na predição do Transtorno do Espectro Autista (TEA) em crianças. Portanto, aborda a necessidade de métodos de triagem eficientes e acessíveis para mitigar os custos substanciais de saúde e os longos tempos de espera para diagnóstico associados à condição. Aproveitando um conjunto de dados que inclui dez características comportamentais (AQ-10-Child) e dez características individuais disponíveis publicamente no UCI ML, o projeto busca melhorar a eficiência em termos de recall do rastreamento do TEA. O estudo explora e otimiza vários algoritmos de aprendizado de máquina, incluindo kNN com o valor de k otimizado, SVM com o parâmetro C otimizado, Nayve Bayes, árvores de decisão com* max_depth *otimizado e Random Forests com* $n\_estimators$ *otimizados. O desempenho de cada modelo é avaliado usando matrizes de confusão e de recall em uma validação de hold-out de 3 vias, com ênfase particular no recall como a métrica primária para selecionar o modelo mais eficiente. Por fim, todas as etapas do projeto, do carregamento de dados à avaliação do modelo, são encapsuladas em um ambiente Docker, garantindo uma execução consistente e reproduzível em diferentes ambientes.*

1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by challenges in social interaction, communication, and restricted or repetitive

patterns of behavior [Thabtah 2017a]. A crucial aspect of managing ASD is early diagnosis, which has been shown to reduce the impact and associated costs by enabling timely interventions. However, current diagnostic procedures are often lengthy, resource-intensive, and inaccessible to many, leading to extended waiting lists and delayed support [Thabtah 2017a]. This scenario suggests a need for alternative solutions for reliable screening methods.

This work addresses this need by proposing a complete machine learning project aimed at developing a predictive model for ASD screening data in children. The primary objective is to enhance the efficiency of screening processes, specifically in terms of maximizing recall, therefore ensuring that as many true positive cases of ASD as possible are identified early. Using a publicly available dataset comprising key behavioral and individual characteristics, this project explores and optimizes various supervised machine learning algorithms [Thabtah 2017b]. Lastly, the project was build keeping in mind the reproducibility of the entire pipeline used leveraging Docker containers and GitHub, promoting transparent and verifiable research outcomes.

2. Theoretical Background

2.1. Autism Spectrum Disorder (ASD)

ASD is a complex neurodevelopmental condition characterized by persistent deficits in social communication and interaction, alongside restricted, repetitive patterns of behavior, interests, or activities [Hirota and King 2023]. While historically conceptualized as distinct disorders, the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) adopted a dimensional approach, consolidating these conditions under a single spectrum. This shift provides an opportunity to better characterize heterogeneous clinical presentations and potentially identify ASD subtypes, which is crucial for advancing research and clinical understanding [Grzadzinski et al. 2013]. Early signs often manifest within the first three years of life and include reduced eye contact, limited use of gestures, and lack of imaginative play [Hirota and King 2023, Park et al. 2016]. Although commonly associated with atypical social looking patterns, such as less eye contact and more mouth gaze, emerging evidence, particularly in children, indicates that this generalization may not consistently hold across all ages and contexts, suggesting that underlying mechanisms relate more to general attentional preferences for social stimuli rather than specific facial features [Falck-Ytter and von Hofsten 2011]. Furthermore, impairments in the self-system, particularly the psychological self, have been observed in individuals with ASD, correlating with their social and cognitive functioning levels [Huang et al. 2017]. Given the rising prevalence and the substantial impact of ASD, early diagnosis and intervention, primarily through behavioral therapies, are paramount, though pharmacological treatments may address co-occurring psychiatric conditions [Hirota and King 2023, Park et al. 2016]

2.2. Dataset: Autistic Spectrum Disorder Screening Data for Children

The dataset used in this project was the ”Autism-Child-Data.arff” file, publicly available through the UCI Machine Learning Repository. This dataset was specifically compiled for autism screening in children and is designed to identify individuals with ASD based on a combination of behavioral traits and individual characteristics. It comprises 20 features

derived from a Autism Spectrum Questionaire (AQ-10) and demographic and personal attributes such as age and gender [Thabtah 2017a]. The target variable, 'Class/ASD', indicates whether a child has ASD or not. The dataset is very comprehensive and focus on behavioral traits which makes it particularly useful for developing predictive models.

2.3. Machine Learning Algorithms

This project uses a set of machine learning algorithms for the classification of ASD screening data. They were prioritized in order to practice and review the concepts learned from the CMP263 subject.

