



UNIVERSIDADE D
COIMBRA



Dani Ferreira Silva

**MODELOS DE CLASSIFICAÇÃO PARA DIAGNÓSTICO
PRECOCE DO TRANSTORNO DO ESPETRO AUTISTA**

VOLUME 1

Dissertação no âmbito do Mestrado Integrado em Engenharia Biomédica,
Especialização em Instrumentação Biomédica orientada pelo Professor Doutor
Francisco José Santiago Fernandes Amado Caramelo e apresentada à Faculdade de
Ciências e Tecnologia.

Dezembro de 2018



UNIVERSIDADE DE
COIMBRA

FACULDADE
DE CIÊNCIAS
E TECNOLOGIA

Dani Ferreira Silva

Modelos de Classificação para Diagnóstico precoce do Transtorno do Espectro Autista

Tese submetida à
Universidade de Coimbra para o grau de
Mestre em Engenharia Biomédica

Supervisor:
Prof. Dr. Francisco Caramelo (Coordenador)

Coimbra, 2018

Este trabalho foi desenvolvido em colaboração com:

Coimbra Institute for Clinical and Biomedical Research



Centro Hospitalar e Universitário de Coimbra (CHUC)



Esta cópia da tese é fornecida na condição de que quem a consulta reconhece que os direitos de autor são pertença do autor da tese e que nenhuma citação ou informação obtida a partir dela pode ser publicada sem a referência apropriada.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognize that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without proper acknowledgement.



Agradecimentos

Agradeço ao Professor Doutor Francisco Caramelo, pelo incentivo constante e inestimável e pela confiança demonstrada. Graças às suas reconhecidas competências científicas e académicas, colocou-me perante desafios vários e estimulou o meu interesse pelas temáticas abordadas, oferecendo-me uma oportunidade ímpar de enriquecimento académico e pessoal.

Exprimo a minha gratidão à Professora Doutora Guiomar Oliveira, do Hospital Pediátrico de Coimbra, que gentilmente compartilhou os dados utilizados neste projeto e, para além disso, toda a disponibilidade para esclarecer questões essenciais para a compreensão e evolução do trabalho.

A todos os meus amigos com os quais compartilhei tantas gargalhadas, projetos de última hora, semanas de exames exaustivas, tradições académicas únicas e todos os bons e maus momentos que se sucederam ao longo dos últimos anos.

Finalmente à minha mãe, à minha irmã, e a toda a minha família por todo o apoio ao longo destes anos, durante os altos e baixos, e por sempre me encorajarem a perseguir os meus sonhos e aventuras. Quero agradecer por todos os grandes momentos que passámos juntos, por me ajudarem a crescer de todas as formas e por tornarem tudo isto possível.

Acknowledgments

Resumo

O presente estudo tem como objetivo obter um modelo preditivo capaz de distinguir precocemente entre crianças com Transtorno do Espectro Autista (ASD) e crianças sem esse transtorno, recorrendo a informações familiares e a resultados de dois testes específicos de rastreio, a Lista de Verificação Modificada para Autismo em Crianças (M-CHAT) e o Questionário de Comunicação Social (SCQ). Para tal, procedeu-se à execução de metodologias de *data science* entre as quais a aplicação de classificadores e técnicas de imputação de *missings* ou a seleção de *features*, por forma a obter os melhores resultados de classificação. A seleção de *features* permitiu identificar os melhores discriminadores do Transtorno do Espectro Autista entre os itens do M-CHAT, do SCQ e de outros dados, examinando-se a fiabilidade, a validade e a precisão diagnóstica do conjunto de *features* selecionado.

Os participantes deste estudo foram encaminhados para a Unidade de Neurodesenvolvimento e Autismo (UNDA) do Hospital Pediátrico do Centro Hospitalar e Universitário de Coimbra (CHUC) devido a suspeitas de distúrbios do desenvolvimento neurológico. Entre 2009 e 2015, todas as crianças foram extensivamente e longitudinalmente avaliadas pela mesma equipa clínica especializada e multidisciplinar para o diagnóstico do ASD. Os diagnósticos finais (ASD, $n = 135$, versus Não-ASD, $n = 66$) foram rigorosamente realizados, não apenas com avaliação clínica mas também com outros testes objetivos bem estabelecidos para a idade das crianças que entretanto cresceram. Os resultados do M-CHAT, SCQ e outros dados foram analisados e as suas propriedades psicométricas comparadas.

Para a implementação do classificador foram testados três algoritmos diferentes: *Support Vector Machine*, *Random Forest* e *Logistic Regression*. O classificador *Random Forest* obteve o melhor desempenho ao distinguir a classe ASD da classe não-ASD com valores médios de 95,7% de exatidão, 97,4% de sensibilidade, uma especificidade de 94,2% e uma área abaixo da curva ROC (AUC) de 95,8%.

Pode-se considerar que os resultados obtidos para o modelo preditivo fornecem um valioso contributo para o diagnóstico diferencial precoce e para a tomada de decisão

clínica, permitindo uma rápida e correta atualização das famílias e consequente planejamento do tratamento, particularmente ao avaliar as crianças em situação de risco. Este fator é de uma importância crucial, pois os encaminhamentos antecipados aumentam as intervenções precoces, melhorando o prognóstico ao longo da vida e atenuando, de forma substancial, o impacto que o Transtorno do Espectro Autista tem no cotidiano das crianças e das respectivas famílias.

Palavras-chave: transtorno do espectro autista, M-CHAT, SCQ, classificação.

Abstract

The main goal of this study is to obtain a predictive model able to distinguish between children with ASD and children without ASD, by using parental information and the results of two specific tests Modified Checklist for Autism in Toddlers (M-CHAT) and Social Communication Questionnaire (SCQ). We applied data science methodologies such as missing data imputation and feature selection to obtain better results. Feature selection permitted to identify the best discriminants for ASD among all the features, including M-CHAT and SCQ items, when the precision, accuracy and validity were examined.

The participants were forwarded to the Unit for Neurodevelopment and Autism (UNDA) of the Pediatric Hospital of Coimbra (CHUC), because they were suspected of having neurodevelopmental disorders. Between 2009 and 2015, all the children were thoroughly evaluated by the same multidisciplinary and specialized clinical team for ASD. The final diagnosis (ASD, $n = 135$ versus Non-ASD, $n = 66$) was achieved resorting not only to clinical evaluation but also using well established tests. The M-CHAT and SCQ results were analysed and its psychometric properties compared.

For the implementation of the classifier three different algorithms were tested: Support Vector Machine, Random Forest and Logistic Regression. Random Forest was the best classifier achieving 95.7 % for accuracy, 97.4 % for sensitivity, a specificity of 94.2 % and an *Area Under the Receiver Operating Characteristic (ROC) Curve* (Área abaixo da curva ROC) (AUC) of 95.8%.

The results obtained for the predictive model provide a valuable contribution to the differential diagnosis and clinical decision, allowing a fast and correct update of information to the families and consequent treatment planning, particularly for children in risky situations. This factor is crucial because early referrals increase early interventions, improving the prognosis throughout life, and substantially reducing the impact that Autistic Spectrum Disorder has on the daily lives of children and their families.

Keywords: autism spectrum disorder, M-CHAT, SCQ, classification.

Acrónimos

- AAP** *American Academy of Pediatrics*. 4, 5
- ADI-R** *Autism Diagnostic Interview - Revised*. 6, 20, 21, 50
- ADOS-G** *Autism Diagnostic Observation Schedule - Generic*. 6
- ANN** *Artificial Neural Network*. 36
- ASD** *Autism Spectrum Disorder*. 1–11, 13–17, 19–22, 37, 40, 41, 43, 44, 48–52, 58, 60–62, 65–70
- AUC** *Area Under the ROC Curve* (Área abaixo da curva ROC). 33, 34, 36, 43, 45–47, 52, 53, 56, 57, 59, 61, 62, 64, 65
- CARS** *Childhood Autism Rating Scale*. 6
- CHAT** *Checklist for Autism in Toddlers*. 15
- CHUC** Centro Hospitalar e Universitário de Coimbra. 10
- DFA** *Discriminant Function Analysis*. 19
- DSM** *Diagnostic and Statistical Manual of Mental Disorders*. 5, 6, 21, 50
- GARS-2** *Gilliam Autism Rating Scale - Second Edition*. 6
- IC** Intervalo de Confiança. 8, 46
- LDA** *Linear Discriminant Analysis*. 31, 33, 35
- LR** *Logistic Regression*. 33, 34, 36, 44, 45, 53, 59, 62
- M-CHAT** *Modified Checklist for Autism in Toddlers*. 14–19, 48–50, 52, 63, 65–68, 70, 75, 78

PCA *Principal component analysis.* 29

PDD-NOS *Pervasive Developmental Disorder – Not Otherwise Specified.* 5, 67

RF *Random Forest.* 31, 33–36, 44, 45, 48, 53, 58, 60–62

RFE *Recursive Feature Elimination.* 28, 29

ROC *Receiver Operating Characteristic.* 20, 33, 34, 45, 46

SCQ *Social Communication Questionnaire.* 14, 15, 20–23, 48–50, 52, 53, 63, 65, 66, 70, 79

SVM *Support Vector Machine.* 28, 29, 31, 33–36, 42, 44, 53, 58, 62

SVM-RFE *Support Vector Machine - Recursive Feature Elimination.* 28–30, 40–43, 48, 52, 53, 57–59, 61, 67, 68

