UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

LAURA GALANT SPEGGIORIN

Development of a machine learning model for molecular signature identification and early diagnostic support of Autism Spectrum Disorder

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Science

Advisor: Prof. Dr. Mariana Recamonde Mendoza
Coadvisor: Prof. Dr. Thayne Woycinck Kowalski

Porto Alegre
August 2024

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos André Bulhões Mendes
Vice-Reitora: Prof^{a}. Patricia Helena Lucas Pranke
Pró-Reitora de Graduação: Prof^{a}. Cíntia Inês Boll
Diretora do Instituto de Informática: Prof^{a}. Carla Maria Dal Sasso Freitas
Coordenador do Curso de Ciência de Computação: Prof. Marcelo Walter
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

ABSTRACT

Autism Spectrum Disorder (ASD) is a spectrum of prevalent, highly heritable, and heterogeneous neurodevelopmental disorders that manifest through impairments in social communication and interaction, sensory sensitivities, repetitive behaviors, and varying degrees of intellectual disability. Because of the increase in the diagnosis of ASD in recent years and the fact that its molecular mechanisms are not completely understood, better diagnosis and better understanding of its origins is pivotal. It is known that the occurrence of the disorder is influenced by both genetic and environmental factors, making biological data that combine both genetic and environmental influences such as genome-wide gene expression levels a good candidate for this study. Machine Learning can be used to learn complex underlying patterns in ASD through classification algorithms. Considering the difficulty and impacts in life when the diagnosis is delayed and the fact that ASD may be influenced by prenatal, perinatal, and very early postnatal environmental factors, this work proposes a support machine learning model to aid in the early diagnosis of ASD with gene expression information from samples of blood collected from the umbilical cord at the time of birth, taking advantage of both genetic information as well as a unique insight to the neonate's environment in its most susceptible period. That is achieved through a two-step ML classifier, where the first part classifies instances in typical development (TD) or not, and the second part tries to separate those classified as not being TD in ASD or not. An ensemble approach was used in each part, combining several of the most known algorithms tuned for each individual problem alongside dimensionality reduction techniques. The classifiers had their hyperparameters tuned to each problem, and the models were validated inside a k-fold cross-validation before being tested on previously separated data. Several metrics were extracted to characterize the proposed model's performance, and they indicated subtle but promising results that validate the idea behind this work, suggesting this model could be an important step into building a reliable and robust pre-diagnostic tool.

Keywords: Machine learning. autism spectrum disorder. ensemble. bioinformatics. gene expression. transcriptomics. neurodevelopment.

Desenvolvimento de modelo de aprendizado de máquina para identificação de assinatura molecular e auxílio ao diagnóstico precoce de Transtorno do Espectro Autista

RESUMO

O Transtorno do Espectro do Autismo (TEA) é um grupo bastante frequente de transtornos hereditários e heterogêneos do neurodesenvolvimento que se manifestam por meio de déficits na comunicação e interação social, sensibilidades sensoriais, comportamentos repetitivos e variados níveis de deficiência intelectual. Com o aumento recente dos diagnósticos de TEA e a falta de compreensão aprofundada sobre seus mecanismos, é crucial melhorar o diagnóstico e entender melhor suas origens. Sabe-se que o TEA é influenciado por fatores genéticos e ambientais, e dados biológicos que combinam influencias genéticas e ambientais como dados de expressão gênica são promissores para investigar essas influências. O aprendizado de máquina pode ser usado para aprender padrões sutis e complexos em TEA, por meio de algoritmos de classificação. Considerando a dificuldade e os impactos na vida quanto mais tardio o diagnóstico, e o fato de que o TEA pode ser influenciado por fatores ambientais pré-natais, perinatais e pós-natais, este estudo propõe um modelo de aprendizado de máquina para auxiliar no diagnóstico precoce do TEA utilizando informações de expressão genética obtidas de amostras de sangue do cordão umbilical no momento do nascimento, aproveitando tanto a informação genética quanto uma visão única do ambiente do bebê em seu período mais suscetível. O modelo é composto por duas etapas: a primeira classifica os indivíduos como desenvolvimento típico (DT) ou não, e a segunda distingue TEA dentre as amostras classificadas como nao sendo DT. Utilizou-se uma abordagem ensemble, com múltiplos algoritmos de classificação para cada etapa, combinada com técnicas de redução de dimensionalidade. Os classificadores tiveram seus parametros ajustados e validados através de validação cruzada k-fold antes de serem testados em dados independentes. Diversas métricas foram extraídas para caracterizar o desempenho do modelo proposto, e indicaram resultados sutis, mas promissores, validando a hipotese deste estudo sugerindo que o modelo pode ser um passo importante para o desenvolvimento de uma ferramenta diagnóstica mais confiável e robusta para o TEA.

Palavras-chave: aprendizado de máquina, transtorno do espectro autista, ensemble. bioinformática, expresssão gênica, transcriptômica, neurodesenvolvimento..

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
CV	Cross-validation
LR	Logistic Regression
ML	Machine Learning
NB	Naive Bayes
RF	Random Forest
TD	Typical Development
ASD	Autism Spectrum Disorder
EFB	Exclusive Feature Bundling
KNN	K-Nearest Neighbors
MLP	Multi-layer Perceptron
MRI	Magnetic Resonance Imaging
NTD	Non-Typical Development
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
XGB	Extreme Gradient Boosting
ADHD	Attention deficit hyperactivity disorder
ADI-R	Autism Diagnostic Interview-Revised
ADOS	Autism Diagnostic Observation Schedule
DSM-5	Diagnostic and Statistical Manual of Mental Disorders—5th edition
ESAT	Early Screening of Autistic Traits
GOSS	Gradient-based One-Side Sampling
LGBM	Light Gradient-Boosting Machine

MSEL	Mullen Scales of Early Learning
EARLI	Early Autism Risk Longitudinal Investigation
M-CHAT	Modified Checklist for Autism in Toddlers
MARBLES	Markers of Autism Risk in Babies-Learning Early Signs

CONTENTS

1 INTRODUCTION

An increase in the diagnosis of Autism Spectrum Disorder (ASD) has been observed in recent years, with studies mentioning it could be as prevalent as 1 in 36 people (ZABLOTSKY; BLACK; BLUMBERG, 2017). This has gradually encouraged more research on the causes and impacts of this disorder. ASD is a group of neurodevelopmental disorders characterized by impairment in social communication and interaction as well as repetitive patterns of behaviors and it frequently occurs alongside other conditions. Almost three-quarters of children with ASD also have another medical, psychiatric, or neurological disorder, leading to additional physical or mental challenges, higher treatment costs, and increased demands on their families (SHARMA; GONDA; TARAZI, 2018). Some of these challenges can be lessened by early interventions, which also justify the need for early diagnosis. However, the heterogeneous origin of ASD is not completely understood. It is known that the occurrence of the disorder is influenced by both genetic and environmental factors, being a highly heritable disorder.

Several studies have been proposed to further understand the relationship between the genetic and environmental components of ASD, like Tylee et al. (2017) and Hertz-Picciotto et al. (2018), and one tool for such is the use of transcriptomic data. Gene expression alterations by default offers a mix of genetic and enviromental factors, making it promising for the study of such a complex, heterogeneous disorder. Alterations have been observed in pathways involved in immune response and neuronal activity functions (HODGES; FEALKO; SOARES, 2020), making blood and brain expression ideal for the study of ASD.

ASD is associated with the role of hundreds of gene variants, hence to this day diagnosis is given by standardized neurodevelopmental assessments applied by professionals. Some machine learning (ML) studies propose the idea of using supervised learning to learn complex underlying patterns in ASD through classification algorithms, essentially offering a preliminary diagnosis based on some specific data (OMAR et al., 2019) (LIU; LI; YI, 2016) (ZHANG et al., 2018). The power of ML combined with the information of gene expression has been shown to give promising results, but mostly rely on postmortem samples of brain tissue or blood samples collected into toddlerhood.

Considering the difficulty and impacts in life the later the diagnosis, and the fact that ASD may be influenced by prenatal, perinatal, and very early postnatal environmental factors, this work proposes a support ML model to aid in the early diagnosis of ASD with gene expression information from samples of blood collected from the umbilical cord at the time of birth, taking advantage of both genetic information as well as a unique insight to the newborn's environment in its most susceptible period.

That is achieved through a two-part ML classifier, where the first part classifies subjects in typical development (TD) or not, and the second part tries to separate those classified as not being TD in ASD or not. An ensemble approach was used in each part, combining several of the most known algorithms tuned for each individual problem, alongside dimensionality reduction techniques. The models were trained, validated and tested in a k-fold cross validation (CV) inside of a holdout.

The proposed architecture offered subtle but promising results in the chosen data that validate the idea behind this work and could be used to improve the method into a reliable and robust model. The idea of a two-step approach takes advantage of the differences between both atypically developing groups and the control group to first separate the samples and later identify ASD subjects from a smaller pool. The plan of combining multiple classifiers yielded a final model that could learn some of the ASD patterns even though some of the individual classifiers could not, and even more, suggested which types of algorithms are better fit for the problem based on their way of learning.

The remainder of this work is structured as follows. Chapter 2 expands on the main concepts of this work in the topics of ASD, ML and gene expression. Chapter 3 presents works on a similar scope to ours. Chapter 4 introduces the data used in this study, alongside the followed methodology and the evaluation approach. Chapter 5 presents and analyses the results obtained through the implementation. Finally, Chapter 6 discusses the conclusions obtained through the whole analysis.

2 BACKGROUND

This Chapter presents the main concepts necessary to understand this research. As the main topic of the problem being tackled, the Section 2.1 offers an overview of Autism Spectrum Disorder, followed by some necessary knowledge in bioinformatics in Section 2.2 to give some insight to the data being used and the interpretation of the model, and finishing with some basic notions of ML to understand the proposed architectures in Section 2.3.

2.1 Autism Spectrum Disorder

ASD, often referred to as autism, is a prevalent, highly heritable, and heterogeneous neurodevelopmental disorder characterized by a variety of cognitive characteristics and frequently occurring alongside other conditions (LORD et al., 2020). The term *autism* was first introduced by KANNER et al. as an independent disorder to a syndrome observed in young children with communication and social impairments, as well as repetitive behaviors 1943. Nowadays, the current definition and diagnostic criteria come from the Diagnostic and Statistical Manual of Mental Disorders—5th edition (DSM-5), that defines it as (ASSOCIATION, 2013):

> Autism spectrum disorder (ASD) is a group of neurodevelopmental disorders characterized by persistent impairment in social communication and interaction and restricted and repetitive patterns of behavior, interests, or activities.

The DSM-5 introduces the idea of a spectrum of disorders, meaning the term ASD is used as an umbrella term to define a myriad of disorders that previously were separate diagnoses, and offers additional clinical descriptors to better define the level of severity and help categorize the level of support needed by an individual with the disorder.

As defined, autism manifests through impairments in social communication and interaction, sensory sensitivities, repetitive behaviors, and varying degrees of intellectual disability. In addition to these core symptoms, almost three-quarters of children with ASD also have another medical, psychiatric, or neurological disorder, leading to additional physical or mental challenges, higher treatment costs, and increased demands on their families (SHARMA; GONDA; TARAZI, 2018). Common conditions include hyperactivity and attention disorders like attention deficit hyperactivity disorder (ADHD), as well as anxiety, depression, and epilepsy (LORD et al., 2020). ASD is also more frequent in patients with chromosomal abnormalities, the most mentioned being Down syndrome and Fragile X syndrome (GENOVESE; BUTLER, 2020) (SHARMA; GONDA; TARAZI, 2018).

The origin of ASD is fairly uncertain, but it is known that both genetic and environmental factors play a part in the presence of the disorder. In Subsection 2.1.1 we will explore some of the known origins and risk factors associated to ASD. Following that, the diagnostic methods and criteria will be explored in Subsection 2.1.2 alongside some prevalence aspects in general population.

2.1.1 Causes of ASD

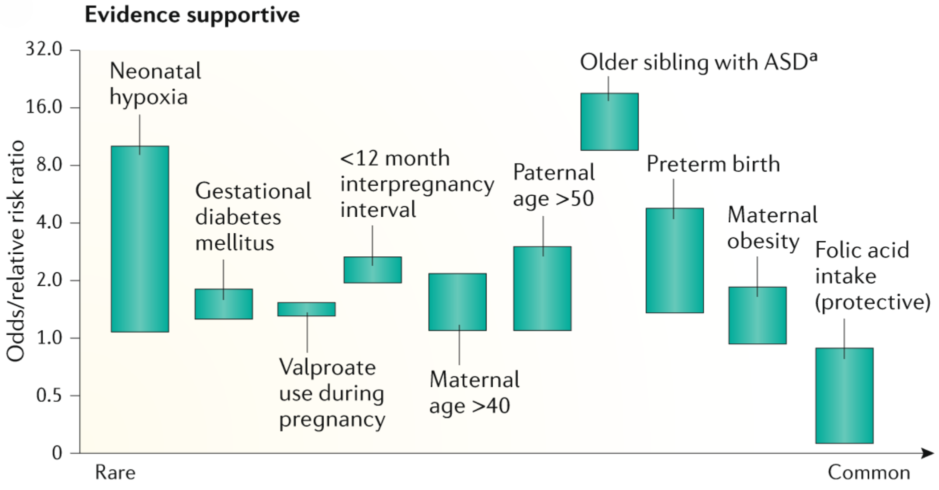
Genetics plays a prominent role in ASD susceptibility. Heritability is the proportion of variation in a population trait that can be attributed to inherited genetic factors. The heritaility of ASD risk has been well established with twin and family studies, with the risk rate ranging from study to study, but agreeing on increased risk of diagnosis when compared to population norms (that risk being much higher, although not absolute, in monozygotic twins) (HODGES; FEALKO; SOARES, 2020). A study on the heritability of medical conditions based on insurance claims for over a third of the entire United States population even demonstrated that autism is among the most heritable common medical conditions, and had the highets estimate in their liability-scale heritability (WANG et al., 2017).

Although ASD is considered one of the most genetically heterogeneous neuropsychiatric disorders, having been associated with the role of hundreds of gene variants with risk effects highly variable and related to other conditions, genomics studies have broadened our understanding of ASD susceptibility genes. Moreover, the study of these genes' functions can further enlighten us on potential biological mechanisms. Some studies suggest, for example, alterations in pathways involved in immune response and neuronal activity functions (HODGES; FEALKO; SOARES, 2020).