- **K-Nearest Neighbors (kNN):** kNN algorithm operates by classifying new data points based on the majority class among their k nearest neighbors. Its performance is significantly influenced by the choice of distance measure, and it demonstrates a degree of robustness to noise [Abu Alfeilat et al. 2019].
- **Support Vector Machines (SVM):** This algorithm construct an optimal hyperplane to separate data points into different classes. SVMs offer theoretical tractability grounded in computational learning theory and consistently achieve strong performance across various real-world applications, including text categorization and face detection [Hearst et al. 1998].
- **Gaussian Naive Bayes (GNB):** GNB is a probabilistic algorithm which often competes effectively with more sophisticated classifiers in practice. Empirical studies indicate that GNB performs well with low-entropy feature distributions and certain types of functional feature dependencies, with its accuracy being more correlated with the information lost due to the independence assumption rather than the direct degree of feature dependency [Rish et al. 2001].
- **Decision Trees (DT):** Decision Trees defines classification systems by segmenting populations into tree-like structures. They are capable of efficiently handling large and complex datasets without imposing restrictive parametric assumptions, facilitating both classification and prediction tasks [Ying et al. 2015].
- **Random Forests (RF):** Random Forests was chosen as an ensemble learning approach, by combining multiple decision tree predictors, where the construction of each tree is based on a random vector. This method demonstrates robust performance, with its generalization error converging as the number of trees increases, and exhibits greater robustness to noise compared to single tree models. Additionally, internal estimates provide insights into error rates and feature importance [Breiman 2001].

3. Methodology

The methodology adopted in this project follows a simple machine learning pipeline. Figure 1 shows the followed steps which encompasses data loading and initial analysis, first step of preprocessing, data splitting, second step of preprocessing, model training with hyperparameter optimization and, finally, model evaluation.

3.1. Data collection and preparation

The initial phase involved a exploration of the dataset to understand its structure, characteristics, and potential challenges. This step was crucial for informing subsequent preprocessing decisions. As was described in the Section 2.2 the Dataset comprehends be-

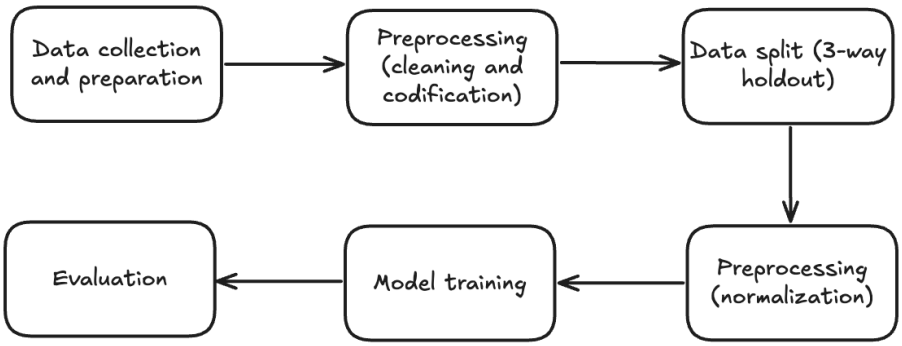

Figure 1. Methodology steps.

havioral and individual characteristics of each participant. The test could be taken by the child itself, or a related care taker.

During the analyses was noticed that the $Relation$ and $Ethnicity$ columns have missing values, thus, in order to prevent missinput data, these columns weren't considered in the training of the models. Moreover, the column age_desc, helds the same information for all of the rows and consequentially was also marked for deletion. $Result$ attribute was also marked for deletion since it represented the sum of positive answers from the questionnaire, therefore it wouldn't contribute for the model generalization. For this same reason, neither $used\_app\_before$ nor $country\_of\_residence$ were considered also.

Finaly, looking at the target attribute distribution produced another crucial insight about the data split strategy. As shown in the Figure 2, the classes are pretty much balanced and therefore a simple 3-way holddout would be enough to split the data.