UNDA *Unidade de Neurodesenvolvimento e Autismo.* 10

VPN *Valor Preditivo Negativo.* 17, 20

VPP *Valor Preditivo Positivo.* 14, 16, 17, 20

Lista de Figuras

3.1	Resumo do procedimento experimental.	37
3.2	Representação esquemática detalhada do procedimento experimental.	38
3.3	Procedimento executado para a implementação dos Testes Estatísticos à base de dados.	42
4.1	Número de <i>missings</i> por <i>feature</i> presentes na base de dados original. A linha vertical traçada identifica o valor correspondente a 30% de <i>missings</i>	54
4.2	<i>Heatmap</i> da correlação entre <i>features</i> . De referir que os eixos não apresentam os nomes de todas as <i>features</i>	55
4.3	Comparação entre os classificadores gerados a partir de dados não imputados e de dados imputados através do modelo apresentado. Nota: os pontos verdes assinalados correspondem ao valor da média para cada conjunto de dados.	56
4.4	Comparação entre os modelos preditivos gerados. Nota: os pontos verdes assinalados correspondem ao valor da média para cada conjunto de dados.	60
4.5	Comparação entre os classificadores gerados a partir de dados: sem seleção de <i>features</i> , com seleção de <i>features</i> efetuada pelos Testes Estatísticos, com seleção de <i>features</i> efetuada pelo <i>Support Vector Machine - Recursive Feature Elimination</i> (SVM-RFE) e com seleção de <i>features</i> efetuada com a conjugação dos Testes Estatísticos e SVM-RFE. Nota: os pontos verdes assinalados correspondem ao valor da média para cada conjunto de dados.	61
4.6	<i>Heatmap</i> da distribuição das respostas aos testes <i>Modified Checklist for Autism in Toddlers</i> (M-CHAT) e <i>Social Communication Questionnaire</i> (SCQ) em função do diagnóstico real de cada indivíduo (representado na última coluna).	62
4.7	Porcentagem relativa correspondente ao número de vezes que o modelo preditivo falha o diagnóstico de cada indivíduo (com 500 repetições). 63	

4.8	Relação entre o número de indivíduos com mais de 10% de falhas no diagnóstico, o resultado de cada item dos testes M-CHAT e SCQ e o diagnóstico real.	65
-----	---	----

Lista de Tabelas

2.1	Resultados de avaliação do M-CHAT.	18
2.2	Resultados de avaliação do SCQ.	23
4.1	Dados demográficos e ocorrência dos principais marcos do neurodesenvolvimento dos participantes do estudo, para toda a amostra e para cada grupo de diagnóstico (<i>Autism Spectrum Disorder (ASD) versus não-ASD</i>).	51
4.2	Literacia dos pais para cada grupo de diagnóstico (<i>ASD versus não-ASD</i>).	51
4.3	Valores médios de AUC e respetivos intervalos de confiança para os classificadores <i>Support Vector Machine</i> , <i>Random Forest</i> e <i>Logistic Regression</i> , tendo em conta os diversos métodos de seleção de <i>features</i> e o método de imputação de <i>missings</i> . Os dados são apresentados da seguinte forma: média AUC (intervalo de confiança).	53
4.4	Valor-p de cada <i>feature</i> e <i>ranking</i> do SVM-RFE com e sem implementação prévia dos Testes Estatísticos. * Controlo do esfíncter da bexiga diurno/noturno.	58
4.5	Valores das métricas de avaliação dos classificadores implementados para o conjunto de dados que sofreu imputação e seleção de <i>features</i> recorrendo ao SVM-RFE.	61
4.6	<i>Ranking</i> dos itens mais relevantes do teste M-CHAT para a classificação do ASD, obtidos no estudo efetuado e na literatura consultada.	67

Índice

Lista de Figuras	xiii
Lista de Tabelas	xv
1 Introdução	1
1.1 Contextualização - Transtorno do Espectro do Autismo	1
1.1.1 Sintomas	2
1.1.2 Causas e riscos	3
1.1.3 Diagnóstico	3
1.1.4 Ferramentas de diagnóstico	5
1.1.5 Idade para rastreio do ASD	6
1.1.6 Intervenção antecipada	7
1.1.7 Epidemiologia	8
1.1.8 Tratamento e terapias	8
1.2 Motivação e Objetivo	9
1.3 Estrutura	10
2 Revisão Bibliográfica	13
2.1 Testes de rastreio do ASD	13
2.1.1 Comparação das propriedades psicométricas dos testes	13
2.1.2 Testes disponíveis	14
2.2 M-CHAT	15
2.2.1 Validade de Diagnóstico	15
2.2.2 Itens críticos do M-CHAT	19
2.3 SCQ	20
2.3.1 Validade de diagnóstico	20
2.4 Imputação de dados	24
2.4.1 Imputação pela Moda	25
2.4.2 Imputação pela Regressão Linear	25
2.5 Seleção de <i>features</i>	25

2.5.1	SVM-RFE	28
2.5.2	Testes Estatísticos (Teste-t)	30
2.6	Balanceamento	31
2.7	Classificação	33
3	Metodologia	37
3.1	Imputação de Dados	37
3.2	Seleção de <i>features</i>	40
3.3	Balanceamento	43
3.4	Classificação	44
3.5	Métricas de avaliação	45
3.6	Análise de amostras	47
3.7	<i>Ranking</i> dos itens do M-CHAT	48
4	Resultados e Discussão	49
4.1	Base de dados	49
4.2	Avaliação	52
4.3	Imputação de <i>missings</i>	52
4.3.1	Avaliação	55
4.4	Seleção de <i>features</i>	57
4.4.1	Avaliação	59
4.5	Classificação	59
4.6	Análise de indivíduos	63
4.7	<i>Ranking</i> dos itens do M-CHAT	66
5	Conclusão	69
	Bibliografia	71
	Anexos	85
A	Modified Checklist for Autism in Toddlers (M-CHAT) — Versão Portuguesa	87
B	Social Communication Questionnaire (SCQ) — Versão Portuguesa	91

Introdução

1.1 Contextualização - Transtorno do Espectro do Autismo

O Transtorno do Espectro do Autismo (ASD) corresponde a um conjunto de perturbações complexas do neurodesenvolvimento caracterizado por deficiente interação e comunicação social, comportamentos estereotipados e repetitivos e desenvolvimento intelectual irregular, frequentemente com retardo mental [1, 2]. Embora o autismo possa ser diagnosticado em qualquer idade, é considerado um "distúrbio do desenvolvimento", porque os sintomas, geralmente, aparecem nos dois primeiros anos de vida [3].

O termo "espectro" refere-se à ampla gama de sintomas, destreza e níveis de incapacidade funcional que podem ocorrer nas pessoas com ASD. Algumas crianças e adultos com ASD são totalmente capazes de realizar todas as atividades da vida diária, enquanto outros requerem apoio substancial para execução de atividades básicas [1]. Embora, atualmente, não exista um tratamento padrão e definitivo para o ASD, os tratamentos e serviços podem melhorar os sintomas e a aptidão das pessoas afetadas. Os tratamentos incluem terapias de comportamento e comunicação, fisioterapia e medicamentos para controlar os sintomas [3].

O ASD ocorre em todos os grupos étnicos e em todos os níveis socioeconómicos. No entanto, o sexo masculino é significativamente mais propenso a desenvolver ASD do que o sexo feminino [4].

1.1.1 Sintomas

Geralmente, os sintomas de ASD começam a manifestar-se antes dos 3 anos de idade e permanecem por toda a vida de uma pessoa podendo, no entanto, melhorar com o tempo. Algumas crianças com ASD demonstram sintomas da doença logo nos primeiros meses de vida, mas noutros casos os sintomas podem não aparecer até aos 24 meses ou mais tarde. Usualmente as crianças com ASD têm um desenvolvimento normal até aos 18 a 24 meses de idade e, a partir dessa idade, deixam de adquirir novas capacidades ou perdem as que já tinham [5]. Estas crianças podem tornar-se excessivamente focadas em certos objetos, raramente fazem contacto visual e não iniciam os balbucios normais. Noutros casos, as crianças podem desenvolver-se normalmente até ao segundo ou terceiro ano de vida, mas depois começam a alienar-se e a tornar-se indiferentes ao envolvimento social.

A gravidade do ASD pode variar muito e é baseada no grau em que a comunicação social, a insistência na repetição das atividades e os padrões repetitivos de comportamento afetam o funcionamento diário do indivíduo [1].

Dificuldades sociais e dificuldades de comunicação

A natureza mútua de dar e receber da comunicação e interação normais é muitas das vezes particularmente desafiadora. As crianças com ASD podem não responder aos seus nomes, evitar contacto visual com outras pessoas, iniciar interação somente para atingir objetivos específicos, não entender como brincar ou como se envolver com outras crianças e podendo também preferir ficar sozinhas. Para além disso, poderão ter dificuldades em entender os sentimentos de outras pessoas ou falar sobre os seus próprios sentimentos.

No que concerne às aptidões verbais dos portadores de ASD, estas podem ser muito variáveis, desde a falta de fala até a fala fluente embora desajeitada e inadequada [5].

Comportamentos repetitivos e característicos

Muitas crianças com ASD praticam movimentos repetitivos ou comportamentos incomuns, como bater os braços, balançar de um lado para outro ou girar, podem tornar-se obsessivamente interessadas num tema específico, ficar com raiva ou ter explosões emocionais, especialmente quando inseridas num ambiente novo ou excessivamente estimulante [5].

1.1.2 Causas e riscos

Embora os cientistas não saibam as causas exatas do ASD, pesquisas sugerem que os genes podem agir em conjunto com influências do ambiente de forma a afetar o desenvolvimento. Ainda que existam diversos estudos a decorrer no sentido de perceber o que provoca o desenvolvimento de ASD, foram já identificados alguns fatores de risco:

- Pais que têm um filho com ASD têm 2 a 18% de probabilidade de ter um segundo filho que também seja afetado [6, 7];
- Estudos mostraram que entre gémeos verdadeiros, se uma criança tem ASD, a outra será afetada em 36 a 95% das vezes. Em gémeos não-verdadeiros, se uma criança tem ASD, a outra é afetada até cerca de 31% das vezes [8, 9, 10, 11];
- Crianças com pais mais velhos correm maior risco de ter ASD [12, 13];
- O ASD tende a ocorrer mais frequentemente em pessoas que possuem certas particularidades genéticas ou cromossômicas. Cerca de 10% das crianças com autismo são também identificadas como tendo síndrome de Down, síndrome do X frágil, esclerose tuberosa ou outros distúrbios genéticos e cromossômicos [14, 15];
- Uma pequena percentagem de crianças que nascem prematuramente ou com baixo peso está associado a um maior risco de ter ASD [16].