Genetics certainly play a role, but the genetic risk may be modulated by prenatal, perinatal, and very early postnatal environmental factors in some patients. One evidence-based review of systematic reviews and meta-analyses of environmental risk factors for autism included a comprehensive coverage of the literature and reported several of these factors (MODABBERNIA; VELTHORST; REICHENBERG, 2017), illustrated in Figure 2.1 among other studies. Some of the environmental factors associated with increased risk for autism identified by this review and other studies are:

- Advanced parental age;
- Birth trauma, particularly if due to proxies of hypoxia;
- Maternal obesity;
- a short interval between pregnancies;
- Gestational diabetes mellitus;
- Prenatal exposure to valproic acid;
- Infants born prematurely (demonstrated to carry a higher risk for ASD and other neurodevelopmental disorders);
- Parental history of psychiatric disorders, and in particular schizophrenia and affective disorders;

Figure 2.1: **Environmental risk factors for autism.** Data from studies aiming to identify risk factors for autism with evidence supporting an association. Bars represent ranges. [a] Represents recurrence risk.



Source: Lord et al. (2020)

It is important to explicitly say that all of the factors mentioned above have been reported as having a *correlation* to the presence of ASD, thus no *causal* determinations have been made to date.

Several studies have been proposed to further understand the relationship between the genetic and environmental components of ASD, but two significant to this work are Early Autism Risk Longitudinal Investigation (EARLI) (NEWSCHAFFER et al., 2012) and Markers of Autism Risk in Babies-Learning Early Signs (MARBLES) (HERTZ-PICCIOTTO et al., 2018). Both studies are high-risk pregnancy cohorts that enroll younger siblings of a child previously diagnosed with ASD, evaluating environmental risk factors and assessing the child development from birth to up to 36 months. In both studies, data and biological samples were collected throughout pregnancy, at birth, and until the child's third birthday. The biological samples included umbilical cord blood collected in a PAXgene Blood RNA tube, that contains an additive that yelds immediate long-term stabilization of RNA by reducing post-collection degradation and minimizing gene induction and repression, thereby yielding accurate and reproducible gene expression data when later sent for sequencing (RAINEN et al., 2002). The data was collected through standardized interviews and questionnaires, and the younger child development was assessed by trained, reliable examiners using standardized tests and an algorithmic approach to scoring.

2.1.2 Diagnosis of ASD

Given the complexity, severity, and symptom overlap of ASD with other psychiatric disorders, accurate diagnosis requires the use of appropriate instruments and scales. Effective clinical management of ASD depends on thorough assessments, including interviews with parents or caregivers, direct patient interviews, observational assessments, and an in-depth review of family history for ASD or other neurodevelopmental disorders (SHARMA; GONDA; TARAZI, 2018). Some neural screenings like Magnetic Resonance Imaging (MRI) can facilitate understanding of how the brain structurally and functionally develops differently in people with autism, but, to date, these results are not definitive (LORD et al., 2020).

Standardized neurodevelopmental assessments applied by professionals are the current norm for a diagnosis. Some typical tool utilized in children are Autism Diagnostic Observation Schedule (ADOS) (BRYSON et al., 2008), Autism Diagnostic Interview-Revised (ADI-R) (LORD; RUTTER; COUTEUR, 1994), Modified Checklist for Autism in Toddlers (M-CHAT) (WRIGHT; POULIN-DUBOIS, 2014), Early Screening of Autistic Traits (ESAT) (SWINKELS et al., 2006), and Mullen Scales of Early Learning (MSEL) (MULLEN, 1995).

Diagnostic methods for identifying autism in adulthood are still developing, as this is a relatively recent focus. Currently, clinical approaches primarily extend techniques from methods designed for diagnosing autism in children to adults, often relying on developmental information from their childhood. ADOS Module 4 (HUS; LORD, 2014), for example, has shown to have good specificity and sensitivity. However, since undiagnosed autistic adults seeking an autism assessment are often also dealing with co-occurring mental health disorders, any assessment method must effectively distinguish between symptoms and behaviors related to autism and those arising from other mental health issues (LORD et al., 2020), meaning any diagnostic method developed for adults must be able to differentiate symptoms of ASD from those of other co-occurring disorders.

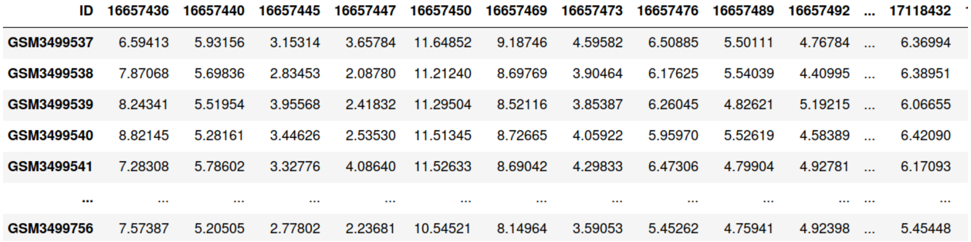
2.2 Gene expression

Genes encode proteins and proteins dictate cell function. Gene expression is the process by which the information in a gene is transformed into a functional product, usually through the transcription of RNA molecules that code for proteins or non-coding RNAs with various roles. Gene expression regulates the timing, location, and quantity of RNA and protein production. It is strictly regulated, changing significantly across cell types and different conditions. Additionally, the RNA and protein products of some genes often regulate the expression of other genes. Transcriptomics is the study an organism's transcriptome, which encompasses all of its RNA transcripts. The transcriptome provides a snapshot of the total transcripts present in a cell at a specific moment (LOWE et al., 2017).

There are different techniques in this field to study the gene expression, microarrays being the one used in this work. The main objective of most microarray experiments is to analyze gene expression patterns by measuring the expression levels of thousands of genes in a single assay. Typically, RNA is extracted from different tissues, then labeled and hybridized to the arrays. This approach enables the expression levels to be assayed between appropriate sample pairs. After hybridization, the arrays are scanned to produce grayscale images for each sample pair, which must be analyzed to identify the array spots and measure the relative fluorescence intensities for each element. The basis of microarray analysis is that the measured intensities for each gene on the array reflect its relative expression level (QUACKENBUSH, 2002). An example of this can be seen in Figure 2.2, where the rows are patient samples, the columns identify each probe, and each value represent the (already normalized) measured fluorescence intensities that represent a certain gene expression for a certain instance.

Studying biological factors that combine both genetic and environmental influences, such as genome-wide gene expression levels (transcriptome), can offer valuable insights into the developmental pathophysiology of ASD, reason why several studies have

Figure 2.2: Exemple of microarray, where the rows are patient samples, the columns identify each probe, and each value represent the (already normalized) measured fluorescence intensities that represent a certain gene expression for a certain instance.



ID	16657436	16657440	16657445	16657447	16657450	16657469	16657473	16657476	16657489	16657492	...	17118432
GSM3499537	6.59413	5.93156	3.15314	3.65784	11.64852	9.18746	4.59582	6.50885	5.50111	4.76784	...	6.36994
GSM3499538	7.87068	5.69836	2.83453	2.08780	11.21240	8.69769	3.90464	6.17625	5.54039	4.40995	...	6.38951
GSM3499539	8.24341	5.51954	3.95568	2.41832	11.29504	8.52116	3.85387	6.26045	4.82621	5.19215	...	6.06655
GSM3499540	8.82145	5.28161	3.44626	2.53530	11.51345	8.72665	4.05922	5.95970	5.52619	4.58389	...	6.42090
GSM3499541	7.28308	5.78602	3.32776	4.08640	11.52633	8.69042	4.29833	6.47306	4.79904	4.92781	...	6.17093
...	...	...	...	...	...	...	...	...	...	...	...	...
GSM3499756	7.57387	5.20505	2.77802	2.23681	10.54521	8.14964	3.59053	5.45262	4.75941	4.92398	...	5.45448

Source: The Authors

already examined gene expression differences in ASD children specially in blood samples, like Tylee et al. (2017), Enstrom et al. (2009), Gregg et al. (2008) and Kong et al. (2013). That is because immune-related alterations have consistently been observed in ASD patients, including signs of immune dysregulation and autoimmune phenomena. The use of gene expression in blood, then, is expected to provide insight into ASD-related transcriptional differences in circulating immune cells.

Like mentioned in Section 2.1, environmental factors may play a role in prenatal, perinatal, and very early postnatal development associated to ASD. Umbilical cord blood provides a snapshot of fetal blood and the exchanges across the fetoplacental unit, offering a unique perspective into prenatal development. It contains a very specific mixture of cells that reflects the immune response, as well as endocrine and cellular communication crucial for fetal development around the time of birth. This idea comes from Mordaunt et al. (2019), inspiration for this research, who proposed a meta analysis on this data to further our understanding of perinatal transcriptional changes that occur before an ASD diagnosis in high-risk children.

2.3 Machine Learning

According to Goodfellow, Bengio and Courville (2016), Machine Learning systems are those capable of learning by extracting patterns from raw data and making decisions based on them. But to fully define a machine that learns it is important to define learning in this context. One such definition is that any program is considered to be learning if it improves its performance through experience acquired from data (MITCHELL, 1997).

ML is usually divided into three categories, based on the kind of learning that is being applied: supervised learning, unsupervised learning and reinforcement learning. This work applies a supervised learning approach, meaning the program receive a dataset where each instance has its defining information (called its features or attributes) and a target, and the objective of the program is to learn how to best map each input to a suitable target based on its features (MURPHY, 2012). Supervised learning can be divided in two kinds of learning tasks, classification, where the target is categorical and the objective is to assign a category to each item, and regression, where the target is a real-valued scalar and the objective is predicting a real value for each item (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012).

This work is projected as a classification problem. Specifically, we develop a binary model for classification on biomedical data, where the input data is gene expression information from patients in an ASD study, the features being each individual probe, and the target labels, or classes, for the output are related to the subjects final diagnosis.

ML has been used in biology for decades, and currently it is used in nearly every field of biology. Recently, though, there has been na effort to better match methods to certain scenarios. There are two main goals when using ML in biology, which are to make accurate predictions, and to further our understanding of biology (GREENER et al., 2022).

2.3.1 Algorithms

In total, eight ML classification algorithms were used inin each classification step of this work. Each classification algorithm was trained twice, once for each step, in the same data, but on different target sets.

K-Nearest Neighbors (KNN): Is an instance-based method that assumes all instances correspond to points in the n-dimensional space (n being the number of features), and attributes a class to the predicted instance based on a vote of the classes of the K nearest instances in that space, the nearness being given by a distance function (MITCHELL, 1997).

Naive Bayes (NB): A probability-based classifier that assigns to a new instance the most probable target value based on the assumption that the attribute values are conditionally independent, meaning that given the target value of the instance, the probability of observing a certain combination of attribute values is the product of the probabilities for each individual attribute (MITCHELL, 1997). In the scope of the current research, since all the feature are real-valued, the Gaussian distribution was used, meaning the likelihood of the features is assumed to be Gaussian function (MURPHY, 2012).

Logistic Regression (LR): A method that uses a sigmoid function that takes input as independent variables and produces a probability value between 0 and 1, to classify an instances based on a threshold for that value (MURPHY, 2012).

Support Vector Machine (SVM): A kernel method where each data point is a vector so that the input vectors are mapped to a high-dimension feature space, where a decision boundary (whose location is determined by a subset of the data points known as support vectors) is chosen by maximizing the margin, which is defined as the smallest distance between the decision boundary and any of the samples. It performs linear classification, but also non-linear classification with the use of kernel functions (CORTES; VAPNIK, 1995) (BISHOP, 2006).

Multi-layer Perceptron (MLP): Also known as the feed-forward neural network, consists of a input layer that represents the input features, one or more hidden layers that transforms the values from the previous layer with a weighted linear summation followed by a non-linear activation function, and an output layer that receives the values from the last hidden layer and transforms them into output values (BISHOP, 2006).

Random Forest (RF): An ensemble learning method based on the aggregation of numerous decision trees built on independent identically distributed random vectors, where in the end the predicted target for a certain input is defined through a vote of all trees' outputs (BREIMAN, 2001). Decision trees refer to a hierarchical model of decisions and their consequences. They are used to classify an instance into a predefined set of classes based on a division of their attributes values, in hierarchical order of attribute importance (ROKACH; MAIMON, 2014).

Extreme Gradient Boosting (XGBoost): An ensemble tree based method that utilizes an optimized version of gradient boosting. Gradient boosting is an approach where multiple weak models (in this case decision trees) are combined in a sequence so that each new model corrects mistakes of the previous ones by optimizing the loss function. XGBoost optimizes the approach for both hardware and software, having as some key improvements the parallelization of trees, regularization to reduce overfitting, and its flexibility by having a large number of adjustable hyperparameters (CHEN; GUESTRIN, 2016).

Light Gradient-Boosting Machine (LightGBM): Similar to XGBoost, it is a variation of gradient boosting decision trees that aims at an optimized model with a more efficient hardware usage. LightGBM has some of the same advantages as XGBoost as well as two novel techniques, Gradient-based One-Side Sampling (GOSS), which selectively retains instances with large gradients during training, and Exclusive Feature Bundling (EFB), a near-lossless method to reduce the number of effective feature (KE et al., 2017).

2.3.2 Model evaluation

In this Subsection, the methods used in this work for evaluating an ML model's performance are discussed. This includes the evaluation metrics, described in Section 2.3.3 and cross-validation, in Section 2.3.4.

2.3.3 Performance metrics

It is important to track measures of a model's performance in order to assess its abilities, usually through quantitative measures. To characterize the performance of this model, six metrics were used: accuracy, precision, recall, F1 score, F-beta score and the ROC AUC (area under the receiver operating characteristic curve). All these metrics are better understood through the concept of a confusion matrix, that can be formally defined as following (OPITZ, 2024):

$$m_{ij}^{f,S} = |\{s \in S | f(s_1) = i \wedge s_2 = j\}|^2 \tag{2.1}$$

where

- $m_{ij}^{f,S}$ is the confusion matrix $m_{ij}^{f,S} \in \mathbb{R}_{\leq 0}^{n \times n}$
- f is any classifier $f : D \to C = \{1, \cdots, n\}$
- S is the sample set, a finite set $S \subseteq D \times C$
- i is the predicted labels
- j is the true labels
- D is the set of possible data entries

- C is the set of target classes
- n is the number of classes in the problem

In simpler terms, a confusion matrix is such that each entry t, p is equal to the number of observations known to be in group t and predicted to be in group p. In Table 2.1 an example can be seen, where **TP** is the True Positive, the inputs from class Positive predicted correctly as class Positive, **TN** is the True Negative, the inputs from class Negative predicted correctly as class Negative, **FP** is the False Positive, the inputs from class Negative predicted incorrectly as class Positive, and **FN** is the False Negative, the inputs from class Positive predicted incorrectly as class Negative. The sum of all 4 values results in the complete population $n = TP + TN + FP + FN$.