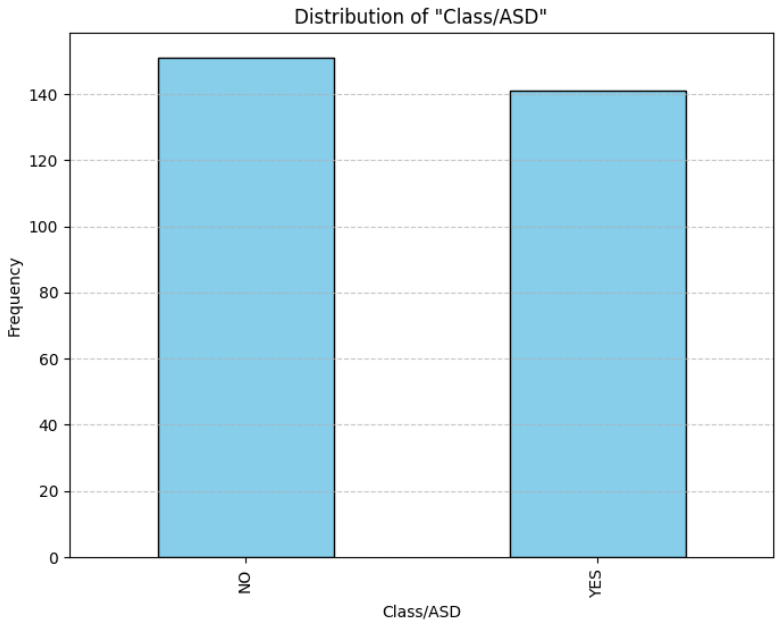

Figure 2. Target attribute histogram.

3.2. Preprocessing

Following the Figure 1, the next step is the preprocessing. This step was designed in two distinct phases to effectively prevent data leakage. Initially, prior to the data split, identified columns marked for deletion were removed, and categorical features possessing only two possible values were binarized. Additionally, rows containing empty values were eliminated. Notably, no data imputation was performed to prevent the potential introduction of bias or information from the validation and test sets into the training process.

Moreover, after the data split phase (Section 3.3), another round of preprocessing was applied in order to normalize the Age column to have a single scale for its values. It's important to note that this column were normalized for each set (train, validation and test) separatedly in order to prevent data leakage.

3.3. Data Split

The preprocessed dataset was divided into three distinct subsets: training, validation, and test sets. This 3-way hold-out validation strategy is valuable for tuning the hyperparameters of each model, except for Naive Bayes. By evaluating on a separate validation set, we prevent overfitting to the training data and ensure that the chosen model generalizes well to unseen data. The test set was reserved exclusively for the final evaluation of the best-performing model. This set provides an unbiased estimate of the model's performance on new, real-world data, as it has not been used in any part of the training or hyperparameter tuning process.

3.4. Hyperparameters Optimization

Hyperparameter optimization is a crucial step in the machine learning pipeline, involving the selection of the optimal set of hyperparameters for a learning algorithm that maximizes model performance on the validation set. For each of the following algorithms the metric used was recall to ensure that as many true positive cases of ASD as possible are identified early.

3.4.1. kNN

The optimal value for k (the number of neighbors) was determined by evaluating the model's recall on the validation set across a range of k values. Figure 3 shows the results of each number of k: for accuracy, precision and recall (even though recall was the main metric, precision was used for tiebreakers). With $k = 11$ the recall metric was maximized, hence it was chosen the value for kNN.

3.4.2. SVM

The regularization parameter C, which controls the trade-off between maximizing the margin and minimizing classification errors, was optimized for the SMV model. Similar to kNN, the Figure 4 shows the metrics behavior across multiple C values. For this case, specifically, all three metrics were maximized with $C = 1$ demonstrating a great result for SVM.

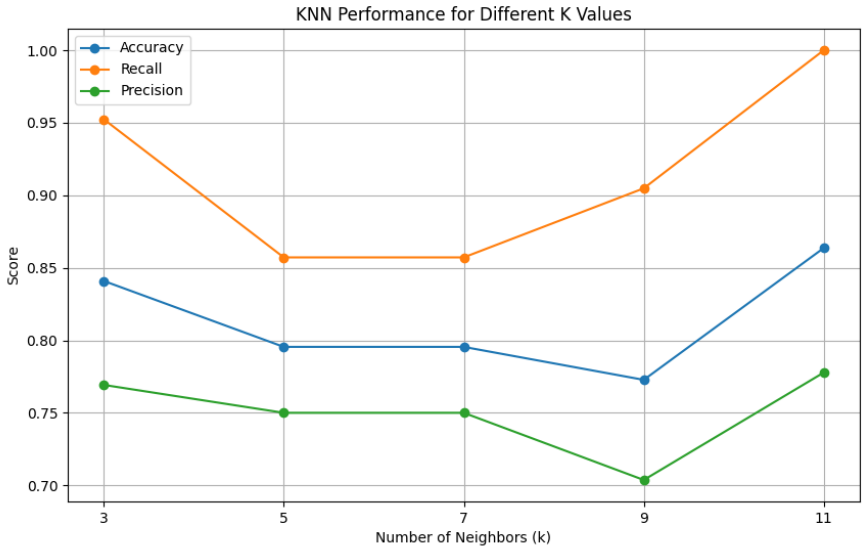

Figure 3. kNN performance across multiple k values.