1.1.3 Diagnóstico

À primeira vista, algumas pessoas com autismo podem parecer ter uma deficiência intelectual, problemas de processamento sensorial ou problemas de audição ou visão. Para complicar ainda mais, estas disfunções podem co-ocorrer com o autismo e, nestas situações, é importante distinguir o ASD de outras doenças, já que um diagnóstico preciso e precoce pode fornecer a base para um programa educacional e de tratamento apropriado [17].

Dois métodos frequentemente utilizados na triagem do desenvolvimento de ASD são os questionários aos pais e as observações feitas por clínicos especializados, sendo que cada um tem as suas vantagens e desvantagens. A informação fornecida pelos pais é incentivada por muitos, devido à sua maior proximidade com os filhos. Este

método permite conhecer uma variedade de estados de espírito da criança, seja no comportamento em ambientes diferentes, seja na reação aquando da interação com vários indivíduos. Por conseguinte, os questionários preenchidos pelos progenitores não se baseiam numa amostra breve e atípica de comportamentos, permitindo a observação de situações que poderão não ser observáveis na maioria dos consultórios de atenção primária.

Por outro lado, as vantagens da observação profissional incluem o conhecimento de toda a gama do que constitui o desenvolvimento normal de uma criança. Os médicos são mais objetivos do que os pais e podem ter menor probabilidade de sobrestimar ou subestimar as aptidões ou os comportamentos problemáticos de uma determinada criança. O significado de um determinado comportamento é, na grande maioria das vezes, mais claro para um clínico treinado quando comparado com um pai. Finalmente, em muitos casos, a configuração usada para realizar as observações pode ser padronizada.

Neste seguimento, o diagnóstico em crianças é geralmente realizado por um processo composto por duas fases:

Fase 1: Rastreio geral de desenvolvimento durante as consultas periódicas

A Academia Americana de Pediatria (*American Academy of Pediatrics* (AAP)) recomenda que todas as crianças sejam examinadas no que concerne a atrasos no desenvolvimento nas consultas pediátricas dos 9, 18, 24 ou 30 meses e, especificamente para o autismo, nas consultas dos 18 e 24 meses. Uma triagem adicional poderá ser essencial se a criança for identificada com alto risco de ter ASD (como as crianças que correspondam às principais causas referidas em 1.1.2) ou problemas de desenvolvimento [18].

Existem muitas ferramentas para avaliar o ASD em crianças, mas nenhuma ferramenta isolada deve ser usada como base para o diagnóstico. As ferramentas de diagnóstico geralmente dependem de duas principais fontes de informação: as descrições dos pais ou cuidadores sobre o desenvolvimento dos seus filhos e a observação de um profissional sobre o comportamento da criança.

Nesta fase, caso sejam identificados problemas de desenvolvimento, as crianças são encaminhadas para uma segunda etapa de avaliação [18].

Fase 2: Avaliação Adicional

Esta segunda avaliação é realizada por uma equipa de médicos e outros profissionais de saúde com experiência no diagnóstico de ASD, podendo incluir:

- Um pediatra de desenvolvimento — um médico especializado em desenvolvimento infantil;
- Um psicólogo infantil e/ou psiquiatra infantil — um médico especializado no desenvolvimento e comportamento do cérebro;
- Um neuropsicólogo — um médico que se concentra na avaliação, diagnóstico e tratamento de distúrbios neurológicos, médicos e do desenvolvimento neurológico;
- Um fonoaudiólogo — um profissional de saúde que tem treino específico em dificuldades de comunicação.

De acordo com o Manual Diagnóstico e Estatístico de Perturbações Mentais (*Diagnostic and Statistical Manual of Mental Disorders* (DSM)), um guia criado pela AAP, a análise médica de deteção do ASD foca-se nos seguintes aspetos:

- Nível cognitivo ou destreza de pensamento;
- Destreza da linguagem;
- Comportamentos apropriados para a idade, necessários para executar as atividades diárias de forma independente.

A avaliação efetuada resulta num diagnóstico formal e recomendações para o tratamento [18].

A última versão do DSM-V, de 2013 [2], alterou a forma como o autismo é classificado e diagnosticado. De acordo com esta última atualização, as pessoas com ASD podem ser diagnosticadas com uma das seguintes doenças: transtorno autista, síndrome de Asperger ou transtorno global do desenvolvimento sem outra especificação (*Pervasive Developmental Disorder – Not Otherwise Specified* (PDD-NOS)).

1.1.4 Ferramentas de diagnóstico

Existem diversos testes de diagnóstico de ASD que podem ser executados em diversas idades e tendo em conta diversos fatores do desenvolvimento [19]:

Entrevista de Diagnóstico do Autismo - Edição Revista [20] — Versão portuguesa do *Autism Diagnostic Interview - Revised* (ADI-R), um instrumento de diagnóstico clínico para avaliar o autismo em crianças e adultos. Este instrumento foca-se no comportamento em três áreas principais: interação social recíproca, comunicação e linguagem, e interesses e comportamentos restritos e repetitivos. O ADI-R é apropriado para crianças e adultos com idades mentais de aproximadamente 18 meses ou mais.

Escala de Observação para o Diagnóstico de Autismo - Genérica [21] — Versão portuguesa do *Autism Diagnostic Observation Schedule - Generic* (ADOS-G), uma avaliação semi-estruturada e padronizada da interação social, comunicação, brincadeiras e uso imaginativo de materiais para indivíduos suspeitos de ter ASD. A programação observacional consiste em quatro módulos de 30 minutos, cada um projetado para ser administrado a diferentes indivíduos de acordo com o seu nível de linguagem expressiva.

Escala de Avaliação de Autismo Infantil [22] — Versão portuguesa do *Childhood Autism Rating Scale* (CARS), uma avaliação adequada para crianças com mais de 2 anos de idade. O CARS inclui itens de cinco sistemas proeminentes de diagnóstico do autismo. Cada item abrange uma característica, capacidade ou comportamento específico.

Escala de avaliação de autismo de Gilliam [23] — Versão portuguesa do *Gilliam Autism Rating Scale - Second Edition* (GARS-2), auxilia professores, pais e clínicos a identificar e diagnosticar o autismo em indivíduos de 3 a 22 anos. Ajuda também a estimar a gravidade dos distúrbios do indivíduo.

Para além das ferramentas acima apresentadas, também o DSM fornece critérios padronizados para auxiliar no diagnóstico de ASD [19].

1.1.5 Idade para rastreio do ASD

Uma barreira à triagem precoce, e um fator importante na seleção da idade ideal para a triagem, é a dúvida quanto à validade do diagnóstico precoce. Um vasto conjunto de literatura indica que o diagnóstico de ASD é estável ao longo do tempo, mesmo quando o diagnóstico é feito aos 2 anos de idade [24, 25, 26]. Embora haja menos conhecimento sobre a estabilidade do diagnóstico realizado aos 2 anos de idade, o interesse em detetar o autismo o mais cedo possível está a redirecionar os esforços de pesquisa em curso com vista a despistar crianças antes dos 2 anos, e até

mesmo logo no primeiro ano de vida.

Outra questão relativamente à escolha da melhor idade para o rastreio do autismo é a faixa etária do surgimento do transtorno. É vantajoso realizar a triagem o mais cedo possível, de modo a maximizar as oportunidades de intervenção. No entanto, quando a triagem é efetuada demasiado cedo torna-se difícil distinguir o ASD de outro tipo de atraso no desenvolvimento, ou mesmo de desenvolvimento normal. A triagem populacional aos 14 meses mostrou produzir um número significativo de falsos positivos [27], sugerindo que a triagem deve ocorrer após essa idade, possivelmente na consulta dos 18 meses de idade. Em adição, aproximadamente 30% das crianças com autismo apresentam um período de desenvolvimento normal seguido de estagnação ou regressão [28], podendo a triagem precoce ignorar algumas dessas crianças com início tardio. Tendo em conta as informações referidas, a faixa etária entre os 16 e os 30 meses de idade é referida como sendo a ideal para o diagnóstico confiável de ASD.

1.1.6 Intervenção antecipada

A intervenção antecipada ajuda as crianças portadoras de ASD a melhorar o seu desenvolvimento e a adquirir novas capacidades. Evidências crescentes sugerem que a intervenção precoce resulta em ganhos progressivos de desenvolvimento em domínios como a comunicação, a interação social e a capacidade cognitiva, maior independência e melhor qualidade de vida [29, 30, 31, 32].

As conexões no cérebro de um criança são mais adaptáveis nos primeiros três anos de vida. Essas conexões, também designadas por circuitos neuronais, são a base para a aprendizagem, para o comportamento e para a saúde e, com o tempo, tornam-se mais difíceis de mudar [33].

Embora seja relevante diagnosticar a doença antecipadamente, nem sempre é possível. A maioria das crianças não é diagnosticada até aos 4 anos de idade. Não obstante, estudos mostraram que pais de crianças com ASD detetam problemas de desenvolvimento ainda antes do primeiro aniversário do filho. Preocupações sobre a visão e a audição foram as mais frequentemente relatadas no primeiro ano, sendo as diferenças nas aptidões sociais, comunicativas e motoras finas evidentes a partir dos 6 meses de idade [34].