Table 2.1: Example of confusion matrix, where **TP** is the True Positive, **TN** is the True Negative, **FP** is the False Positive, and **FN** is the False Negative.

		Predicted	
		Positive	Negative
True	Positive	TP	FN
	Negative	FP	TN

Source: The Authors

Based on the notions presented for the confusion matrix, the six chosen metrics can be explained as following:

Accuracy: is the rate of instances correctly predicted by the model, defined as

$$Acc = \frac{TP + TN}{n}$$

Precision: shows how often the model is correct when predicting an instance as Positive, defined as

$$Prec = \frac{TP}{TP + FP}$$

Recall: also known as **sensitivity** or **True Positive Rate** (TPR), shows how often the model predicts the Positive instances as such, and is defined as

$$Recall = \frac{TP}{TP + FN}$$

F1 score: also known as **F-measure**, can be interpreted as a harmonic mean of the precision and recall, and is defined as

$$F1 = \frac{1 \times TP}{2 \times TP + FP + FN}$$

F-beta score: can be interpreted as the weighted harmonic mean of precision and recall, where $\beta \geq 0$, $\beta < 1$ gives more weight to precision, $\beta > 1$ gives more weight to recall, and $\beta = 1$ is the same as the F1 score. It is defined as

$$F_\beta = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + FP + \beta^2 FN}$$

ROC AUC: compute Area Under the Receiver Operating Characteristic Curve (ROC AUC), the ROC curve being the plot of the he fraction of true positives out of the positives against the fraction of false positives out of the negatives at various discrimination threshold settings.

Another important concept surrounding the CM is **Specificity**, also called **True Negative Rate** (TNR). It is the probability of a negative test result among the Negative instances, defined as

$$TNR = \frac{TN}{TN + FP}$$

The choice of performance measures typically reflects the expected characteristics and behavior of the model being evaluated. In the context of this work a large set of metrics can be justified in the intention to fully understand the model proposed, but in the sense of enhancing the model's performance a single metric was chosen as optimization goal: **F-beta score.**

2.3.4 Cross-validation

It is important to evaluate how the model performs on data that it has not seen before, since this determines its performance in real-world cases. To properly do that the dataset used must be divided, so that the model can be trained in one part (a Train set) of the dataset and evaluated with the other (a Test set), but the proportion of such division must be so that the Train set is large enough to properly learn the input patterns and the Test set is not too small that the statistical uncertainty of the computed performance metrics may make it difficult to claim that the result is truly representative of the model's capabilities (GOODFELLOW; BENGIO; COURVILLE, 2016).

The performance of any ML algorithm depends heavily on the given representation of the data (GOODFELLOW; BENGIO; COURVILLE, 2016). As Libbrecht and Noble (2015) explain, data from medical or biological applications, like genomic, epigenomic, transcriptomic, proteomic or metabolomic data are high dimensional by default, and as so are prone to some difficulties that come with high dimensionality, like poor generalization. Not only that, but they also tend to have a small number of data points, which is the case of the dataset collected for this work. Besides all that, the proposed model has a two-part classification process, and multiple classifiers have their hyperparameters tuned to each step of said process.

Because of the architecture of this experiment and the low amount of data used, cross-validation was chosen as a solution. Cross-validation (CV) is the process of dividing a whole dataset in K equal parts, called folds, and then for each fold $k = \{1, \cdots, K\}$ the algorithm is trained on all folds but k'th (GOODFELLOW; BENGIO; COURVILLE, 2016). In the context of this work, a holdout 80/20 was applied in the whole dataset, and a nested cross-validation was used on the Train set, all stratified for all 3 classes. That means the whole dataset was divided keeping 80% to train on and 20% to test, and then on the Train dataset was applied a 5-fold CV, followed by a 3-fold CV on the training portion of each iteration of the 5-fold one, always keeping the proportion of samples between classes. The classifiers' hyperparameters are tuned inside a 3-fold CV, and then trained and validated for the two-step process on the outer 5-fold CV. The whole process in then tested on the remaining Test set (separated through holdout).

3 RELATED WORKS

This Chapter presents other works related to this research, some in the study of ASD through ML with gene expression data and some in the findings of works utilizing blood expression to research ASD.

It is impossible to talk about related works without mentioning Mordaunt et al. (2019), that provided the pre-processed data for this research and has served as motivation for our work as the first study to identify gene expression differences in cord blood specific to ASD through a meta-analysis. They used genome-wide transcript levels measured by Affymetrix Human Gene 2.0 array in RNA from cord blood samples from both MARBLES and EARLI, two high-risk pregnancy cohorts previously mentioned in Section 2.1, to compare gene expression between ASD and TD, and between TD and children that were not ASD but not TD either (Non-TD), within each study and via meta-analysis, through differential expression analysis and weighted gene correlation network analysis. The authors discovered that while gene expression differences in cord blood between ASD or Non-TD and TD groups did not reach genome-wide significance, 172 genes were differentially expressed between ASD and TD and 66 genes between Non-TD and TD, including 8 genes that were differentially expressed in both comparisons. They also identified demographic factors and cell type proportions significantly correlated to the gene expression modules. Their final conclusion was that, since umbilical cord blood is not the main affected tissue in ASD, it is composed of many cell types, and ASD is a heterogeneous disorder, the expression differences found were subtle but the enriched gene pathways support involvement of environmental, immune, and epigenetic mechanisms in ASD etiology.

ML has been applied in many ways regarding ASD, including ASD identification like this work, except usually on standardized assessments, kinesthetic data, and neuroimaging data. Omar et al. (2019), for example, proposed a prediction model based on ML technique for identifying ASD in people of any age. The authors used 250 instances collected from people with and without autistic traits and the AQ-10 dataset, that combines personal and social information with the results from the AQ-10 tool (Autism Spectrum Quotient), used to identify whether an individual of any age should be referred for a comprehensive autism assessment. The results showed that the proposed prediction model provides good results in terms of accuracy, specificity, sensitivity, precision, and false positive rate (FPR) for both kinds of datasets.

Liu, Li and Yi (2016) applied ML to classify children as having ASD or not based on the analysis of an eye movement dataset, reporting that their study is preliminary but the results show promising evidence. Zhang et al. (2018) used feature extraction and classification techniques to identify ASD in male children through white matter abnormality via identification of discriminative fiber tracts in diffusion magnetic resonance imaging (dMRI). Neuroimaging has also been used to predict symptom severity, like Moradi et al. (2017), who propose using regression techniques to predict ASD symptom severity based on cortical thickness with smaller samples than other approaches, and demonstrate its usefulness with the Autism Brain Imaging Data Exchange (ABIDE) database.

Regarding gene expression and ML, both Lin et al. (2020) and Gök (2019) employ ML-based approaches to predict ASD risk genes based on brain-related data. LIN et al. used supervised ML methods to predict autism risk genes from their spatiotemporal expression signatures, network topology features, gene-level constraint metrics, and other general gene features. GöK used genes expression data related to brain development and proposed a ML model composed of feature extraction, followed by discretization methods and a classification algorithm. Both works presented interesting results when predicting known and new candidate genes for ASD.

Some work was also found on ASD and blood gene expression, although not all with ML techniques. In Glatt et al. (2012), an investigation for blood-based gene expression signatures in ASD children was performed, finding potential biomarkers for ASD in one half of the sample set through an analysis of covariance (ANCOVA) and using the biomarkers to build a classifier tested in the remaining data. The identified genes went through an enrichment analysis to identify particular functional, ontological, or structural annotations. Tylee et al. (2017) combined the subject-level data from previously published studies in blood transcriptomics by mega-analysis to increase the statistical power, processing data with uniform methods. Covariate-controlled mixed-effect linear models were used to identify gene transcripts and co-expression network modules that were associated with ASD diagnosis. Moreover, permutation-based gene-set analysis was used to identify functionally related sets of genes that were over- and under-expressed among ASD samples, and evidence for sex-differences in the ASD-related transcriptomic signature was explored. Finally, the authors demonstrated that machine-learning classifiers using blood transcriptome data perform with moderate accuracy when data are combined across studies.

Enstrom et al. (2009) performed a gene expression screening and cellular functional analysis on peripheral blood from children with ASD and TD children, finding abnormalities in natural killer cells that could represent a predisposition in ASD to develop autoimmunity and/or adverse neuroimmune interactions during critical periods of development.

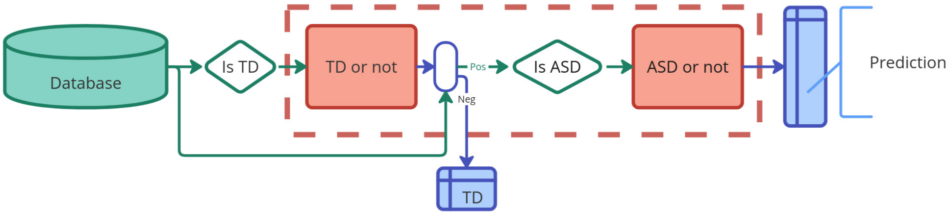
Using classification algorithms in ASD data is not new, and it has shown promising results. However, few works use gene expression for diagnosis classification, especially blood gene expression. Published researches propose the existence of information on this kind of data that could, in theory, be learned. Specific to umbilical cord blood in ASD cases, only Mordaunt et al. (2019) was found, so this work proposes a new kind of exploration in the field of ASD research by using such type of data in a diagnosis classification problem.

4 EXPERIMENTS

In broad terms, the problem being tackled by this work is of trying to use gene expression data to classify samples as belonging to ASD patients or not. As seen previously in Section 2.3, it is already hard to extract relevant information from biological data because of high dimensionality and small number of data points, and ASD is a particularly complex case, being a heterogeneous disorder for which a lot of the genetic characteristics are still unknown, as seen in Section 2.1. For these reasons, the approach used to create an ASD early diagnosis support model was dividing the major problem into two tasks: first classify the patients into typical and atypical development, and then separate those classified as having an atypical development into ASD and not ASD.

Figure 4.1 shows an overview of the proposed architecture, where data is used as input for the first smaller model, whose output is used to filter the same data, that is then used as input of the second smaller model, whose output is the final prediction. In Figure 4.2 we show the inside of each smaller model, where 8 classifiers are trained and then have their predictions combined using soft voting.

Figure 4.1: Architecture of the proposed classifier, where in green we have the dataset, transformed for each individual task, in red the models equivalent to each task, the output of the first serving to filter the input data to the next, only keeping instances not predicted as TD, and finally, in blue, the final prediction.

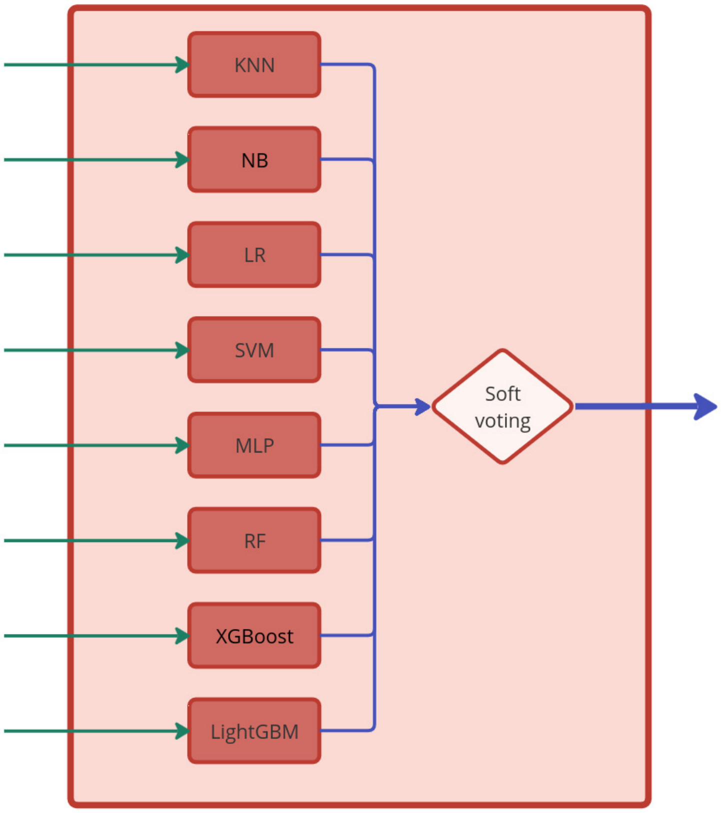


Source: The Authors

To implement that, a preprocessed gene expression dataset from umbilical cord blood samples was used, with the number of features reduced by 86%, as explained in Section 4.1. Section 4.2 details the proposed experiment, that consisted of:

1. Dividing the dataset;
2. Scaling the data;
3. Choosing 8 classifiers;
4. Tuning and training said classifiers for the first task;
5. Combining the predictions into a single one;

Figure 4.2: Inner architecture of each model of the proposed classifier, where 8 distinct classifiers are trained in the same data and then have their predictions aggregated through soft voting into a final prediction.



Source: The Authors

6. Using the combined prediction to filter the dataset for the second task;
7. Tuning and training the classifiers for the second task;
8. Combining the predictions into a final one.

In Section 4.3 it is further explained how the aforementioned model was evaluated, by choosing which metrics to consider, analysing and selecting the best hyperparameters for each classifier, and then training on the Train set and evaluating on the Test set.

4.1 Dataset and dimensionality reduction

Our experiments (described in details in Section 4.2) were conducted using a preprocessed dataset collected from Mordaunt et al. (2019). The data consisted of genome-wide transcript levels measured in umbilical cord blood samples from two studies in early ASD, EARLI (NEWSCHAFFER et al., 2012) and MARBLES (HERTZ-PICCIOTTO et al., 2018). Both studies collected the blood sample at birth and carried 36 month assessments, with an algorithm-based child diagnosis, from high-risk pregnancies based on the knowledge that the families already had a older child previously diagnosed with ASD. The younger siblings were categorized as either ASD, TD, or Non-TD.

The preprocessing applied by the authors consisted of assessing signal distribution within each study, quality control, normalization, probe annotation and filtering, and finally surrogate variable analysis. Signal distribution was assessed in perfect-match probe intensity and robust multi-chip average (RMA) normalized data. For quality control, outliers were identified and those that did not meet certain criteria were excluded, alongside subjects without a diagnosis by the time of the study. Probes were annotated at the transcript level, and those not assigned to a gene were excluded as well. Surrrogate variable analysis was then used to estimate and adjust for factors that may have substantial effects on gene expression. From the resulting 271 samples preprocessed in the study, only the data from those who consented to share was used.