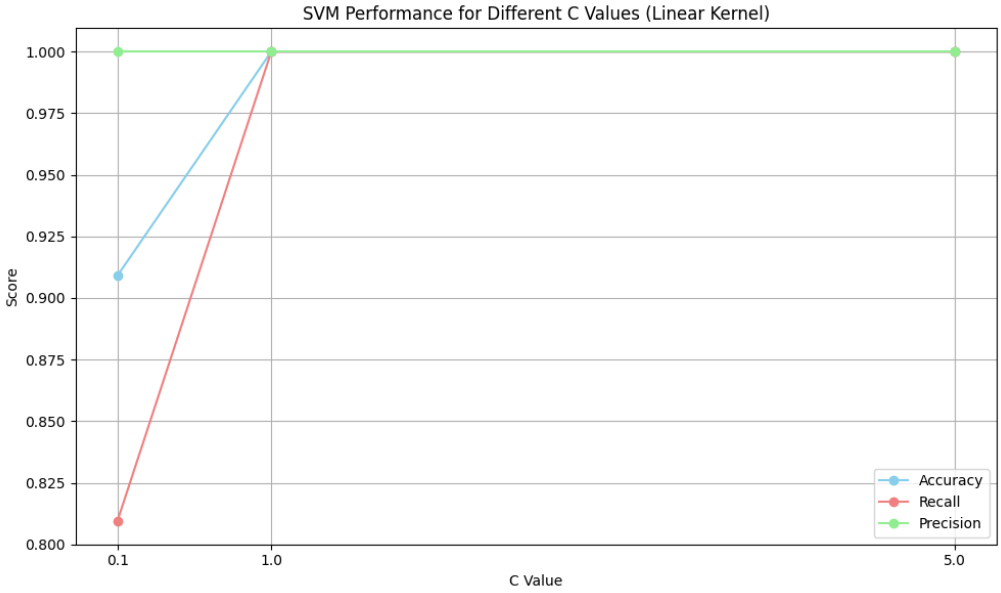

Figure 4. SVM performance across multiple C values.

3.4.3. Decision Trees

For Decision Trees, the ‘max_depth‘ hyperparameter, which limits the maximum depth of the tree, was optimized to prevent overfitting and improve generalization. The performance of each configuration is shown in Figure 5. The chosen value for max_depth was 7 since it had the greater value possible for Recall in this configuration.

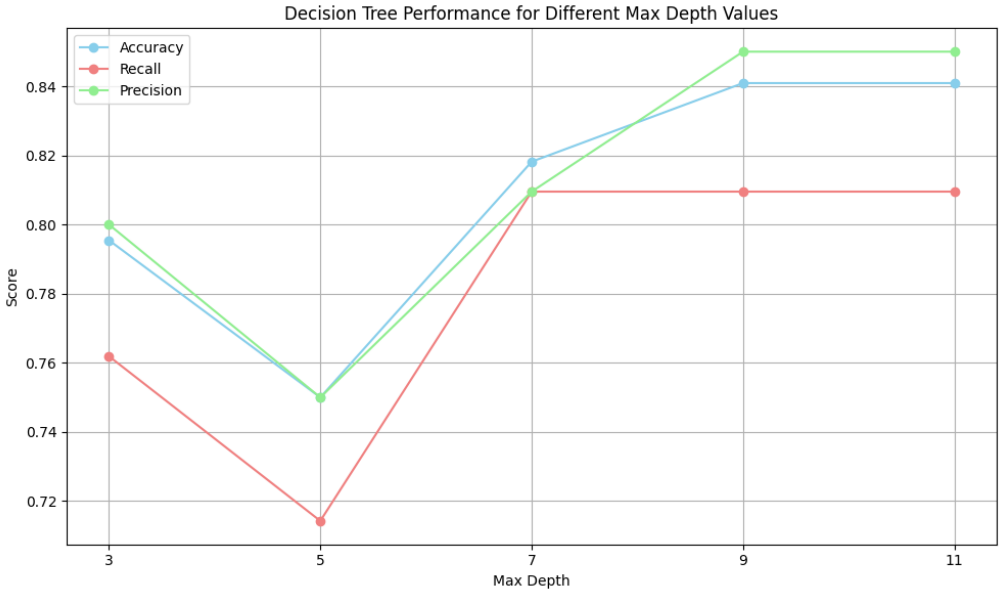

Figure 5. DT performance across multiple max_depth values.