1.1.7 Epidemiologia

Nos últimos 30 anos, o número de casos de autismo registrados aumentou consideravelmente em todos os países onde existem registros de estudos da prevalência da doença. Este aumento pode representar uma real extensão da prevalência, ou pode dever-se a uma maior consciencialização do problema. Essa consciencialização nos últimos anos pode relacionar-se com a combinação com o estabelecimento de critérios diagnósticos mais inclusivos [35], mudanças nos critérios diagnósticos do ASD, diagnóstico antecipado da doença, variáveis demográficas e geográficas, sendo que ainda se acredita que fatores ambientais contribuam para o aumento da prevalência do autismo, embora esta assunção careça de mais investigação.

No último estudo de revisão sistemática de levantamentos epidemiológicos de ASD em todo o mundo, concluiu-se que a média de prevalência da doença é de 1 em 160 crianças [36]. Em Portugal, de acordo com o último estudo realizado, a prevalência global de ASD na população-alvo (crianças entre os 6 e os 9 anos de idade) de Portugal Continental foi de 9,2 em 10.000 crianças (Intervalo de Confiança (IC) 95% 8,1–10,0) [37].

1.1.8 Tratamento e terapias

Embora não exista cura conhecida para o ASD, são desenvolvidas terapias e intervenções comportamentais que diminuem ou melhoram significativamente sintomas específicos dos portadores da doença [1]. Os diferentes tipos de tratamento podem, geralmente, ser divididos nas seguintes categorias:

Abordagens de comportamento e comunicação

O plano ideal de tratamento coordena terapias e intervenções que atendam às necessidades específicas do indivíduo.

As intervenções comportamentais/educacionais precoces têm sido muito bem sucedidas em muitas crianças com ASD. Nessas intervenções, os terapeutas usam sessões de treino altamente estruturadas e intensivas para favorecer o desenvolvimento de características sociais e de linguagem das crianças, como a análise comportamental aplicada, que encoraja comportamentos positivos e desencoraja os negativos. Além disso, o aconselhamento familiar para pais e irmãos de crianças com ASD muitas das vezes ajuda as famílias a lidar com os desafios específicos de viver com um indivíduo

com esta patologia com ASD [1].

Medicação

Não existe medicação que possa curar o ASD ou até mesmo tratar os principais sintomas, mas existem medicamentos que podem diminuir determinados efeitos da doença (irritabilidade, agressão, comportamentos repetitivos, hiperatividade, problemas de atenção, ansiedade e depressão) [38].

Abordagens Dietéticas

Alguns tratamentos dietéticos foram desenvolvidos por terapeutas confiáveis, mas muitos desses tratamentos não têm o apoio científico necessário para uma recomendação generalizada. Um tratamento não comprovado pode ter efeitos positivos em algumas crianças, mas pode não influenciar outras.

Os tratamentos dietéticos baseiam-se na ideia de que alergias alimentares ou falta de vitaminas e minerais causam determinados sintomas do ASD. Neste sentido, as mudanças alimentares incluem a remoção de certos tipos de alimentos da dieta da criança e o uso de suplementos vitamínicos ou minerais. Alguns pais consideram que as mudanças na dieta produzem efeito no modo como a criança age [38].

1.2 Motivação e Objetivo

Embora as discussões continuem a incidir sobre o se e o porquê da incidência do Transtorno do Espectro Autista parecer estar em expansão no mundo, a conscientização certamente aumentou [39]. Vários estudos demonstram que a detecção e a intervenção precoce têm um impacto positivo nos sintomas do ASD [31, 29, 40]. De facto, pesquisas sugerem que o autismo é um distúrbio em que a triagem antecipada é essencial e, portanto, tornou-se fundamental o desenvolvimento de testes de rastreio específicos para este, que sejam apropriados e viáveis para a aplicação em crianças. Neste seguimento, surgem diversas questões às quais se torna essencial responder. Deverá ser uma triagem relativa ao neurodesenvolvimento geral ou específica para o autismo? Qual a melhor idade para a implementação de testes de rastreio? Qual o melhor método de triagem? As informações dos pais/cuidadores devem ser consideradas ou a observação direta é a melhor abordagem?

O presente estudo focou-se na avaliação dos itens específicos de dois testes de rastreio do desenvolvimento neurológico e comportamental, o M-CHAT e o SCQ, relativamente à sua relevância na identificação prospetiva de casos de ASD. Esta análise

incidiu sobre uma amostra de duzentas e uma crianças com suspeitas de perturbações do neurodesenvolvimento.

Estas crianças foram encaminhadas para a Unidade de Neurodesenvolvimento e Autismo do Hospital Pediátrico do Centro Hospitalar e Universitário de Coimbra, onde foram avaliadas longitudinalmente por uma equipa clínica especializada e multidisciplinar para o diagnóstico de ASD, considerando-se este o diagnóstico real para estes indivíduos. A abordagem efetuada no presente estudo consistiu na aplicação dos testes referidos para triagem comportamental, de forma a avaliar o poder discriminativo em crianças com ASD e sem ASD, comparando-as posteriormente com o diagnóstico clínico.

É importante referir que esta abordagem inovadora corresponde ao primeiro estudo de avaliação da capacidade de diagnóstico da versão portuguesa do M-CHAT e do SCQ numa população bem caracterizada de crianças portuguesas.

Considerando as restrições temporais em ambientes de cuidados primários, um breve teste de triagem, específica para o autismo, pode ser útil para facilitar o processo envolvido na avaliação pediátrica nas consultas de rotina. Especificamente, para os clínicos com experiência mínima em ASD, o processo estaria mais facilitado, uma vez que estariam perante um teste com um número finito de itens, cujo tempo de implementação e respetivos custos seriam mais reduzidos.

Desta forma, o objetivo deste estudo exploratório foi obter um modelo preditivo capaz de distinguir entre crianças com ASD e crianças sem ASD. Tornou-se também relevante identificar os melhores discriminadores de ASD entre os itens do M-CHAT, do SCQ e de outras informações familiares, examinando-se a fiabilidade, a validade e a precisão diagnóstica do conjunto de *features* selecionado.

1.3 Estrutura

Primeiramente, foram fornecidos alguns conhecimentos gerais sobre o ASD de forma a contextualizar e proporcionar uma abordagem mais detalhada do modelo preditivo para esta perturbação do desenvolvimento.

O próximo capítulo, Revisão Bibliográfica, descreve os diferentes testes utilizados na triagem do ASD e apresenta as pesquisas mais relevantes acerca dos mesmos. Para além disso, são ainda apresentados os vários algoritmos utilizados ao longo do estudo e a respetiva análise bibliográfica.

Seguidamente, o capítulo da Metodologia, apresenta os procedimentos realizados na imputação de dados, na seleção de *features*, no balanceamento dos dados e na classificação, bem como as métricas utilizadas para avaliar os seus impactos.

No quarto capítulo, Resultados e Discussão, serão apresentados todos os resultados dos diferentes métodos e a respetiva análise e discussão.

Por último, no capítulo da Conclusão, são sugeridos alguns trabalhos futuros, juntamente com uma discussão sobre alguns dos principais desafios do projeto.

Revisão Bibliográfica

2.1 Testes de rastreio do ASD

A triagem pediátrica é frequentemente a única avaliação que as crianças recebem até começarem a pré-escola ou mesmo o jardim de infância, e portanto, os pediatras devem ser sensíveis às preocupações do seu desenvolvimento. Por forma a aliviar as pressões de uma consulta no pediatra é indispensável a existência de um dispositivo de triagem para possíveis casos de autismo que seja objetivo, fácil de executar e breve. Além disso, os médicos nem sempre podem identificar com segurança um atraso no desenvolvimento com base no comportamento de uma criança numa sessão do consultório médico; especialmente para crianças muito pequenas, um atraso social pode ser confundido com timidez e o comportamento no consultório pode não representar o seu comportamento normal. Desta forma, tornam-se essenciais as informações fornecidas pelos pais em qualquer instrumento de triagem [41].

2.1.1 Comparação das propriedades psicométricas dos testes

Segundo a literatura relativa aos testes de triagem do ASD e, tal como Charman e Gotham [42] observaram, os três parâmetros comumente utilizados na avaliação da eficácia e utilidade das ferramentas de triagem são:

Sensibilidade — Proporção de indivíduos com ASD que têm um resultado positivo no teste. Este deve ser alto para que todos os indivíduos com ASD sejam identificados na triagem;

Especificidade — Proporção de indivíduos sem ASD que têm um resultado negativo no teste. Este deve também ser alto, caso contrário, a proporção de crianças falsamente identificadas com ASD na triagem exigiria exames diagnósticos dispendiosos e preocupações desnecessárias aos pais;

Valor Preditivo Positivo (VPP) — Proporção de indivíduos com resultado positivo no teste que apresentam o transtorno. Idealmente deverá ser perto dos 100%.

2.1.2 Testes disponíveis

Existem diversas ferramentas disponíveis para avaliar os sintomas autistas em crianças e cada uma tem as suas vantagens. Não obstante, grande parte desses instrumentos possui um ou mais dos seguintes problemas como dispositivo de triagem: a sua realização é muito longa e pesada; são projetados para crianças em idade escolar, podendo não ser válidos para crianças mais novas; alguns não têm em conta as informações dos pais e apenas dependem de uma amostra de comportamento observada no consultório; alguns confiam em anormalidades no comportamento raramente notórias antes dos 3 anos de idade; muitas das medidas disponíveis não foram padronizadas; têm de ser realizados por um especialista em deficiências de desenvolvimento ou autismo. Mais importante ainda, é o facto de estes testes serem usados com crianças já identificadas como tendo características autistas ou outras preocupações de desenvolvimento, ao invés de serem realizados de forma generalizada em todas as crianças [41].