From the original data the only information kept for this work was the gene expression, the probes' identifiers (used as feature names), a depersonalized identification for each patient (used as each sample's id), and the diagnosis (which was used as the target class). In total there were 36459 attributes and 224 samples, divided in three classes:

- 53 ASD
- 80 Non-TD
- 91 TD

Besides the imbalance seen in the classes (that was judged not significant enough to justify altering through data augmentation, resampling techniques, or any similar approach) the collected data results in a very high dimensional dataset. As mentioned before, transcriptomic data like the one applied to this work are high dimensional by default, and as such are prone to some difficulties. For this reason, a few techniques were applied to lower the dimensionality by reducing the amount of features.

The resulting dataset was first filtered based on information collected about the platform used in the sequencing process, the "[HuGene-2_0-st] Affymetrix Human Gene 2.0 ST Array [transcript (gene) version]". Using the column **"GB_ACC"** on the platform table (that specifies the *GenBank* and *RefSeq* Accessions for each transcript associated to a probe) only the probes annotated with the prefix "NM" were kept as features, since that means that it is associated to mRNA. That resulted in a reduction of 65% in the number of features, leaving 12896 columns in the dataset.

The remaining features were ranked and filtered by the top 5000 features with the highest variance, which is defined as the following:

$$S^2 = \frac{\sum(x_i - \mu)^2}{n - 1} \tag{4.1}$$

Where:

- S^2 is the variance;
- n is the number of samples
- x_i is the ith data point
- μ is the sample mean

4.2 Methods

Besides the preprocessing already implemented in the collected dataset and the transformations done as described previously in Section 4.1, a few other manipulations, as described in Section 4.2.1, were necessary as part of being a dataset used for a ML classification problem.

Each of the eight classification algorithms mentioned in Section 2.3.1 was tuned and trained twice, once for each task, resulting in 16 distinct trained models (2 of each algorithm) for each fold of the cross validation. The hyperparameter tuning explained in Section 4.2.2 was performed with the same hyperparameter list for each task, the only difference being in the data.

Since the original dataset consists of 3 classes and each task consists of 2 distinct classes, a simple transformation was applied to the target list before each step's implementation: for the first step's implementation the class 'TD' was set to target 0, and both classes 'Non-TD' and 'ASD' were set to 1; and for the second step, both 'TD' and 'Non-TD' were set to target 0 and class 'ASD' was set to 1. The second step does not expect a class 'TD', but considering that after the first model's aggregation, described in Section 4.2.3, when the dataset is filtered to keep only the samples classified as being related to an atypical development there is the possibility of misclassified samples, those were simply labeled as not ASD.

4.2.1 Data preprocessing

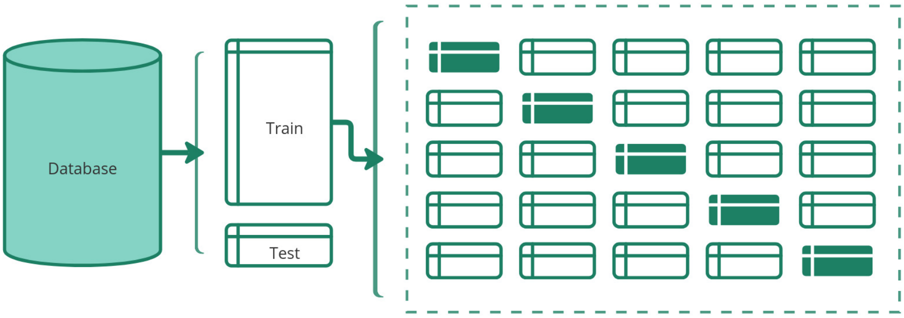
The reduced dataset, with 224 samples and 5000 features, was divided in two subsets using the holdout 80/20 method, meaning the original dataset was divided in a training set of 80% and a test set of 20% of the whole sample set, both with the same ratio of each class as shown in Table 4.1. For the Train set, 5-fold stratified cross validation was used, with a nested 3-fold stratified cross validation for the hyperparameter tuning. All these steps are visible in Figure 4.3, that illustrates the data division through the holdout and then the k-fold CV, and Figure 4.4, that illustrates what takes place at each iteration of the 5-fold CV.

Table 4.1: Number of samples of each class in each dataset before (Total) and after (Train, Test) the stratified holdout 80/20 split.

Dataset	ASD	Non-TD	TD	Samples
Train	42	64	73	179
Test	11	16	18	45
Total	53	80	91	224

Source: The Author

Figure 4.3: Division of the data in train and test with the holdout split, followed by the division of the train set in 5 folds for the CV, where for each iteration a different dataset is used for validation until every fold has been the validation set once.



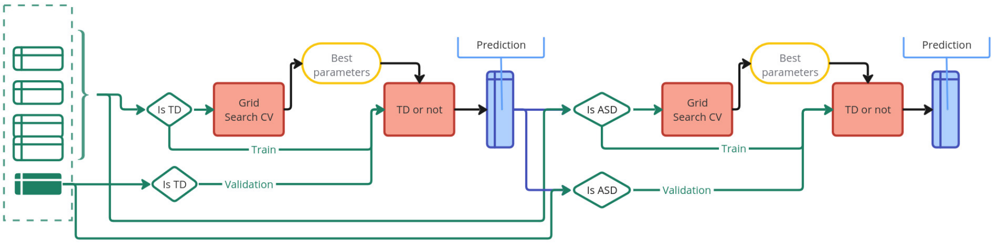
Source: The Authors

All classifiers had their own pipeline consisting of a scaler and the classifier itself. The scaler of choice for all eight classifiers was Scikit Learn's "StandardScaler", that scales each feature independently and centers the values around the mean at unit variance, following the formula bellow for each sample:

$$z = \frac{(x_i - \mu)}{\sigma} \tag{4.2}$$

Where:

Figure 4.4: CV iteration, where for each model 4 folds are used in a grid search, which yields the hyperparameters for each classifier that is then trained on the same 4 folds and validated in the fifth one. The prediction of the first model filters the data, which is transformed for the second model, that follows the same steps as the first.



Source: The Authors

- z is the scaled sample
- x_i is the current sample
- μ is the mean of the training samples
- σ is the standard deviation of the training samples

4.2.2 Hyperparameter tuning

Each classifier also had a set of hyperparameters to be tuned using a grid search, meaning an exhaustive search for the best combination of hyperparameters was performed for each classifier at each task, over a predefined set of values, in a 3-fold cross-validation within each iteration of the 5-fold cross validation.

In order to choose the best hyperparameter for each model it is first important to separate the hyperparameters in 3 types: nominal, integers and real numbers. For each type a different strategy was implemented.

For **nominal** hyperparameters, defined here as the those where each possible value is independent, meaning there is no natural ordering between values. For this kind of hyperparameters the statistical mode was used, so the frequency was calculated and ranked, and the value with the highest frequency was chosen, and in case of a tie the first on the list is chosen.

Integer hyperparameters are here defined as those that have a numeric value, but that value does not allow for fracions, like a discrete quantity of something (eg. number of estimators). For those hyperparameters the median value was used in the final classifier.

As for **real numbers** hyperparameters, these are hyperparameters where its values are any number within an interval, meaning that any value from the set of real numbers $\mathbb{R}$

is allowed, as long as it is within the hyperparameter's limits. For these cases, the mean was used as final value.

4.2.3 Aggregation

The metric used to determine the best estimator from the grid searches was the F-beta score. All estimators, with their determined best hyperparameters, were fit with the whole Train dataset and a prediction was made, resulting in eight independent target values for each sample. In order to combine all outputs into a single prediction for each sample a soft voting was performed, meaning the average predicted probability for each class was calculated based on the predicted probability of each individual model, and the class with the highest probability was selected.

For algorithms like KNN and SVM a probabilistic output is not obvious. KNN takes a voting of the K nearest points' classes to decide a new instance's class. In order to return the probability of the instance being of each class instead of a label prediction the algorithm calculates the ratio of the K nearest instances that belongs to each class. In the case of SVM, the algorithm maps the input data into high-dimensional feature spaces where linear classification can be performed. To return a probability instead of a target label, the algorithm performs a logistic regression on the SVM's scores, fit by an additional cross-validation on the training data.

There were two main reasons for choosing a soft voting approach instead of a traditional hard voting. The first being to better avoid ties, since the voting is performed on a even number of outputs, and the second being that different algorithms may be better at identifying different sets of instances, reporting higher probability in certain cases. A soft voting takes better advantage of each algorithm's strong points, leading to more robust results.

4.3 Evaluation

The main metric of interest chosen to be maximized in any decision making process was the F-beta score, with $\beta = 1.5$. Because of the low proportion of ASD samples it's imperative to judge the models based on metrics that look specifically to the class of interest. By making a model that prioritizes precision and recall, giving a higher weight to recall, the expectation is avoiding the case where a model would just predict most or every case as being negative, in the context of this model meaning classifying most or all cases as not having ASD.

In order to properly evaluate the model with the Test set, the classifiers must be trained with the whole Train set, all 179 samples as shown in Table 4.1. For that to occur the classifiers must have determined hyperparameter values. Since in the cross-validation each fold had its own set of best hyperparameter a selection strategy needs to be applied to determine the best overall value for each hyperparameter of each classifier. That strategy, already described in Section 4.2.2, is applied as a first step in the validation process.

The classifiers are then trained in the Train dataset as described in the previous section (Section 4.2). The set of models for the first task are trained on the whole Train data, with the target being TD or not and their predictions are aggregated. Next, the combined prediction is used to filter the Train dataset for the second task and the set of models for this classification task (i.e., having as target being ASD or not) are trained on the whole Train data, Finally, their predictions are aggregated in a final prediction, used for labeling a new instance.

In every possible step of the evaluation process, all metrics previously mentioned were calculated in order to characterize and judge the whole method.

4.4 Computational resources

The model was developed and tested in Python 3.10 (ROSSUM; JR, 1995), executed in a jupyter Notebook inside a virtual environment. The operating system was a Ubuntu 22.04.4 LTS, 64 bits, GNOME version 42.9. The machine running the code had a motherboard ASUS Prime B450M Gamin/BR, with a multi-core processor AMD Ryzen 7 2700x eight-core processor × 16, 16GB of RAM memory plus 2.1GB of SWAP, disc storage of 2.3TB, and graphics card AMD Caicos (Radeon HD 6450/7450/8450 / R5 230 OEM).

Some specific Python packages were used during the implementation, like pandas (v2.0.3) (TEAM, 2023) for data analysis and manipulation; NumPy (v1.23.0) (HARRIS et al., 2020) for multi-dimensional arrays and matrices, along with mathematical functions to operate on them; Matplotlib (3.7.1) (CASWELL et al., 2023) for visualization of the results; SciPy (v1.8.0) (VIRTANEN et al., 2020) for extracting some statistical metrics; Scikit-learn (v1.2.2) (PEDREGOSA et al., 2011) for most ML related tasks; XGBoost (v2.0.3) (CHEN; GUESTRIN, 2016) for implementing the XGBoost algorithm in python; and LightGBM (v4.1.0) (KE et al., 2017) for implementing the LightGBM algorithm.

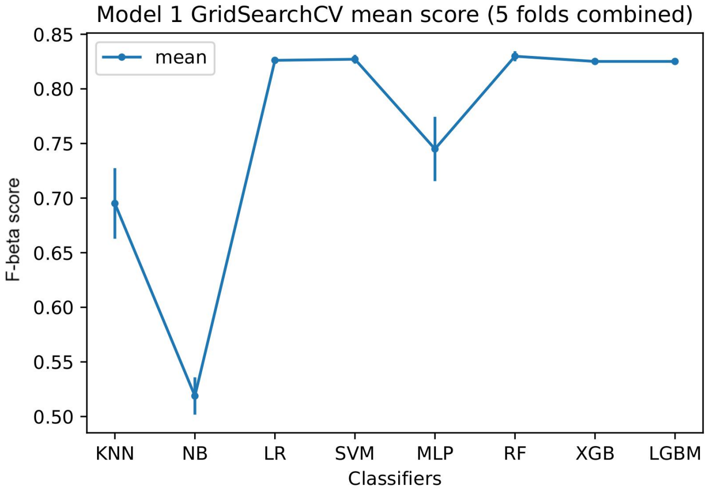
5 RESULTS

This chapter presents the results from the evaluation of the proposed model as described in Section 4.3. They will be reported in the following order: performance in the cross validation in Section 5.1, evaluation of the whole model in Section 5.2, and then a final general analysis of the model's behavior in Section 5.3. Throughout this chapter the part of the model regarding the first part of the problem being tackled (i.e., TD or not TD) will be referred as *model 1*, and the part regarding the second classification task (i.e., ASD or not ASD) will be referred as *model 2*.

5.1 Cross validation performance

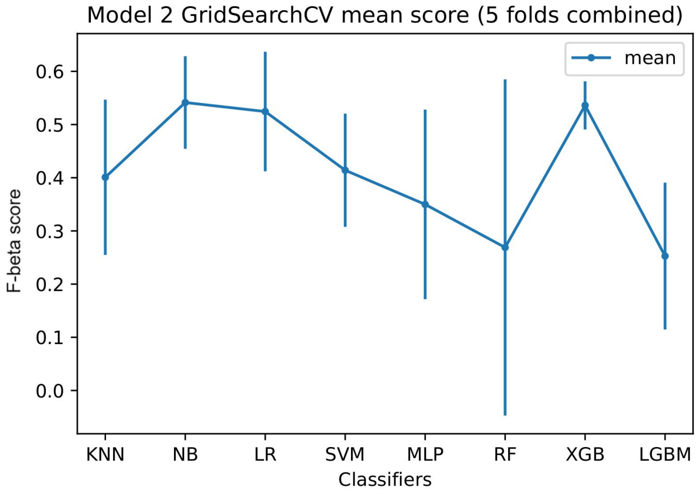
Within the 5-fold cross validation, not only the model was trained and validated but a grid search was performed in a 3-fold nested CV in each iteration in order to better tune the classifiers on the complex problem being tackled. Figure 5.1 and Figure 5.2, respectively referring to the grid search executed for model 1 and the grid search executed for model 2, show the mean and standard deviation of the best F-beta score for each individual classifier in the grid search. The individual score at this point of the training is not that illustrative of the final performance, since it was executed on a very small amount of data, but was used to better tune the models to their respective tasks.

Figure 5.1: Model 1 Grid Search summary, showing the mean (point) and standard deviation (vertical line) of the best score (axis Y) measured through F-beta score of each classifier (axis x) throughout the 5-fold CV.