3.4.4. Random Forests

Finally, Figure 6 highligths the performance of Random Forests for different ‘n_estimators‘ hyperparameter, which represents the number of trees in the forest. A higher number of trees generally improves performance but also increases computational cost.

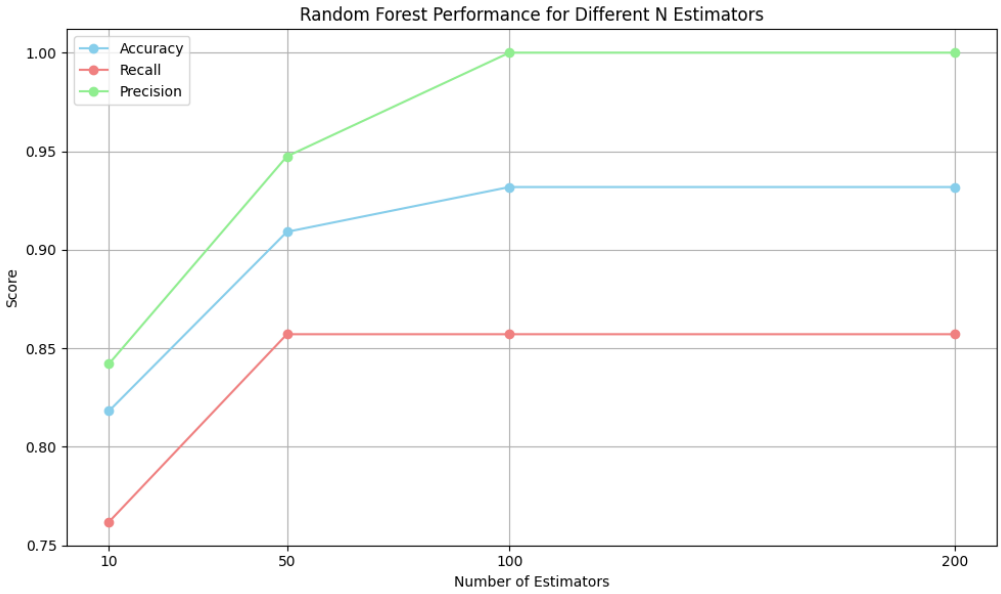

Figure 6. RF performance across multiple n_estimators values.

3.5. Environment (Reproducibility)

Ensuring the reproducibility of machine learning projects is the basis for scientific methods, collaborative development, and deployment. This project achieves high reproducibility by encapsulating the entire execution environment within Docker containers. This approach guarantees that the project can be built and run consistently across different machines, eliminating issues related to conflicting dependencies or environment configurations.

The project is publicly available on Github [Tinin 2025], and it leverages a 'Dockerfile' to define the necessary operating system, software packages, and specific versions of Python libraries required for the project to run correctly. The 'Makefile' simplifies the Docker interactions, providing predefined commands for building and running the project. Specific intructions for running the experiments are detailed in the Github repository.

4. Evaluation

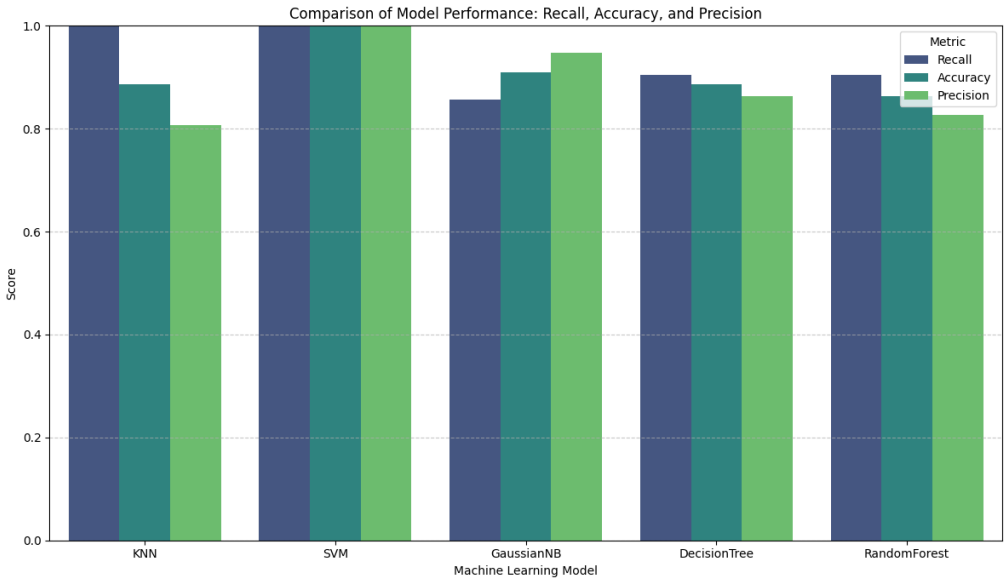

Figure 7. Model performance comparison.