Portanto, é necessário um teste breve e objetivo, que seja fácil de realizar e pontuar e que alerte os médicos para a necessidade de uma avaliação adicional em crianças com os primeiros sinais de autismo.

Os testes de rastreio do ASD aumentam a identificação precoce e o encaminhamento para uma avaliação diagnóstica adequada. Tanto a Lista Modificada de Verificação do Autismo em Crianças - M-CHAT (Anexo A) [41] como o Questionário de Comunicação Social - SCQ (Anexo B) [43] são questionários parentais escritos (sim/não) de identificação de sintomas de interação social recíproca, linguagem e comunicação, e padrões de comportamento repetitivos e estereotipados [44]. Os testes referidos foram elaborados para identificar crianças mais novas com maior probabilidade de ter ASD, de forma a encaminhá-las adequadamente para diagnóstico e intervenção precoce. O M-CHAT e o SCQ são úteis na compilação de informações estruturadas sobre sinais e sintomas, mas não são diagnósticos nem devem ser usados para descartar o diagnóstico de ASD.

Tanto o M-CHAT como o SCQ estão entre os instrumentos de triagem mais utilizados no campo do ASD. No entanto, é importante caracterizar a sua eficácia no

cenário clínico. O tempo de espera por avaliações especializadas de ASD pode ser considerável, portanto, reduzir esse número recorrendo a um processo de triagem preciso seria benéfico para crianças, famílias e serviços locais. Além disso, reduziria também a angústia de encaminhamentos desnecessários e a quantidade de avaliações diagnósticas tardias.

2.2 M-CHAT

O M-CHAT foi desenvolvido a partir da Lista de Verificação do Autismo em Crianças (*Checklist for Autism in Toddlers (CHAT)*) [45], que corresponde à primeira tentativa de desenvolver um instrumento de rastreio prospetivo para o ASD, por prestadores de cuidados primários de saúde, sem requerer treino por parte do examinador nem observação da criança. O M-CHAT inclui os primeiros nove itens do CHAT (secção correspondente às respostas dos pais), mas foi expandido para um questionário de 23 questões, simples e facilmente preenchido pelos pais/cuidadores. As 14 questões adicionadas permitem uma ampla avaliação do comprometimento sensorial e motor, referenciamento social, imitação e reação ao próprio nome. Após avaliação dos itens mais preditivos do diagnóstico de ASD identificaram-se seis itens “críticos” (itens 2, 7, 9, 13, 14, 15). A pontuação do M-CHAT consiste na conversão das respostas do tipo sim/não para as pontuações de aprovação/reprovação. Uma criança falha no teste se falhar em dois ou mais dos seis itens críticos ou quando três dos itens totais falharem [41].

Em Portugal, o Programa Nacional de Saúde Infantil e Juventude recomenda a versão em português deste questionário (Anexo A) para o rastreio de ASD durante os cuidados primários, bem como as avaliações das crianças aos 18 e aos 24 meses [46].

2.2.1 Validade de Diagnóstico

No estudo inicial, Robins *et al.* [41] testaram 1293 crianças com idades entre os 16 e os 30 meses. Das 1293 crianças inquiridas, 1122 crianças foram bem sucedidas no teste e portanto não foram selecionadas, enquanto que 171 crianças foram encaminhadas para serviços de intervenção precoce (consideradas como sendo de alto risco de ASD ou outra deficiência de desenvolvimento). Nos casos de insucesso no teste, a equipa de pesquisa realizava-o novamente via chamada telefónica, e caso a criança falhasse novamente, era realizada uma avaliação clínica minuciosa. Das 58 crianças com insucesso nos dois testes M-CHAT, 39 receberam um diagnóstico de ASD e as

restantes 19 apresentaram atraso na linguagem ou no desenvolvimento global. Apenas 3 das 39 crianças com ASD correspondiam à população não selecionada, sendo que a maioria foi identificada a partir da amostra encaminhada para os serviços de intervenção precoce. Robins *et al.* [41] calcularam a sensibilidade, a especificidade e o VPP para várias combinações de itens do M-CHAT e demonstraram que, para esta amostra, os parâmetros do teste eram razoavelmente fortes (VPP de 68% nos dois estágios).

O resultado do questionário M-CHAT preenchido pelos pais com um acompanhamento especializado para esclarecimento das perguntas, é estimado que obtenha um VPP entre 0,54 a 0,74 em amostras de baixo risco [47, 48, 49]. Pandey *et al.* [40] declararam que o VPP do M-CHAT (usado para triagem de primeiro nível com entrevista de acompanhamento numa amostra comunitária de baixo risco) é menor em crianças mais jovens, com um VPP de 0,28 em crianças entre os dezasseis e os vinte e três meses comparando com um VPP de 0,61 em crianças dos vinte e quatro aos trinta meses.

Kleinman *et al.* [48] e Robins [47] avaliaram o M-CHAT com amostras de 3793 e 4797 crianças, respetivamente, com idades entre os 14 e os 30 meses. Kleinman *et al.* obtiveram um VPP de 0,36 na triagem inicial, que melhorou para 0,74 na triagem precedida de entrevista telefónica de acompanhamento. Robins obteve um VPP muito menor (0,06), mas após a entrevista por telefone, o VPP aumentou para 0,57. Canal-Bedia *et al.* [50] avaliaram uma tradução para o espanhol do M-CHAT e obtiveram um VPP de 0,35 e 0,19 numa amostra combinada de crianças de baixo ($N = 2480$) e alto risco ($N = 2055$), respetivamente.

Considerando apenas os dezanove itens mais preditivos dos testes M-CHAT executados por 84 pais de crianças de dois a três anos de idade, Eaves, Wingert e Ho [51] obtiveram uma sensibilidade alta (0,92), mas uma especificidade muito baixa (0,27). Snow e Lecavalier [52] utilizaram como critério de falha três dos itens do M-CHAT de 56 crianças com idade entre os dezoito e os quarenta e oito meses. A sensibilidade e especificidade foram de 0,88 e 0,38, respetivamente, o VPP foi de 0,83, o Valor Preditivo Negativo (VPN) foi de 0,50 e a AUC de 0,70. Em 2015, Charman *et al.* [44], utilizando todas as respostas fornecidas pelo teste M-CHAT, obtiveram valores semelhantes: sensibilidade e especificidade de 0,82 e 0,50, respetivamente, AUC de 0,77 e VPP de 0,49.

Para determinar se o M-CHAT, realizado durante o exame de rotina dos 18 meses, identifica o transtorno do espectro do autismo, foi executado o teste em 1851 crianças

que compareceram ao *check – up*. Conhecido o diagnóstico final após três anos, este estudo obteve uma sensibilidade de 0,48, uma especificidade de 0,99 e o valor preditivo positivo de 0,455 para crianças com ASD [53]. Com a mesma amostra, Kamio *et al.* [54] voltaram a analisar a prevalência de ASD mas isolando os seis itens altamente discriminativos do M-CHAT. Neste caso, a sensibilidade melhorou significativamente (0,67) enquanto que a especificidade e o VPP diminuíram para 0,90 e 0,17, respetivamente.

O M-CHAT foi ainda testado em crianças com uma média de idades relativamente superior ao recomendado para este teste — 35 meses [55]. A sensibilidade do M-CHAT para o autismo relatada neste estudo foi de 0,82 para a amostra total e de 0,93 para a metade mais jovem da amostra, sendo bastante consistente com outros estudos efetuados. Também Samadi e McConkey [56] implementaram o teste em crianças com idades superiores às recomendadas, entre os dois e os cinco anos. Este estudo obteve resultados bastante satisfatórios em termos de sensibilidade (0,90) e de especificidade (0,82), sendo o resultado não favorável somente para o VPP (0,05).

No geral, estes estudos sobre o teste M-CHAT (resumidos na tabela 2.1) mostraram que é possível identificar prospectivamente ASD, mesmo em crianças sobre as quais os pais e os profissionais não tenham preocupações preexistentes, a partir dos dezoito meses ou mais cedo no segundo ano de vida.

Tabela 2.1: Resultados de avaliação do M-CHAT.

	Idade	Amostra	Sens.	Esp.	VPP	VPN	AUC
Robins <i>et al.</i> (2001) [41]	24 meses	1293	NP	NP	0,68	NP	NP
Wong <i>et al.</i> (2004) [57]	18–24 meses	212	0,931	0,768	NP	NP	NP
Eaves, Wingert e Ho (2006b) [51]	2–6 anos	178	0,92	0,27	0,68	NP	NP
Kleinman <i>et al.</i> (2008) [48]	16–30 meses	3793	NP	NP	0,74	NP	NP
Robins (2008) [47]	16–27 meses	4797	NP	NP	0,57	NP	NP
Pandey <i>et al.</i> (2008) [40]	16–30 meses	6776	NP	NP	0,61	NP	NP
Snow e Lecavalier (2008) [52]	18–48 meses	56	0,88	0,38	0,83	0,50	0,70
Canal-Bedia <i>et al.</i> (2011) [50]	18–36 meses	2055	NP	NP	0,35	NP	NP
Chlebowski <i>et al.</i> (2013) [49]	18–30 meses	18969	NP	NP	0,54	NP	NP
Kamio <i>et al.</i> (2014) [53]	18–21 meses	1851	0,48	0,99	0,46	NP	NP
Kamio <i>et al.</i> (2015) [54]	18–21 meses	1851	0,67	0,90	0,17	NP	NP
Samadi e McConkey (2015) [56]	2–5 anos	2941	0,90	0,82	0,05	NP	NP
Charman <i>et al.</i> (2015) [44]	18–48 meses	120	0,82	0,50	0,49	0,82	0,66

* Sens. = Sensibilidade; Esp. = Especificidade; VPP = Valor Preditivo Positivo; VPN = Valor Preditivo Negativo; AUC = Área sob a curva; NP = *Not Provided*

2.2.2 Itens críticos do M-CHAT

No mesmo estudo em que foi analisada a fiabilidade do teste M-CHAT, Robins *et al.* [41] avaliaram ainda que itens do teste seriam mais discriminativos. Para cada item do teste foi obtido o peso estatístico correspondente através da análise da função discriminante *Discriminant Function Analysis* (DFA). Os itens com os maiores pesos (em ordem decrescente) são 7 (apontamento declarativo), 14 (reação ao próprio nome), 2 (interesse por outras crianças), 9 (trazer objetos para mostrar), 15 (seguir um ponto) e 13 (imitação de ações) [41].