Source: The Authors

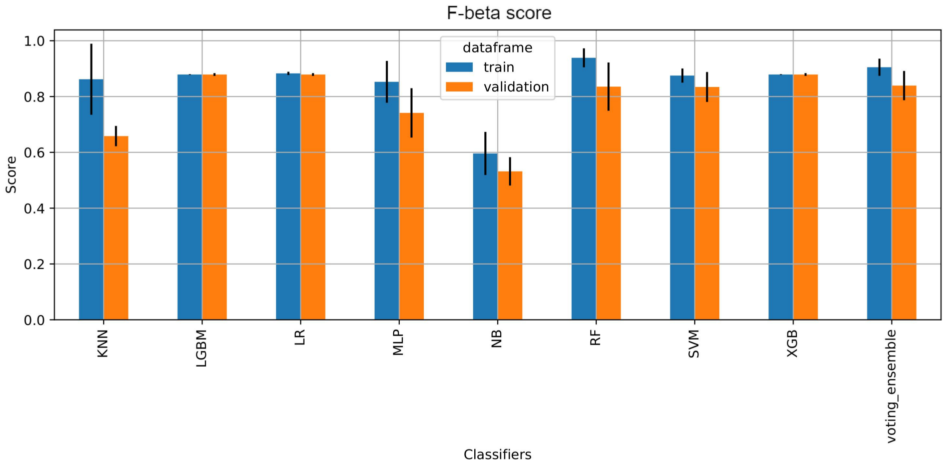
Figure 5.2: Model 2 Grid Search summary, showing the mean (point) and standard deviation (vertical line) of the best score (axis Y) measured through F-beta score of each classifier (axis x) throughout the 5-fold CV.



Source: The Authors

For a better understanding of the performance throughout the CV, we can observe the metrics results for each Train/Validation split. The mean and standard deviation of the F-beta score in the Train and Validation set through the 5-fold CV can be seen in Figure 5.3 for model 1 and Figure 5.4 for model 2. Similar Figures for the other metrics can be seen in Appendix A.

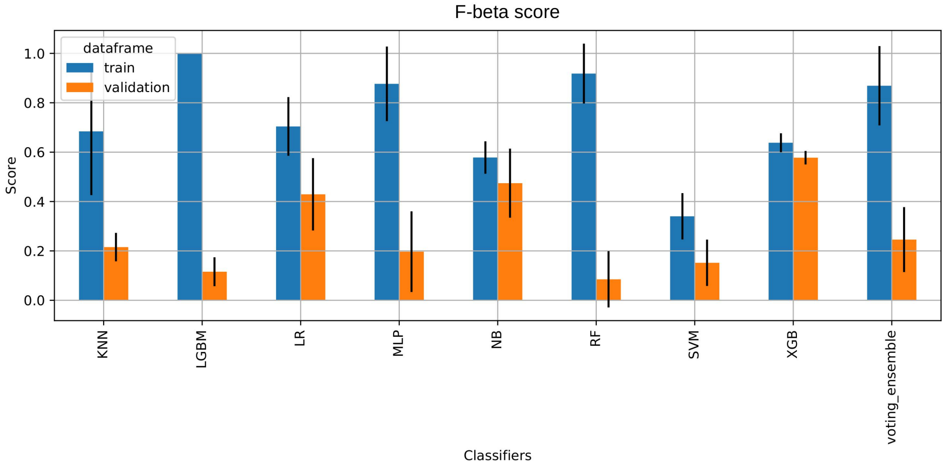
Figure 5.3: Mean (bar) and standard deviation (black line) of the performance measured in F-beta score (axis Y) in the CV for each individual classifier (axis X) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



Source: The Authors

The information collected in the grid search, combined to the performance metrics of each fold and the final measures that will be presented in Section 5.2, can help bring

Figure 5.4: Mean (bar) and standard deviation (black line) of the performance measured in F-beta score (axis Y) in the CV for each individual classifier (axis X) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



Source: The Authors

some better understanding to which classifiers are better suited to solve the proposed problem. More importantly it shows how both parts of the problem are different among themselves. XGBoost, Naive Bayes seem to yield the best performances for model 2 in the CV, both in the grid search and overall, with logistic regression also showing some promise compared to the rest, although not reaching the 0.5 threshold. On the other hand, for model 1, most classifiers present a decent result both in the grid search and overall, some more stable than others, except for Naive Bayes and KNN.

5.2 Final model evaluation

Because of the sequential nature of this two-part classifier, the performance of the whole model is seen in the performance of model 2 – remembering that in the scope of the second model, a positive label means an ASD diagnosis, and a negative label means it's not an ASD case. As can be seen in Table 5.1, the final model had a precision of 0.5 and F-beta score, the main metric of interest, performing in 0.54. Accuracy was the highest reported metric, meaning the model was able to predict correctly 75% of instances never seen. The confusion matrices for the Train set (Table 5.2) and for the Test set (Table 5.3) illustrate better the behavior of the model.

As can be seen, the reason for such high accuracy is the amount of negative instances predicted correctly, getting 27 out of the 33 right, resulting in a TNR of 82%.

Table 5.1: Performance metrics from **model 2** in the Train and Test dataset.

	Accuracy	Precision	Recall	F1 score	**F-beta score**	ROC-AUC
Train	0.98	0.95	0.95	0.95	**0.95**	0.97
Test	0.75	0.5	0.55	0.52	**0.54**	0.68

Source: The Authors

Table 5.2: Confusion matrix of **model 2** in the Train set.

		Predicted	
		Positive	Negative
True	Positive	40	2
	Negative	2	122

Source: The Authors

Given that in this problem 75% of all instances are negative, a higher TNR than TPR is expected, which is one of the reasons to use F-beta score as the main measure, training the models to be less skewed to the negative class. This approach was successful, considering the 75% of instances predicted correctly are well divided between classes. In the end, only 50% of cases predicted as ASD actually have it, but given that a patient actually has ASD, the model correctly predicts 55% of cases.

On the other hand, when looking at Table 5.4 referring to the performance of model 1, the performance appears to be a lot higher overall and interestingly very similar between Train set and Test set. In the scope of this first model, a positive label means a patient is not TD, or have not presented a typical development, and a negative label means it's a TD case. Model 1 presented a F-beta score of 0.89 and perfect recall, meaning every instance of Non-TD and ASD have been correctly identified and were part of the dataset for model 2. The precision of 0.61 can be better illustrated by the confusion matrices for this model, provided in Table 5.5 for Train set and in Table 5.6 for Test set.

Since a priority has been given to make sure all positive cases be identified so model 2 can properly classify them, the model is skewered to the positive class, which was expected from using F-beta score with the same value of β for both problems with different proportions between classes, considering in model 1 the positive label appears in 67% of instances and in model 2 only in 25%. That lead to a lower precision of 0.61 and an extremely low TNR of 12%, which means a higher proportion of noise in the dataset for model 2 caused by mislabeled instances from model 1, since the entry data to model 2 is the Train dataset filtered by model 1's prediction, and in turn difficulty to learn correct generalized patterns in model 2.

The predictions of both model 1 and 2 are the result of eight other models trained for each problem, so we can also look at the individual performance of each model as well. Figure 5.5 and Figure 5.6 show the F-beta score of each trained classifier for, respectively,

Table 5.3: Confusion matrix of **model 2** in the Test set.

		Predicted	
		Positive	Negative
True	Positive	6	5
	Negative	6	27

Source: The Authors

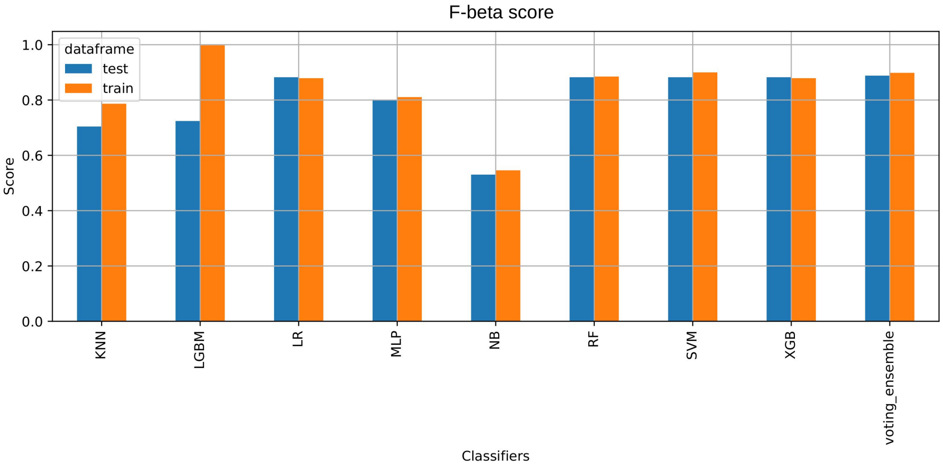
Table 5.4: Performance metrics from **model 1** in the Train and Test dataset.

	Accuracy	Precision	Recall	F1 score	**F-beta score**	ROC-AUC
Train	0.66	0.64	1.0	0.78	**0.90**	0.59
Test	0.62	0.61	1.0	0.76	**0.89**	0.53

Source: The Authors

model 1 and model 2. A characteristic observed both in the performance metrics of the whole and in the individual graphics for each metric is that model 1 shows very little difference from the Train set to the Test set, maybe because it highly prioritizes the positive label in both datasets. It can be seen in Figure 5.5 that characteristic extends through most classifiers used. Other similar Figures for each of the unreported metrics can be seen in appendix A.

Figure 5.5: Performance measured in F-beta score (axis Y) of each individual classifier (axis X) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



Source: The Authors

Figure 5.6, on the other hand, shows that model 2 owes its high difference from Train to Test to some of the classifiers used. Still it managed to perform better than most models in the Test set across all performance measures, showing a very stable behavior, as can be seen in Appendix A. The two models that presented a better F-beta score are the two ensemble methods, XGBoost and LightGBM, with the third ensemble method RF still amongst the better ones, which suggests either the data may be too noisy for simpler models or this specific problem might need a different architecture composed of a

Table 5.5: Confusion matrix of **model 1** in the Train set.

		Predicted	
		Positive	Negative
True	Positive	106	0
	Negative	60	13

Source: The Authors

Table 5.6: Confusion matrix of **model 1** in the Test set.

		Predicted	
		Positive	Negative
True	Positive	27	0
	Negative	17	1

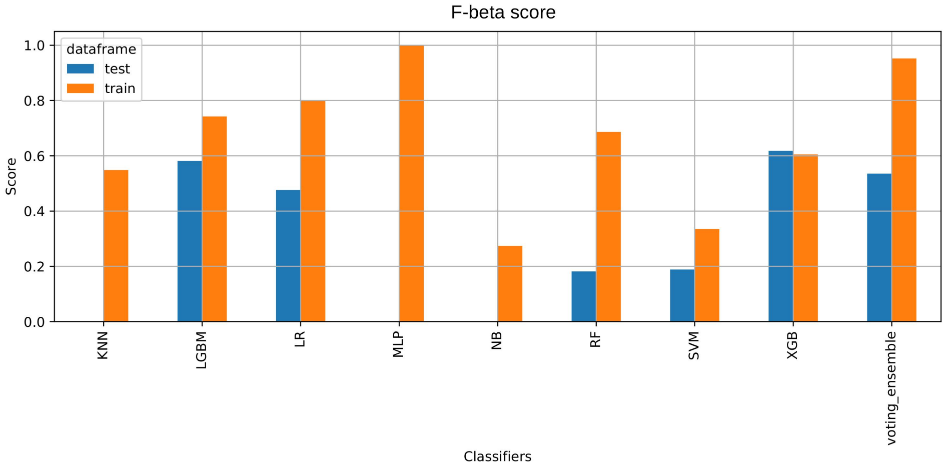
Source: The Authors

more specific set of classifiers in order to achieve precise and robust results. Interestingly, LightGBM was one of the worst performing classifiers in CV, suggesting it needed more data to properly learn. On the other hand, Naive Bayes was one of the best performing in CV, but on the final model had a F-beta score of 0 in the Test set, possibly due to classifying every instance as negative. The goal of using an ensemble approach was to take advantage of each classifier's strengths and avoid their issues. That is better illustrated in the graphs for all metrics in appendix A, but compared to each individual model the final model managed to perform decently in all metrics, as we've seen in Table 5.1. The other two models worth mentioning here are LR, better performing than RF and the only other classifier to offer a F-beta score comparable to the final model, and SVM with a sigmoid kernel, with the lowest F-beta score above 0. The only two kinds of classifiers that managed identify some ASD samples were the ensemble tree classifiers and the ones based on logistic function.

5.3 General analysis

This Section will elaborate in the overall behavior of the proposed model taking into account the main problem, and not its parts. Table 5.7 presents a view of the final classification of each instance as a multiclass confusion matrix. It can be seen that 55% of ASD subjects were correctly classified, and the other 45% was classified as Non-TD. Interestingly, only 17% of instances classified as ASD were actually TD, meaning most patterns found in model 2 may not be exclusive to ASD, but are consistent with atypical development. Of the Non-TD, 75% was correctly implicitly assigned, by not being classified as TD in model 1 or ASD in model 2. The focus of this work is exclusively

Figure 5.6: Performance measured in F-beta score (axis Y) of each individual classifier (axis X) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



Source: The Authors

to aid in the diagnosis of ASD, so this classification of Non-TD should not be taken as a pre-diagnosis of any sort, specially considering that 71% of all subjects were predicted as being Non-TD, including 15 out of the 18 TD cases.

Table 5.7: True and predicted class of all instances based on the original label.

		Predicted		
		ASD	Non-TD	TD
True	ASD	**6**	5	0
	Non-TD	4	**12**	0
	TD	2	15	**1**

Source: The Authors

6 CONCLUSION

Due to the increasing prevalence of ASD and the amount of recent discoveries in the field, several studies have been proposed investigating its genetic characteristics, the environmental components taking part in it, or the relationship between both parts. This work proposed a ML classifier for ASD based on gene expression data from umbilical cord blood, aiming at a possible support tool to aid in an early diagnosis of ASD.

Overall, the proposed model shows a relatively low performance to be used as a tool, with only 55% of all ASD cases correctly predicted, and those being only 45% of the cases predicted as being ASD, but managed to identify patterns in the data that imply the possibility of building one. The better performance in the first part of the model when compared to the second, may be partly due to a better proportion between classes, but the final results are consistent with the findings of Mordaunt et al. (2019). The authors found only subtle differences between classes, and those were mostly between ASD and TD and between Non-TD and TD.

This work has brought some interesting insight, though, through the results found. Three main points to be taken into account when judging the results are that the data may not be informative enough, the model may not be powerful enough, or the patterns in the data may be too complex to be learned though the ML algorithms proposed. As seen in Section 2.1, ASD is a very complex disorder, and a lot is unknown about its genetic components. That is corroborated in Chapter 3, where we see that most patterns found are subtle.