The evaluation phase is critical for assessing the performance of the trained machine learning models. Given the context of ASD screening, where early detection is paramount, a strong emphasis was placed on the recall score. In addition to individual metrics, confusion matrices were generated to provide a comprehensive understanding of each model's predictive capabilities. Figure 7 aggregates the performance results of each model, displaying their scores for accuracy, precision, and recall.

The findings of this analysis indicates that, even when compared against other robust machine learning methods, the Support Vector Machine (SVM) model highlights as the most effective option for the characteristics of this specific dataset. It achieved greater performance for each of the three metrics, maximizing them indicating a great generalization. The SVM's capability in establishing optimal decision boundaries within the feature space of the ASD screening data positions it as a highly reliable candidate for predictive

tasks in sensitive clinical domains, contributing significantly to the development of more efficient and accurate early screening methods.

5. Conclusion

This project successfully developed a comprehensive machine learning pipeline for the early prediction of Autistic Spectrum Disorder in children, emphasizing the critical need for efficient and accessible screening methods. By leveraging a publicly available dataset from UCI ML and applying a rigorous methodology, including data exploration, preprocessing, and a 3-way hold-out validation strategy, the study explored and optimized various machine learning algorithms.

The evaluation, focusing primarily on recall, demonstrated the potential of these models to accurately identify ASD cases, thereby minimizing critical false negatives that could lead to delayed interventions. An important point of this work is the commitment to reproducibility, achieved through the systematic encapsulation of the entire project within a Docker environment.

Future work could involve exploring more advanced feature engineering techniques, integrating larger and more diverse datasets, investigating deep learning architectures for improved predictive power, and conducting a further analysis of the interpretability of the models to understand the most influential features for ASD prediction.

6. Use of Artificial Inteligence in this work

In the development of this project, Artificial Intelligence (AI) generative tools were used to various writing stages, ensuring adherence to ethical and transparency guidelines. These tools primarily helped to structure writing of this article. Furthermore, AI contributed to refining the clarity and conciseness of the academic language, correcting LaTeX notation, and formatting project documentation. Crucially, the AI did not generate any scientific data or research outcomes; instead, it served as an assistant to enhance the presentation and communicability of content originally developed by the researcher, thereby upholding the integrity and authorship of the scientific work.

References

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., and Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data*, 7(4):221–248.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Falck-Ytter, T. and von Hofsten, C. (2011). Chapter 12 - how special is social looking in asd: A review. In Braddick, O., Atkinson, J., and Innocenti, G. M., editors, *Gene Expression to Neurobiology and Behavior: Human Brain Development and Developmental Disorders*, volume 189 of *Progress in Brain Research*, pages 209–222. Elsevier.

Grzadzinski, R., Huerta, M., and Lord, C. (2013). Dsm-5 and autism spectrum disorders (asds): an opportunity for identifying asd subtypes. *Molecular autism*, 4:1–6.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Hirota, T. and King, B. H. (2023). Autism spectrum disorder: a review. *Jama*, 329(2):157–168.

Huang, A. X., Hughes, T. L., Sutton, L. R., Lawrence, M., Chen, X., Ji, Z., and Zeleke, W. (2017). Understanding the self in individuals with autism spectrum disorders (asd): A review of literature. *Frontiers in Psychology*, Volume 8 - 2017.

Park, H. R., Lee, J. M., Moon, H. E., Lee, D. S., Kim, B.-N., Kim, J., Kim, D. G., and Paek, S. H. (2016). A short review on the current understanding of autism spectrum disorders. *Experimental neurobiology*, 25(1):1.

Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. Seattle, USA.

Thabtah, F. (2017a). Autism spectrum disorder screening: Machine learning adaptation and dsm-5 fulfillment. pages 1–6.

Thabtah, F. (2017b). Autistic Spectrum Disorder Screening Data for Children . UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5659W.

Tinin, C. A. (2025). Final project for cmp263.

Ying, L. et al. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.