Wong *et al.* [57], recorrendo ao coeficiente discriminante padronizado, determinaram a ordem das questões que foram mais úteis na diferenciação entre os grupos de crianças com autismo e sem autismo. Sete questões-chave foram identificadas com este critério, pela ordem apresentada: pergunta 13 ("Imita o adulto?"), pergunta 5 ("Brinca ao faz-de-conta, por exemplo, falar ao telefone ou dar de comer a uma boneca, etc.?"), pergunta 7 ("Aponta com o indicador para mostrar interesse em alguma coisa?"), pergunta 23 ("Procura a sua reação facial quando se vê confrontada com situações desconhecidas?"), pergunta 9 ("Alguma vez lhe trouxe objetos (brinquedos) para lhe mostrar alguma coisa?"), pergunta 15 ("Se apontar para um brinquedo do outro lado da sala, a criança acompanha com o olhar?"), e a pergunta 2 ("Interessa-se pelas outras crianças?").

Também Kamio *et al.* [54] utilizaram este critério para identificar os seis itens críticos que mostraram ser os mais discriminativos para crianças japonesas de 18 meses com ASD em comparação com crianças com desenvolvimento normal: apontamento imperativo, imitação de ações, fingir jogo, seguir um ponto, compreensão de linguagem, trazer objetos para mostrar (6, 13, 5, 15, 21 e 9, respetivamente).

Srisinghasongkram *et al.* [58] analisaram os elementos do teste M-CHAT segundo a sua sensibilidade e especificidade para discriminar o ASD de não-ASD numa amostra de 841 crianças tailandesas entre os dezoito e os quarenta e oito meses de idade. Desta forma, identificaram os itens que forneciam sensibilidade aceitável e especificidade mais alta: interesse por outras crianças, apontamento imperativo, apontamento declarativo, jogo funcional, trazer objetos para mostrar, imitação de ações e seguimento do olhar (itens 2, 6, 7, 8, 9, 13 e 17).

2.3 SCQ

O Questionário de Comunicação Social (SCQ) [59] é um interrogatório de quarenta perguntas de resposta "sim" ou "não", baseado no questionário mais extenso ADI-R [60] e destinado aos pais das crianças. Este teste pode ser preenchido em menos de dez minutos, e o resultado é obtido em menos de cinco minutos [43].

O primeiro item do SCQ refere-se ao nível de linguagem atual da criança e permite diferenciar crianças verbais e não verbais. Nos casos de crianças não verbais, seis das perguntas são eliminadas e não são consideradas no resultado final. Para cada um dos itens aplicáveis é dado um valor de 1 ou 0, de acordo com a presença ou ausência do comportamento anormal, respetivamente. O resultado total varia de 0 a 33 para os não verbais e de 0 a 39 para os verbais, sendo que o resultado maior corresponde a uma maior gravidade clínica. Um ponto de corte de 15 para crianças verbais e não verbais é geralmente usado para identificar aqueles que devem receber uma avaliação clínica especializada mais completa, pois serão mais propensos a ter ASD [59].

2.3.1 Validade de diagnóstico

O SCQ foi desenvolvido e testado em duzentas indivíduos de idades compreendidas entre os 4 e os 40 anos (160 com ASD e 40 com distúrbios não ASD, como distúrbios de linguagem ou deficiência intelectual) [59].

Na amostra de desenvolvimento, os indivíduos com ASD tiveram uma pontuação significativamente maior do que aqueles sem ASD em 33 dos 39 itens. A curva ROC indicou que a SCQ foi capaz de discriminar entre diagnósticos de ASD e não-ASD, sendo o seu valor igual a 0,86. A sensibilidade foi de 0,85, a especificidade foi de 0,75, o VPP foi de 0,93 e o VPN foi igual a 0,55, quando um ponto de corte de 15 foi usado para diferenciar os ASD dos não-ASD [59].

O SCQ foi sujeito a um elevado escrutínio, com um largo número de estudos adicionais que analisaram a validade diagnóstica em diversas idades (dezassete meses até quarenta anos) e amostras. A tabela 2.2 resume algumas evidências de validade disponíveis para o SCQ. Corsello *et al.* [61] examinaram o SCQ em crianças entre os 2 a 16 anos (N = 590) através de diagnósticos confirmados por consenso de dois médicos experientes, utilizando os critérios do DSM-IV, após uma avaliação clínica, que incluiu a ADI-R e a ADOS. Os valores de sensibilidade e especificidade foram

avaliados ao longo das idades e entre ASD e não-ASD. Estes foram inferiores aos apresentados na amostra de desenvolvimento, e a sensibilidade para o grupo final foi muito baixa (0,45). Neste estudo, ficou comprovado que o SCQ poderá ser melhor na identificação de crianças mais velhas (mais de sete anos) em comparação com crianças mais jovens. Wiggins *et al.* [62] descobriram também que o SCQ foi menos efetivo na identificação correta de sujeitos mais jovens da amostra entre dezassete a quarenta e cinco meses de idade. A sensibilidade obtida foi de apenas 0,47 analisaram também especificidade de 0,89. Recentemente, Charman *et al.* (2015) [44] analisaram também o teste em crianças muito pequenas, com idades compreendidas entre os dezoito e os quarenta e oito meses, tendo apenas obtido valores de 0,64 e de 0,75 de sensibilidade e especificidade, respetivamente. Quando combinado com uma amostra de trinta e setenta meses de idade ($N = 65$) [52], o SCQ teve também sensibilidade (0,70) e especificidade (0,52) mais baixas do que o relatado no manual. Neste estudo, os diagnósticos foram confirmados com base nos critérios do DSM-IV por consenso de especialistas, utilizando diversos testes para classificação.

O SCQ apresentou melhor desempenho numa amostra de crianças dos quatro aos quatorze anos ($N = 49$) [63]. Nesta publicação obteve-se uma sensibilidade de 0,92 e uma especificidade de 0,62. Eaves, Wingert e Ho [51] examinaram a validade discriminante em crianças de quatro a seis anos ($N = 94$). A sensibilidade obtida aproximava-se de um valor aceitável (0,74), em contraste com a baixa especificidade (0,54). Chandler *et al.* [64] encontraram altos valores de sensibilidade e especificidade ao diferenciar crianças com ASD de crianças sem ASD (0,88 e 0,72, respetivamente). Charman *et al.* [65] encontraram valores semelhantes numa amostra de nove a treze anos de idade ($N = 119$).

Eaves *et al.* [66] examinaram a validade diagnóstica numa amostra de crianças com idade média de cinco anos ($N = 151$), revelando uma sensibilidade geral de 0,71 e uma especificidade de 0,62. Allen *et al.* [67] encontraram valores igualmente baixos numa amostra de oitenta e uma crianças dos dois aos seis anos que tinham sido encaminhadas a uma clínica para um problema inespecífico (sensibilidade 0,60, especificidade 0,70). Lee *et al.* [68] examinaram o SCQ em duzentas e sessenta e oito crianças com idades entre os três e os cinco anos tendo sido obtida uma sensibilidade baixa (0,54), ao invés da excelente especificidade (0,92). Examinando valores dentro de uma subamostra de três anos de idade ($N = 78$) os autores encontraram uma sensibilidade de 0,70 e uma especificidade de 0,88.

Concluindo, o SCQ é um instrumento de triagem de ASD bastante analisado segundo

a sua validade diagnóstica em amostras muito diferenciadas. Esses estudos indicam que o SCQ tem utilidade como instrumento de triagem, mas embora seja apropriado para uma ampla faixa etária, não é ideal quando usado com populações mais jovens (ou seja, crianças de dois a três anos). Este teste foi projetado para indivíduos com idade acima de quatro anos e, aparentemente, terá melhor desempenho em indivíduos acima dos sete anos.

Tabela 2.2: Resultados de avaliação do SCQ.

	Idade	Amostra	Sens.	Esp.	VPP	VPN	AUC
Berument <i>et al.</i> (1999) [59]	4-40 anos	200	0,85	0,75	0,93	0,55	0,86
Eaves, Wingert e Ho (2006a) [51]	4-6 anos	94	0,74	0,54	0,65	NP	NP
Eaves, Wingert e Ho (2006b) [66]	32-82 meses	151	0,71	0,79	0,65	NP	NP
Chandler <i>et al.</i> (2007) [64]	10-14 anos	255	0,88	0,72	0,64	0,91	0,88
Corsello <i>et al.</i> (2007) [61]	2-16 anos	590	0,71	0,71	0,88	0,45	0,77
Allen <i>et al.</i> (2007) [67]	2-6 anos	81	0,60	0,70	NP	NP	NP
Charman <i>et al.</i> (2007) [65]	9-13 anos	119	0,86	0,78	0,74	0,88	0,90
Lee <i>et al.</i> (2007) [68]	3-5 anos	268	0,54	0,92	0,62	0,89	0,88
Wiggins <i>et al.</i> (2007) [62]	17-45 meses	37	0,47	0,89	NP	NP	NP
Witwer <i>et al.</i> (2007) [63]	4-14 anos	49	0,92	0,62	0,87	0,73	0,89
Snow <i>et al.</i> (2008) [52]	30-70 meses	65	0,70	0,52	0,70	0,52	0,67
Charman <i>et al.</i> (2015) [44]	18-48 meses	120	0,64	0,75	0,60	0,78	0,70

* Sens. = Sensibilidade; Esp. = Especificidade; VPP = Valor Preditivo Positivo; VPN = Valor Preditivo Negativo; AUC = Área sob a curva;

2.4 Imputação de dados

Atualmente, existem diferentes sistemas de recolha de dados, como por exemplo, a entrevista pessoal realizada em computador, a entrevista por telefone ou até mesmo a entrevista pessoal. No entanto, não há nenhum sistema que conceda conjuntos de dados perfeitos, e um certo risco de gerar erros estará sempre presente. Portanto, é comum os conjuntos de dados conterem erros na forma de valores ausentes ou inconsistentes (*missings*). Os *missings* devem-se à ausência de resposta, enquanto os valores inconsistentes são produzidos quando a resposta não é registada com precisão [69].