The first point talks about the data used. A frequent problem when using ML techniques in biological data, especially genetic, is the relatively low amount of data available and its high dimensionality. Some techniques were used to counteract its impacts in this work, like the dimensionality reduction, the separation of the problem in two parts, and the use of F-beta score, specially in the second part of the model, and these were important factors into the patterns that the model did learn. Still, considering the subtle patterns of the disorder, more data could better validate our approach for ASD. A different kind of data could be tried as well, like blood or saliva expression from the subject as a baby, instead of umbilical cord blood that has information about the mother as well, but the sample collected should still be from the baby if the intention is to suggest the observation and evaluation of the child from as early as possible. Another alternative to more or different data is also a deeper study in dimensionality reduction, including feature selection algorithms. Feature selection also brings the possibility of identifying ASD biomarker candidates in newborn babies.

The second point mentioned is regarding the choice of classifier. The idea of using multiple classifiers through an ensemble approach is promising, because it manages to take advantage of each classifier's different learning techniques, identifying different patterns and presenting more robust results, but there is a need to better choose the classifiers for each problem. The results show how both parts of the problem are different among themselves and in their needs from a classifier, suggesting that maybe proposing independent approaches for each problem might yield better results. There is here space to also conduct a deeper study on which algorithm to use. The ones chosen used in this work were chosen for being well established in the literature and mostly taking very different approaches to learning, and have brought insight as to which kind of learning is better suited for the problem. For the second part of our classifier, the best algorithms were the ensemble ones based on decision trees, and the ones based on a logistic distribution. That suggests different algorithms based on these kinds of learning may be a good choice for an ensemble approach, or even simply the use of just these classifiers, ignoring the other three selected.

Finally, the last point mentions the possibility of the patterns being too complex for traditional ML algorithms, which is not mentioned as saying ML could not be used, but to suggest that maybe more complex approaches, such as more elaborate ensemble approaches, graph-based approaches like gene co-expression networks, or deep learning algorithms may be able to learn these patterns better or more easily. These will still run into similar issues regarding the low amount of data and its high dimensionality, as all these issues are not independent, but it is an idea worth being explored.

The current work has its limitations as seen, but shows a promising starting point for other projects that aim to aid in the early diagnosis of ASD. Every limitation identified gives way to another possible future work to be explored. Working with something as complex as human beings is a naturally challenging task, especially when talking about something as complex as genetic material, so decisive results are rare, but that is only more reason to investigate and deepen our collective knowledge about it.

REFERENCES

ASSOCIATION, A. P. **Diagnostic and Statistical Manual of Mental Disorders**. American Psychiatric Association, 2013. ISBN 0890425574. Available from Internet: <http://dx.doi.org/10.1176/appi.books.9780890425596>.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 978-0-387-31073-2.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, oct. 2001. ISSN 1573-0565. Available from Internet: <https://doi.org/10.1023/A:1010933404324>.

BRYSON, S. E. et al. The autism observation scale for infants: Scale development and reliability data. **Journal of Autism and Developmental Disorders**, v. 38, n. 4, p. 731–738, abr. 2008. ISSN 1573-3432. Available from Internet: <https://doi.org/10.1007/s10803-007-0440-y>.

CASWELL, T. A. et al. **matplotlib/matplotlib: REL: v3.7.1**. Zenodo, 2023. Available from Internet: <https://zenodo.org/record/7697899>.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2016. (KDD ’16), p. 785–794. ISBN 978-1-4503-4232-2. Available from Internet: <http://doi.acm.org/10.1145/2939672.2939785>.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, sep. 1995. ISSN 1573-0565. Available from Internet: <https://doi.org/10.1007/BF00994018>.

ENSTROM, A. M. et al. Altered gene expression and function of peripheral blood natural killer cells in children with autism. **Brain, behavior, and immunity**, NIH Public Access, v. 23, n. 1, p. 124, jan. 2009. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636576/>.

GENOVESE, A.; BUTLER, M. G. Clinical assessment, genetics, and treatment approaches in autism spectrum disorder (asd). **International Journal of Molecular Sciences**, Multidisciplinary Digital Publishing Institute, v. 21, n. 1313, p. 4726, jan. 2020. ISSN 1422-0067.

GLATT, S. J. et al. Blood-based gene expression signatures of autistic infants and toddlers. **Journal of the American Academy of Child and Adolescent Psychiatry**, v. 51, n. 9, p. 934–44.e2, sep. 2012. ISSN 0890-8567.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GREENER, J. G. et al. A guide to machine learning for biologists. **Nature Reviews Molecular Cell Biology**, v. 23, n. 1, p. 40–55, jan. 2022. ISSN 1471-0080.

GREGG, J. P. et al. Gene expression changes in children with autism. **Genomics**, v. 91, n. 1, p. 22–29, jan. 2008. ISSN 0888-7543. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0888754307002327>.

GöK, M. A novel machine learning model to predict autism spectrum disorders risk gene. **Neural Computing and Applications**, v. 31, n. 10, p. 6711–6717, oct. 2019. ISSN 1433-3058. Available from Internet: <https://doi.org/10.1007/s00521-018-3502-5>.

HARRIS, C. R. et al. Array programming with NumPy. **Nature**, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, sep. 2020. Available from Internet: <https://doi.org/10.1038/s41586-020-2649-2>.

HERTZ-PICCIOTTO, I. et al. A prospective study of environmental exposures and early biomarkers in autism spectrum disorder: Design, protocols, and preliminary data from the marbles study. **Environmental Health Perspectives**, v. 126, n. 11, p. 117004, nov. 2018. ISSN 0091-6765. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371714/>.

HODGES, H.; FEALKO, C.; SOARES, N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. **Translational Pediatrics**, v. 9, n. Suppl 1, p. S55–S65, feb. 2020. ISSN 2224-4336.

HUS, V.; LORD, C. The autism diagnostic observation schedule, module 4: Revised algorithm and standardized severity scores. **Journal of autism and developmental disorders**, v. 44, n. 8, p. 1996–2012, aug. 2014. ISSN 0162-3257. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4104252/>.

KANNER, L. et al. Autistic disturbances of affective contact. **Nervous child**, New York, v. 2, n. 3, p. 217–250, 1943.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Available from Internet: <https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>.

KONG, S. et al. Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. **Neurogenetics**, v. 14, n. 2, p. 143–152, may 2013. ISSN 1364-6745. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3686296/>.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning in genetics and genomics. **Nature reviews. Genetics**, v. 16, n. 6, p. 321–332, jun. 2015. ISSN 1471-0056.

LIN, Y. et al. A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates. **Frontiers in Genetics**, Frontiers, v. 11, sep. 2020. ISSN 1664-8021. Available from Internet: <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.500064/full>.

LIU, W.; LI, M.; YI, L. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. **Autism Research**, v. 9, n. 8, p. 888–898, 2016. ISSN 1939-3806. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1002/aur.1615>.

LORD, C. et al. Autism spectrum disorder. **Nature Reviews Disease Primers**, v. 6, n. 1, p. 1–23, jan. 2020. ISSN 2056-676X.

LORD, C.; RUTTER, M.; COUTEUR, A. L. Autism diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. **J. Autism Dev. Disord.**, Springer Science and Business Media LLC, v. 24, n. 5, p. 659–685, oct. 1994.

LOWE, R. et al. Transcriptomics technologies. **PLOS Computational Biology**, Public Library of Science, v. 13, n. 5, p. e1005457, may 2017. ISSN 1553-7358. Available from Internet: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005457>.

MITCHELL, T. **Machine Learning**. New York, NY: McGraw-Hill Professional, 1997. (McGraw-Hill series in computer science).

MODABBERNIA, A.; VELTHORST, E.; REICHENBERG, A. Environmental risk factors for autism: an evidence-based review of systematic reviews and meta-analyses. **Molecular Autism**, v. 8, n. 1, p. 13, mar. 2017. ISSN 2040-2392.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. London, England: MIT Press, 2012. (Adaptive Computation and Machine Learning series).

MORADI, E. et al. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. **NeuroImage**, v. 144, p. 128–141, jan. 2017. ISSN 1053-8119. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1053811916305298>.

MORDAUNT, C. E. et al. A meta-analysis of two high-risk prospective cohort studies reveals autism-specific transcriptional changes to chromatin, autoimmune, and environmental response genes in umbilical cord blood. **Molecular Autism**, v. 10, n. 1, p. 36, oct. 2019. ISSN 2040-2392. Available from Internet: <https://doi.org/10.1186/s13229-019-0287-z>.

MULLEN, E. M. **Mullen Scales of Early Learning**. Circle Pines, MN: American Guidance Services, Inc, 1995.

MURPHY, K. P. **Machine learning: a probabilistic perspective**. Cambridge, MA: MIT Press, 2012. (Adaptive computation and machine learning series). ISBN 978-0-262-01802-9.

NEWSCHAFFER, C. J. et al. Infant siblings and the investigation of autism risk factors. **Journal of Neurodevelopmental Disorders**, v. 4, n. 1, p. 7, abr. 2012. ISSN 1866-1955. Available from Internet: <https://doi.org/10.1186/1866-1955-4-7>.

OMAR, K. S. et al. A machine learning approach to predict autism spectrum disorder. In: **2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)**. [s.n.], 2019. p. 1–6. Available from Internet: <https://ieeexplore.ieee.org/abstract/document/8679454>.

OPITZ, J. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. **Transactions of the Association for Computational Linguistics**, v. 12, p. 820–836, jun. 2024. ISSN 2307-387X.

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

QUACKENBUSH, J. Microarray data normalization and transformation. **Nature Genetics**, Nature Publishing Group, v. 32, n. 4, p. 496–501, dec. 2002. ISSN 1546-1718. Available from Internet: <https://www.nature.com/articles/ng1032z>.

RAINEN, L. et al. Stabilization of mrna expression in whole blood samples. **Clinical Chemistry**, v. 48, n. 11, p. 1883–1890, nov. 2002. ISSN 0009-9147.

ROKACH, L.; MAIMON, O. **Data Mining with Decision Trees: Theory and Applications**. 2. ed. WORLD SCIENTIFIC, 2014. (Series in Machine Perception and Artificial Intelligence, v. 81). ISBN 978-981-4590-07-5. Available from Internet: <https://www.worldscientific.com/worldscibooks/10.1142/9097>.

ROSSUM, G. V.; JR, F. L. D. **Python reference manual**. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, 1995.

SHARMA, S. R.; GONDA, X.; TARAZI, F. I. Autism spectrum disorder: Classification, diagnosis and therapy. **Pharmacology Therapeutics**, v. 190, p. 91–104, oct. 2018. ISSN 0163-7258.

SWINKELS, S. H. N. et al. Screening for autistic spectrum in children aged 14 to 15 months. i: The development of the early screening of autistic traits questionnaire (esat). **Journal of Autism and Developmental Disorders**, v. 36, n. 6, p. 723–732, aug. 2006. ISSN 1573-3432. Available from Internet: <https://doi.org/10.1007/s10803-006-0115-0>.

TEAM, T. P. D. **pandas-dev/pandas: Pandas**. Zenodo, 2023. Available from Internet: <https://zenodo.org/record/8092754>.

TYLEE, D. S. et al. Blood transcriptomic comparison of individuals with and without autism spectrum disorder: A combined-samples mega-analysis. **American journal of medical genetics. Part B, Neuropsychiatric genetics: the official publication of the International Society of Psychiatric Genetics**, v. 174, n. 3, p. 181–201, abr. 2017. ISSN 1552-4841. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5499528/>.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**, v. 17, p. 261–272, 2020.

WANG, K. et al. Classification of common human diseases derived from shared genetic and environmental determinants. **Nature genetics**, v. 49, n. 9, p. 1319–1325, sep. 2017. ISSN 1061-4036.

WRIGHT, K.; POULIN-DUBOIS, D. Modified checklist for autism in toddlers (m-chat): Validation and correlates in infancy. In: ____. **Comprehensive Guide to Autism**. New York, NY: Springer, 2014. p. 2813–2833. ISBN 978-1-4614-4788-7. Available from Internet: <https://doi.org/10.1007/978-1-4614-4788-7_167>.

ZABLOTSKY, B.; BLACK, L. I.; BLUMBERG, S. J. Estimated prevalence of children with diagnosed developmental disabilities in the united states, 2014-2016. **NCHS data brief**, n. 291, p. 1–8, nov. 2017. ISSN 1941-4927.

ZHANG, F. et al. Whole brain white matter connectivity analysis using machine learning: an application to autism. **NeuroImage**, v. 172, p. 826–837, may 2018. ISSN 1053-8119. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5910272/>.

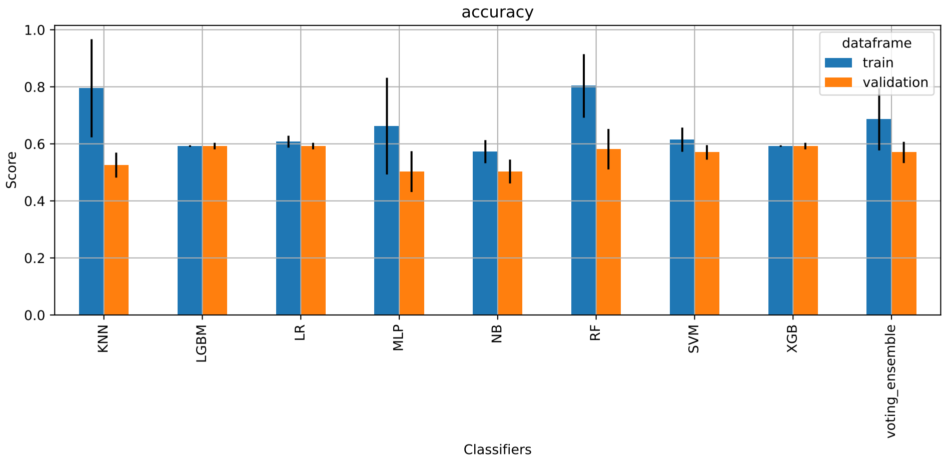
APPENDIX A — EXPANDED RESULTS

This appendix presents the graphic representation of the performance measured in the metrics other than Fbeta score, of both parts of the model, for the CV in Section A.1 and for the final performance in the Test set in Section A.2.

A.1 Cross Validation performance

The Figures for model 1 and model 2 will be given for the CV performance in each metric that was not reported in Chapter 5. Every Figure shows the mean performance across 5 folds and its standard deviation for both train and validation sets, of each individual classifier in each metric, comparing to the final classifier for each part, the voting_ensemble. Figures A.1 and A.2 present the accuracy, Figures A.3 and A.4 present the precision, Figures A.5 and A.6 present the recall, Figures A.7 and A.8 present the F1 score, and finally, Figures A.9 and A.10 present the ROC AUC.