Existem três problemas principais que podem surgir ao trabalhar com bases de dados incompletas. Primeiro, há uma perda de informação e, conseqüentemente, uma perda de eficiência. Em segundo lugar, o trabalho de computação e análise dos dados torna-se mais complexo devido às irregularidades na estrutura dos dados e à impossibilidade de usar o *software* padrão. Em terceiro lugar, e mais importante, os resultados podem ser enviesados devido às diferenças sistemáticas entre os dados observados e não observados [70]. Desta forma, torna-se essencial efetuar o tratamento dos *missings*, sendo uma das estratégias possíveis para lidar com esse problema a execução da imputação de dados, definida como o processo pelo qual os *missings* de um conjunto de dados são estimados por valores apropriadamente calculados. Por outras palavras, a imputação de dados é capaz de preencher as lacunas do conjunto de dados, produzindo uma base de dados completa.

O impacto dos *missings* no resultado da análise estatística depende do mecanismo que fez com que os dados fossem omissos e da forma como estes são trabalhados. A escolha de um método apropriado para substituir os dados em falta depende, entre outros, do padrão dos dados perdidos e do motivo pelo qual eles se encontram em falta [71]. Em grande parte dos estudos epidemiológicos, a falta de dados é um problema, especialmente com questionários que contêm várias perguntas. Os métodos comumente utilizados para solucionar este tipo de problema são os métodos de tendência central ou a regressão simples [72].

Tanto a imputação através da moda como a imputação por regressão linear correspondem a métodos de imputação simples. Este tipo caracteriza-se pelo facto de gerarem um único valor de substituição para cada *missing*. Isso contrasta com a imputação múltipla (não abordada nesta análise), que cria várias cópias do conjunto de dados e imputa cada cópia com diferentes estimativas plausíveis dos valores

omissos [73].

2.4.1 Imputação pela Moda

É um método simples em que qualquer *missing* de uma variável qualitativa é substituído pela moda dos valores observados para essa variável. Portanto, se uma variável apresentar vários *missings* para amostras diferentes, todos eles serão imputados com o mesmo valor — a moda [69].

Embora a substituição dos *missings* pela moda possa diminuir a correlação de uma variável com todas as restantes no conjunto de dados [74], este método não altera a moda da *feature* com *missings*.

2.4.2 Imputação pela Regressão Linear

Esta abordagem usa as informações das *features* completas para preencher as *features* incompletas. As *features* tendem a ser correlacionadas, portanto, as imputações são geradas através da informação dos dados observados [73].

Neste tipo de imputação é necessário ajustar um modelo de regressão à *feature* de interesse, como as *features* de resposta e as *features* relevantes. Desta forma, os coeficientes são estimados e, em seguida, os *missings* podem ser previstos pelo modelo ajustado [75].

As imputações recorrendo à moda são simples, mas subestimam a variância e ignoram a relação com outras *features*. O método de regressão preserva a correlação com outras *features*, mas a variabilidade dos *missings* é subestimada. A variabilidade pode ser ajustada adicionando erros aleatórios ao modelo de regressão [75].

Segundo Widaman [74], se a quantidade de *missings* no conjunto de dados for reduzida, uma única imputação poderá ser suficiente, já que o problema dos *missings* é tão pequeno que os benefícios associados à imputação múltipla provavelmente não justificam o esforço adicional.

2.5 Seleção de *features*

A classificação de dados é um esforço multidisciplinar para extrair informação de classes tendo em conta o conhecimento dos dados. O aumento de grandes conjuntos de dados em diversos domínios coloca desafios sem precedentes à classificação dos

dados [76]. Para além dos conjuntos de dados serem cada vez maiores, novos tipos de dados estão a tornar-se predominantes. Por estes motivos, para obter uma classificação de dados bem-sucedida, a seleção de *features* é uma componente cada vez mais indispensável.

A seleção de *features* corresponde ao processo de seleção de um subconjunto de *features* originais de acordo com determinados critérios, sendo uma técnica importante e frequentemente usada na classificação de dados para a redução da dimensão. Esta técnica reduz o número total de *features* removendo as irrelevantes, redundantes ou que possuam erros associados, produzindo efeitos palpáveis em determinadas aplicações. Deste modo, esta etapa torna-se fundamental e motivada por: acelerar o algoritmo de classificação de dados, melhorar a precisão do modelo preditivo e obter uma melhor compreensão das classes [77].

Vários estudos mostraram que algumas *features* podem ser removidas sem a deterioração do desempenho e, por isso, a seleção de *features* tem sido, há décadas, uma área ativa de pesquisa na classificação de dados, e amplamente aplicada a novas áreas. Com o surgimento de novas aplicações, conseqüentemente, surgem também muitos desafios [76].

Para selecionar *features* relevantes, é necessário determinar a qualidade das mesmas usando um critério de seleção. A separabilidade da classe é frequentemente usada como um dos critérios básicos de seleção, ou seja, quando um conjunto de *features* maximiza a separabilidade das classes, é considerado adequado para classificação [77].

Os algoritmos de seleção de *features* podem ser agrupados em duas categorias, do ponto de vista do resultado do método: classificação de *features* ou seleção de um subconjunto de *features*. A primeira cria um *ranking* de *features* de acordo com uma métrica de avaliação; a segunda escolhe um conjunto mínimo de *features* que funcione nesse mesmo conjunto e que satisfaça um critério de avaliação, sem necessariamente determinar que *features* individuais são mais significativas do que outras [78].

Da perspectiva do *design* do modelo preditivo, a seleção de *features* é uma etapa essencial, dado que o classificador é aplicado ao subconjunto selecionado de *features*. Os métodos *filters* [79] atribuem uma pontuação a cada conjunto de *features* com base nas estatísticas dos dados, independentemente de qualquer modelo preditivo, e tendem a ser computacionalmente mais eficientes do que os métodos *wrappers*

e os métodos *embedded*, pois não exigem a criação de classificadores. Os métodos *wrappers* [80], por outro lado, avaliam cada subconjunto de *features* com base no erro do modelo preditivo e, portanto, são específicos do modelo. Se tivermos d *features*, existem $2^d - 1$ conjuntos de *features* possíveis para avaliar, atribuindo uma pontuação ou construindo um modelo preditivo para cada subconjunto possível, sendo que este último é dispendioso em termos computacionais. Por esse motivo, atualmente utilizam-se também os métodos *filters* e *wrappers* combinados como estratégia de pesquisa, permitindo avaliar apenas alguns dos subconjuntos de *features*. Esta estratégia designa-se por métodos *embedded* [81].

Uma revisão mais completa sobre os métodos de seleção de *features* e as suas aplicações na classificação está disponível em Bolon-Canedo *et al.* [82] e Saeys *et al.* [83].

Um algoritmo de seleção de *features* pode obter três tipos diferentes de resultados [79]:

1. Uma ponderação das *features* que dá a relevância de cada uma relativamente às classes (por exemplo, a informação mútua de cada *feature* com a respetiva classe ou os coeficientes de uma regressão linear);
2. Um *ranking* das *features* que dá importância relativa de cada uma no que concerne às classes (por exemplo, *Recursive Feature Elimination*);
3. Um subconjunto de *features* que indicam quais as que são relevantes para as classes (por exemplo, com qualquer método de *wrapper*). Um conjunto de *features* pode ser obtido colocando um limite num *ranking*, mas o contrário não é possível.

Muitos estudos comparativos de métodos existentes para seleção de *features* foram executados, por exemplo, um estudo experimental de oito métodos *filters* (usando informações mútuas) usado em 33 conjuntos de dados [84]. Sete métodos *filters*, dois métodos *wrappers* e dois métodos *embedded* foram aplicados a 11 conjuntos de dados sintéticos (testados por quatro classificadores), sendo utilizados para estudo comparativo do desempenho da seleção de *features* na presença de *features* irrelevantes, ruído nos dados, redundância e pequena razão entre o número de *features* e o número de amostras [82].

2.5.1 SVM-RFE

O *Recursive Feature Elimination* é um processo iterativo que consiste em três etapas principais: treinar o classificador, estabelecer a classificação das *features* e remover as *features* com menor importância [85].

O SVM-RFE é um conhecido método *embedded* e foi primeiramente introduzido em 2002 na seleção de genes para a classificação do cancro [86]. Este método classifica recursivamente as *features* com o *Support Vector Machine* (SVM), selecionando-as de acordo com os pesos obtidos no classificador. Em seguida, para o caso linear, as *features* são classificadas de acordo com o valor absoluto do seu peso no vetor de ponderação do hiperplano — $D(\vec{x}) = (\vec{w} \cdot \vec{x}) + b$. A variável com o menor peso w^2 tem uma menor contribuição para o hiperplano resultante e pode ser descartada [87]. O procedimento acima é então repetido até que seja obtido um conjunto de *features* vazio.