Figure A.1: Mean (bar) and standard deviation (black line) of the performance measured in accuracy (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.

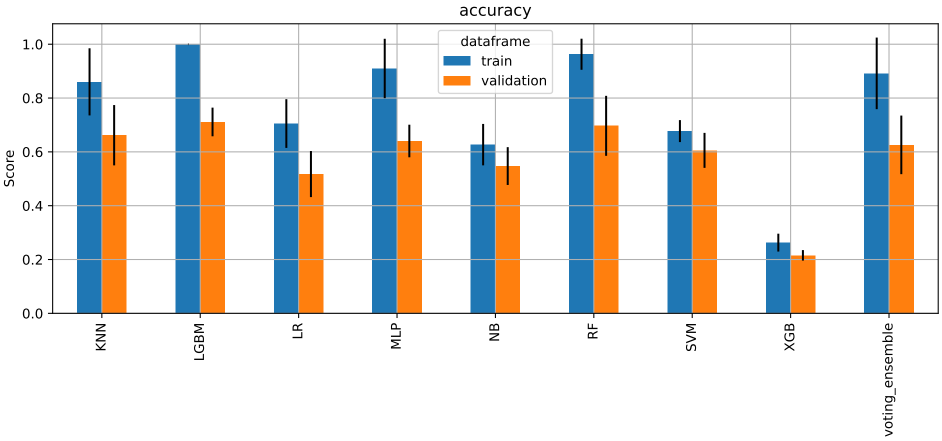


Source: The Authors

A.2 Model's performance

The Figures for model 1 and model 2 will be given for the final performance in each metric that was not reported in Chapter 5. Every Figure shows the final performance

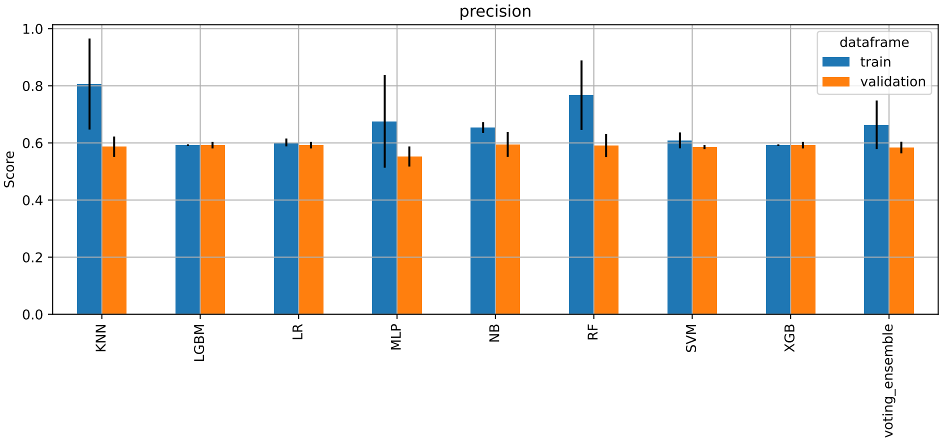
Figure A.2: Mean (bar) and standard deviation (black line) of the performance measured in accuracy (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



Source: The Authors

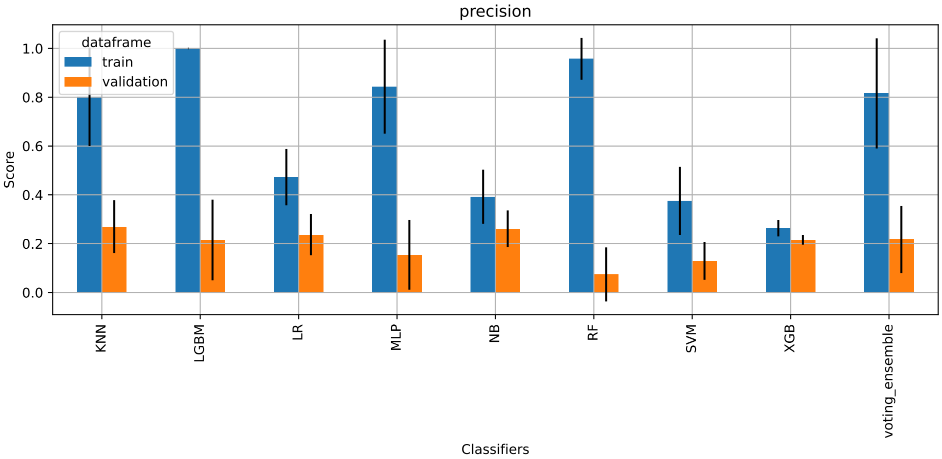
of each individual classifier in each metric for both train and validation sets, comparing to the final classifier for each part, the voting_ensemble. Figures A.11 and A.12 present the accuracy, Figures A.13 and A.14 present the precision, Figures A.15 and A.16 present the recall, Figures A.17 and A.18 present the F1 score, and finally, Figures A.19 and A.20 present the ROC AUC.

Figure A.3: Mean (bar) and standard deviation (black line) of the performance measured in precision (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



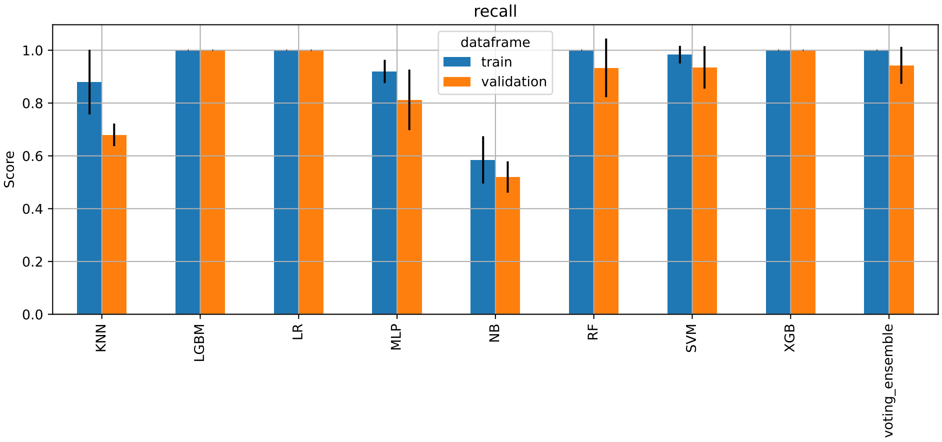
Source: The Authors

Figure A.4: Mean (bar) and standard deviation (black line) of the performance measured in precision (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



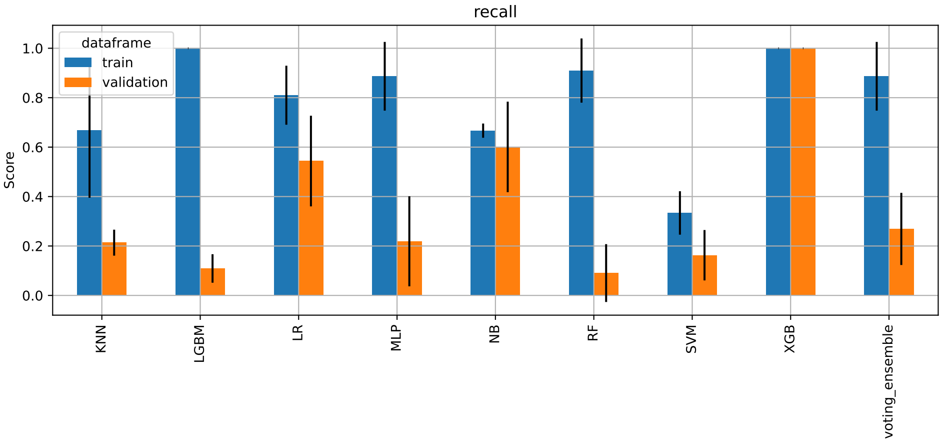
Source: The Authors

Figure A.5: Mean (bar) and standard deviation (black line) of the performance measured in recall (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



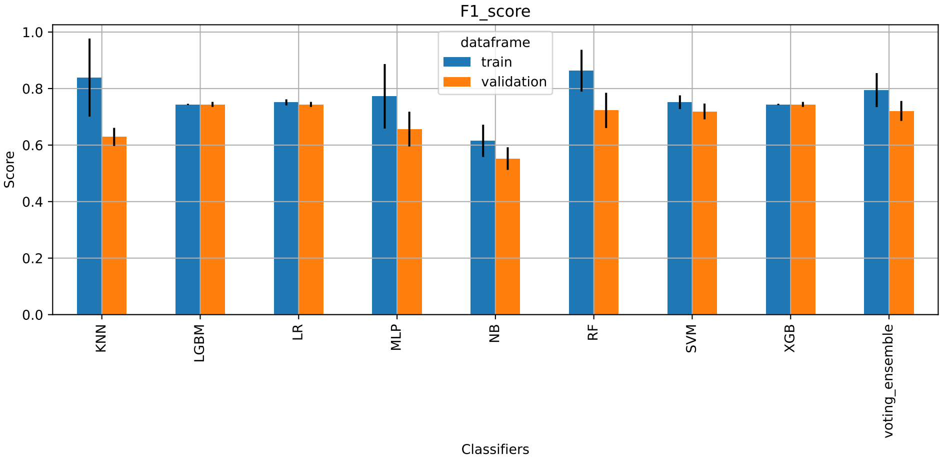
Source: The Authors

Figure A.6: Mean (bar) and standard deviation (black line) of the performance measured in recall (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



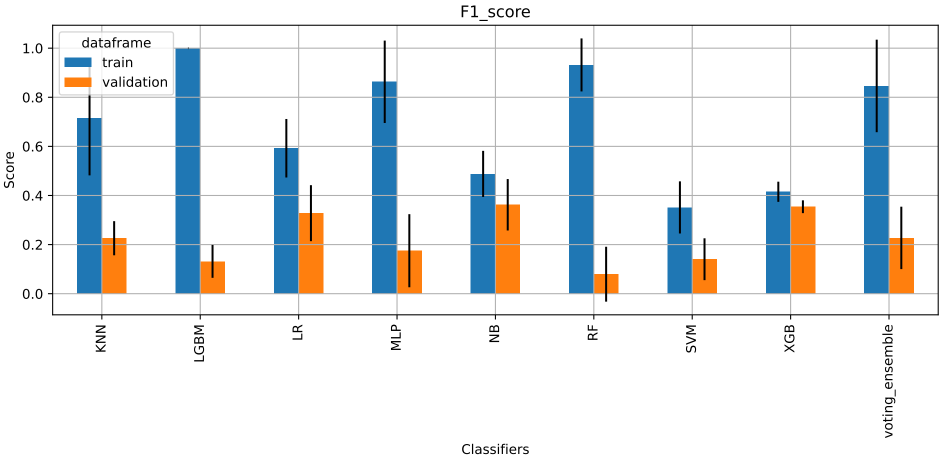
Source: The Authors

Figure A.7: Mean (bar) and standard deviation (black line) of the performance measured in F1-score (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



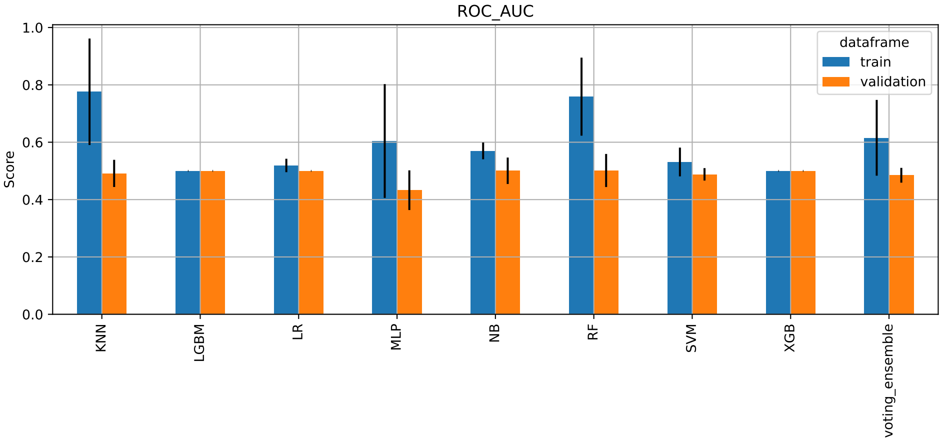
Source: The Authors

Figure A.8: Mean (bar) and standard deviation (black line) of the performance measured in F1-score (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



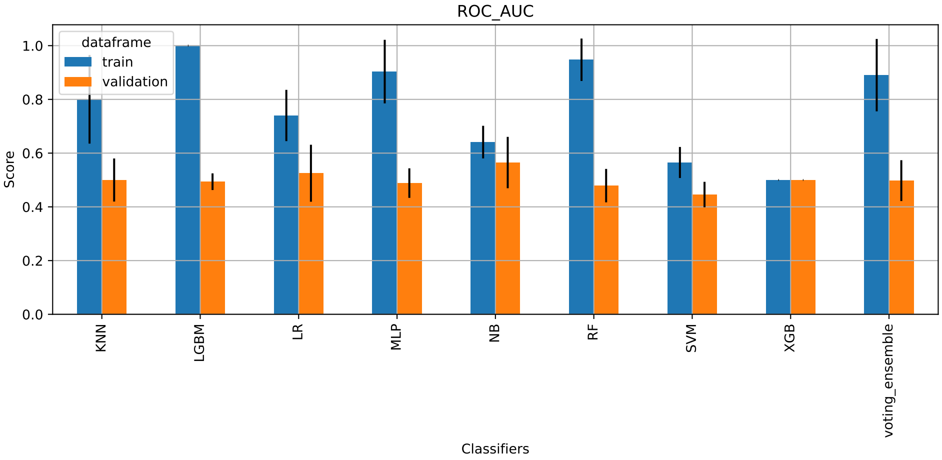
Source: The Authors

Figure A.9: Mean (bar) and standard deviation (black line) of the performance measured in ROC AUC (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



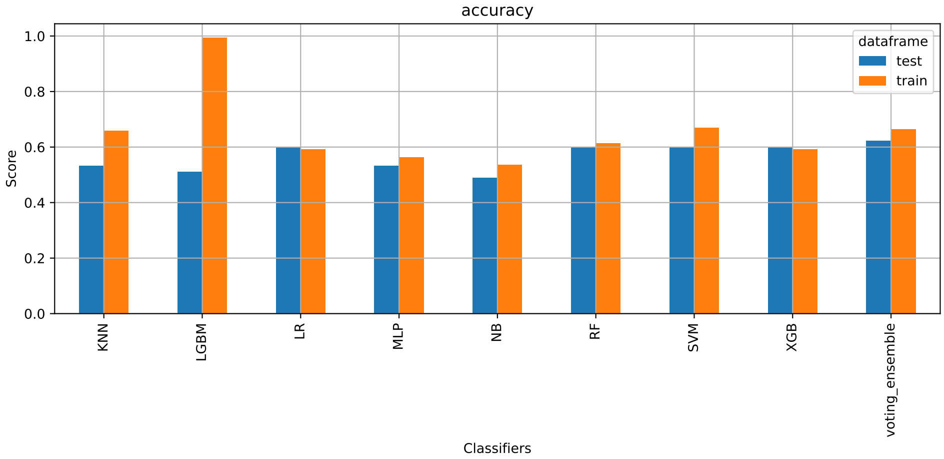
Source: The Authors

Figure A.10: Mean (bar) and standard deviation (black line) of the performance measured in ROC AUC (Y axis) in the CV for each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



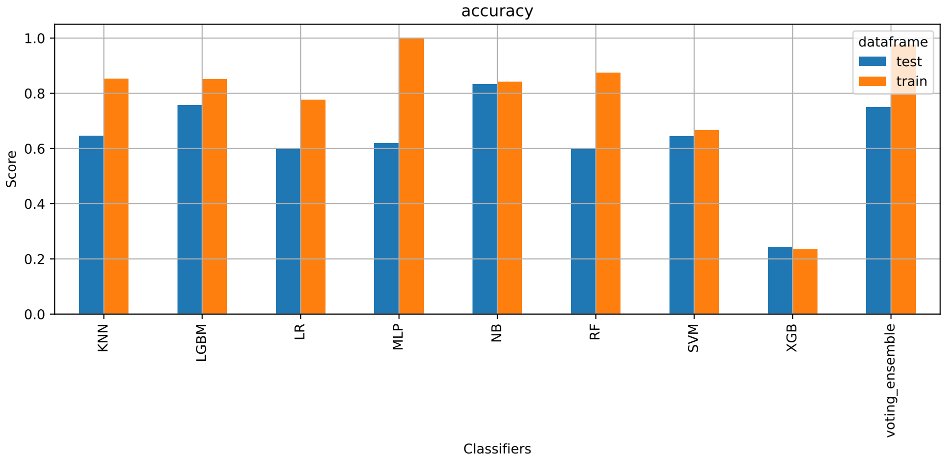
Source: The Authors

Figure A.11: Performance measured in accuracy (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



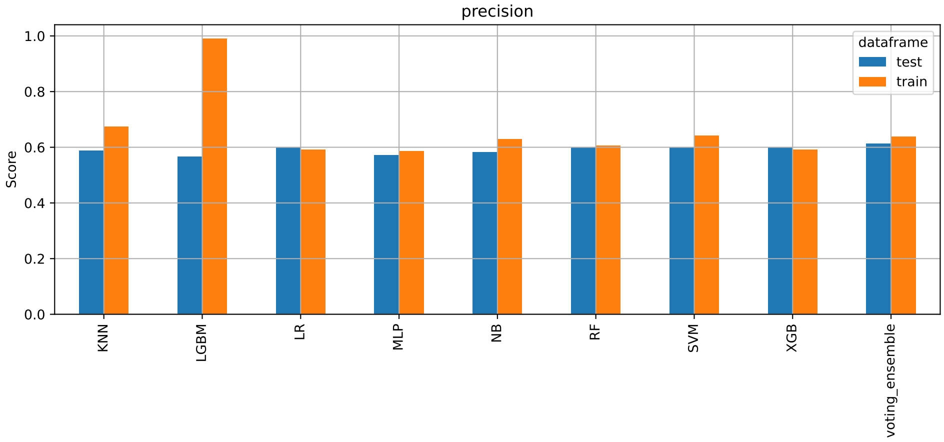
Source: The Authors

Figure A.12: Performance measured in accuracy (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



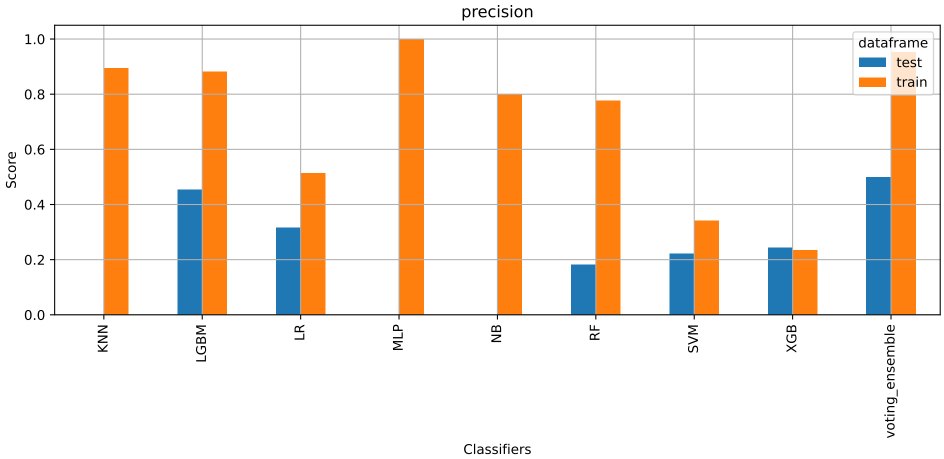
Source: The Authors

Figure A.13: Performance measured in precision (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



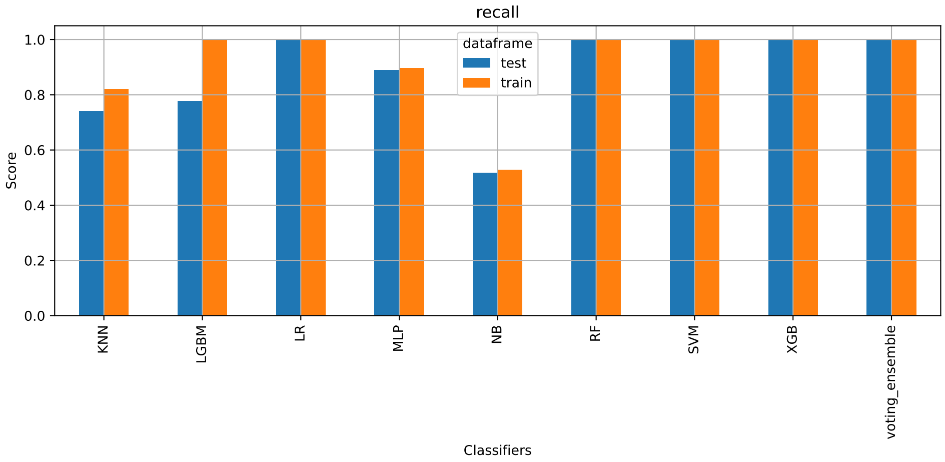
Source: The Authors

Figure A.14: Performance measured in precision (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



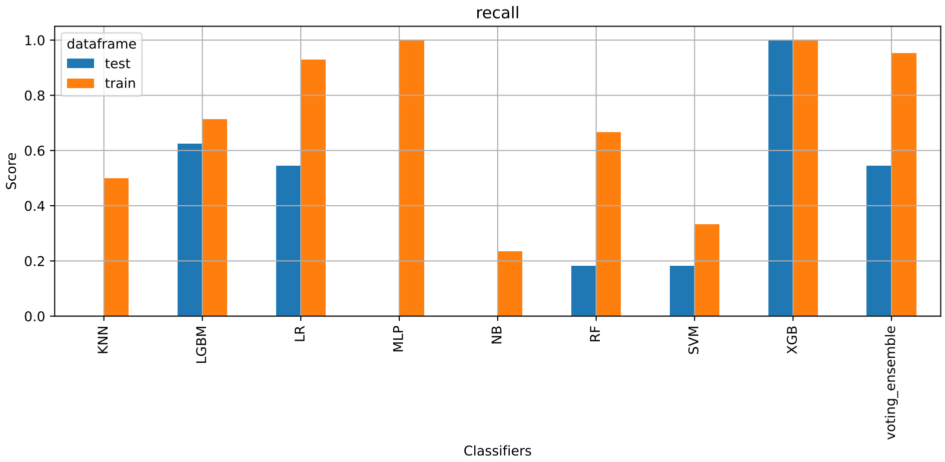
Source: The Authors

Figure A.15: Performance measured in recall (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



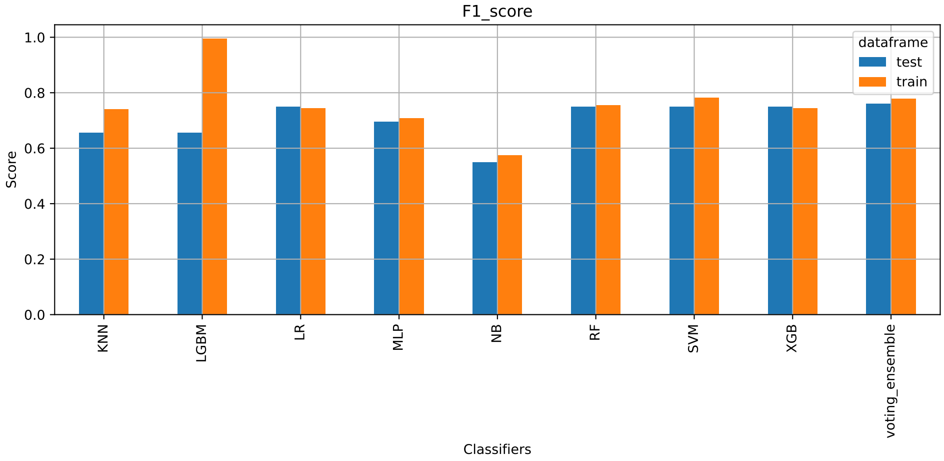
Source: The Authors

Figure A.16: Performance measured in recall (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



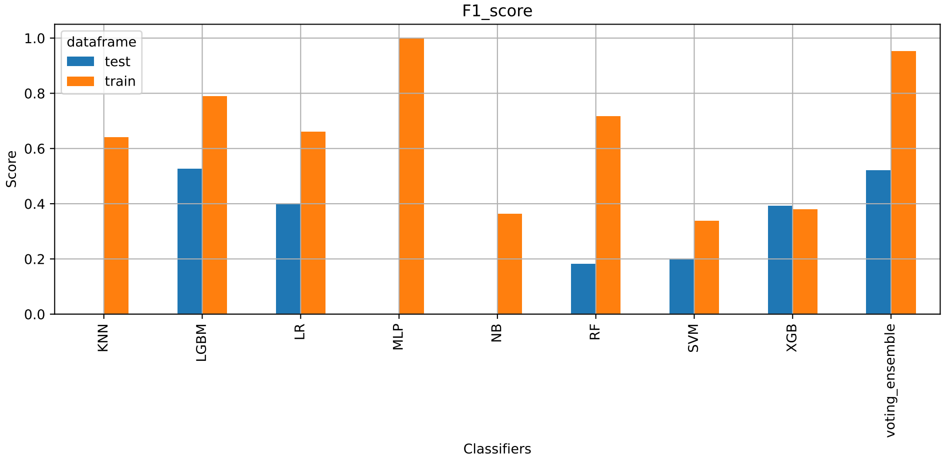
Source: The Authors

Figure A.17: Performance measured in F1-score (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



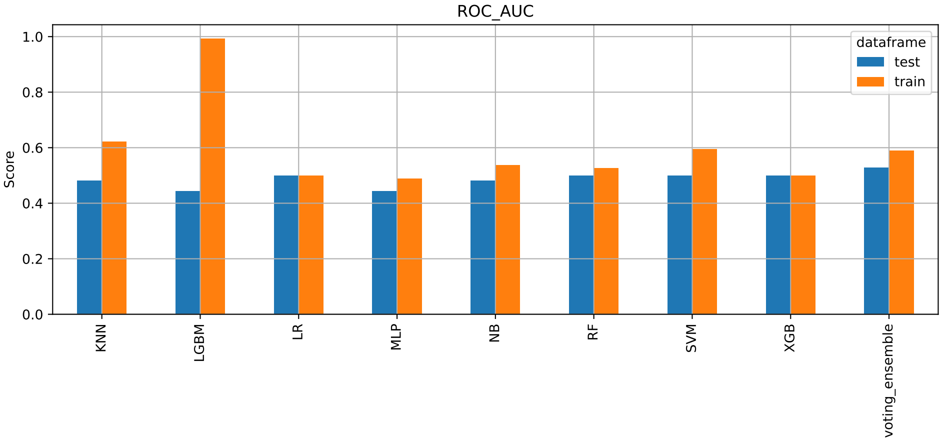
Source: The Authors

Figure A.18: Performance measured in F1-score (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



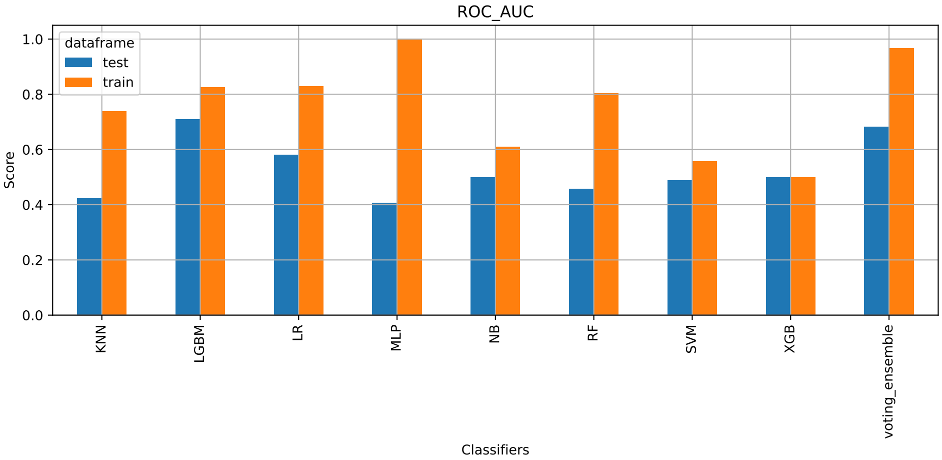
Source: The Authors

Figure A.19: Performance measured in ROC-AUC (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 1**.



Source: The Authors

Figure A.20: Performance measured in ROC-AUC (Y axis) of each individual classifier (X axis) in both Train (orange) and Test (blue) datasets, comparing to the final classifier (voting_ensemble) for **model 2**.



Source: The Authors