O procedimento de *Recursive Feature Elimination* (RFE) comumente utilizado é o seguinte:

1. Início: *ranking* de *features* $R = []$;
subconjunto selecionado $S = [1, \dots, d]$;
2. Repete-se para todas as *features* a classificar:
 - (a) Treinar o SVM linear nos dados de treino fornecidos e *features* em S ;
 - (b) Calcular o vetor de peso recorrendo à equação 2.1:

$$w = \sum_{i=1}^l \alpha_i y_i z_i, \quad (2.1)$$

Onde x_i denota o vetor de exemplo i nos dados iniciais e z_i denota o vetor correspondente no espaço das *features*, para um problema típico de classificação com l vetores de suporte $(x_1, y_1), \dots, (x_l, y_l)$, onde y_1 indica a classe de x_i ;

- (c) Calcular as pontuações de classificação para *features* em S : $c_i = (w_i)^2$;
- (d) Encontrar a *feature* com a menor pontuação no *ranking*: $e = \operatorname{argmin}_i c_i$;

- (e) Atualizar R : $R = R[e, R]$;
- (f) Atualização S : $S = S - [e]$.

3. Lista de *features* do *ranking* R .

Guyon *et al.* [86] propuseram e aplicaram o método SVM-RFE na seleção de genes através de duas bases de dados de cancro. Os resultados obtidos mostraram melhorias significativas sobre o método base e, a relevância dos genes obtida pelo SVM-RFE apresenta uma relação plausível com o cancro. Pelo contrário, outros métodos testados selecionaram genes que estão correlacionados, mas não são relevantes para o diagnóstico de cancro. O método de RFE foi demonstrado para classificadores lineares, incluindo o SVM. O estudo efetuado indica que o SVM-RFE é muito mais robusto ao *overfitting* de dados do que outros métodos.

Para analisar a influência que a seleção de *features* produz no desempenho do classificador SVM, Duan e Rajapakse [88] testaram os Testes Estatísticos e SVM-RFE. Pelos resultados de elevada precisão de classificação obtidos, verifica-se que o SVM-RFE tem um desempenho superior aos Testes Estatísticos e, evidentemente, também quando comparado com a situação em que não é aplicado nenhum método de seleção.

Outro estudo que analisa alguns métodos de seleção de *features* é apresentado por Saeys *et al.* [89]. Os métodos *SymmetricalUncertainty*, *Relief*, *RandomForest* e SVM-RFE foram aplicados a dez subamostras de cada uma das seis bases de dados (cada subamostra continha 90% dos dados do total da base de dados) e, posteriormente, procedeu-se à média do parâmetro utilizado para a avaliação dos métodos (robustez). Para todos os métodos de seleção de *features* foram ainda testados três modelos preditivos. Os melhores resultados obtidos após a implementação de métodos de seleção de *features* corresponderam ao SVM-RFE. A execução deste método obteve resultados estáveis para os diversos classificadores provando ser uma técnica robusta e com bom desempenho de classificação. Em 2013, o desempenho do SVM-RFE foi também analisado e comparado, desta vez com o método *Principal component analysis* (PCA) e com uma base de dados relativa ao cancro da mama [90]. Uma vez mais, os resultados mostram que o SVM-RFE é mais adequado para realizar a seleção de *features* a classificar.

Embora muitos algoritmos tenham sido desenvolvidos, o algoritmo *Support Vector Machine - Recursive Feature Elimination* é um dos melhores algoritmos de seleção

de *features*, provando ser robusto ao ajuste de dados comparativamente com outras técnicas [86] e mostrando também o seu excelente desempenho em diversas áreas, por exemplo, genómica [91] e proteómica [92].

2.5.2 Testes Estatísticos (Teste-t)

O uso de testes paramétricos requer a verificação de dois pressupostos: i) distribuição normal; ii) homogeneidade de variância. Para testar se uma amostra segue uma distribuição normal, é realizado o teste de Shapiro-Wilk [93]. Para determinar se duas *features* têm variância homogénea, é utilizado o teste de Levene [94].

O teste t-Student [95] para amostras independentes é um teste paramétrico que compara médias populacionais para um par de *features*. A hipótese nula do teste afirma que não há diferença estatística entre a média das duas populações. Para *features* que possuem variâncias diferentes utilizou-se o teste t de Welch.

O teste de Mann-Whitney [96] é um teste não paramétrico aplicado para duas *features* independentes — é uma alternativa não paramétrica ao teste t-Student para amostras emparelhadas. Quando usado para dados emparelhados, a hipótese nula do teste afirma que a diferença entre a mediana dos pares de observações é zero.

O teste de Fisher é um teste não paramétrico aplicado para *features* qualitativas e, cuja hipótese nula afirma que não existe relação entre elas.

O facto de estes testes estatísticos derivarem da suposição de que as *features* são independentes pode corresponder simultaneamente a um benefício e a uma desvantagem. Por um lado, o pressuposto da independência torna este método muito rápido. Por outro lado, a suposição de independência pode não ser válida para todos os conjuntos de dados [85].

Tradicionalmente, o valor $p < 0,05$, ou $p < 5\%$, tem sido recomendado como um limiar no qual se pode rejeitar a hipótese nula [97]. É importante ainda referir que o valor p de um teste t usado para comparar dois grupos apenas nos diz se a diferença média entre os grupos é estatisticamente significativa ou não. Com esta abordagem não é possível avaliar se existe alguma diferença significativa ou alguma relevância entre *features*.

Esta metodologia tem sido amplamente utilizada na seleção de *features* para di-

versas bases de dados e em diferentes procedimentos. A abrangência dos testes estatísticos na investigação é tão elevada que estes são utilizados sem uma avaliação direta. Seguidamente, serão apresentados alguns estudos efetuados diretamente à avaliação desta abordagem.

Em 2003, Wu *et al.* [98] utilizaram testes estatísticos para selecionar *features* e compararam essa seleção com o *Random Forest* (RF). Na aplicação do Teste-t para a seleção, os classificadores de SVM, *Linear Discriminant Analysis* (LDA) e *Random Forest* obtiveram os melhores resultados (uma precisão de aproximadamente 85%).

Os testes estatísticos foram também avaliados para amostras de tamanho extremamente pequeno. Os resultados desse estudo mostraram que o Teste-t provoca percentagens de erro próximas de 5% na maioria dos casos [99].

Num estudo extensivo, Haury *et al.* [100] compararam 32 métodos de seleção de *features* em 4 conjuntos de dados de expressão de genes para o prognóstico do cancro da mama. Essa comparação foi efetuada tendo por base os critérios de desempenho preditivo, estabilidade e interpretabilidade das *features* identificadas. Com esta análise, corroboraram que o método de seleção de *features* tem uma influência significativa na precisão e estabilidade dos modelos preditivos. Para além disso, verificaram que, entre todos os métodos testados, apenas o Teste-t responde corretamente quanto à precisão e estabilidade dos dados. Desta forma, chegou-se à conclusão de que o Teste-t era o que oferecia o melhor desempenho e robustez.

2.6 Balanceamento

O conjunto de dados desequilibrados, um problema frequentemente encontrado na aplicação no mundo real, pode causar um efeito seriamente negativo no desempenho de classificação dos algoritmos de *machine learning*.

Um conjunto de dados é designado por desequilibrado se contiver muito mais amostras de uma classe comparativamente às restantes classe. Nesse cenário, os classificadores podem ter um bom acerto na classe maioritária, mas um acerto muito baixo na classe minoritária, devido à influência que a maioritária tem nos critérios de treino. A maioria dos algoritmos de classificação originais procura minimizar a taxa de erro — a percentagem da previsão incorreta das classes — mas ignora a diferença entre o número de amostras de cada classe. Em particular, é assumido implicitamente que todos os erros de classificação incorreta têm o mesmo peso [101].

Muitos trabalhos de pesquisa sobre conjuntos de dados desequilibrados concordaram que, devido à distribuição desigual de classes, o desempenho dos classificadores existentes é tendencioso em relação à classe majoritária. As razões para o mau desempenho dos algoritmos de classificação existentes, em conjuntos de dados desequilibrados, são essencialmente causados por se assumir que: há distribuição igual de dados para todas as classes e, também que os erros de classes diferentes têm o mesmo peso [101].

Várias soluções para o problema do desequilíbrio de classes foram propostas anteriormente, tanto ao nível dos dados como ao nível do algoritmo. Ao nível dos dados, essas soluções incluem muitas formas diferentes de re-amostragem, como *oversampling* aleatório com substituição, *undersampling* aleatório, *oversampling* direcionado (na qual nenhum novo exemplo é criado, mas a escolha de amostras a serem substituídas é informada e não aleatória), *undersampling* direcionado e combinações das técnicas referidas. Ao nível algorítmico, as soluções incluem ajustar os pesos das várias classes para contrabalançar o seu desequilíbrio [102].

Oversampling

O método mais simples para aumentar o tamanho da classe minoritária corresponde ao *oversampling* aleatório, ou seja, um método que equilibra a distribuição de classes através da replicação aleatória de amostras positivas. No entanto, como este método replica os exemplos existentes na classe minoritária, o *overfitting* tem uma maior probabilidade de ocorrer [103].

Undersampling

O *undersampling* usa um subconjunto da classe majoritária para treinar o classificador. Como muitos exemplos de classes majoritárias são ignorados, o conjunto de treino torna-se mais equilibrado e o processo de treino mais rápido. A técnica de pré-processamento mais comum é o *undersampling* aleatório. Amostras da classe majoritária são descartadas aleatoriamente do conjunto de dados [103].

A principal desvantagem do *undersampling* é o facto de descartar dados potencialmente úteis. A principal desvantagem do *oversampling* é que a cópia exata de exemplos existentes torna o *overfitting* provável, sendo uma segunda desvantagem o aumento do número de casos para treino e, conseqüentemente, do tempo de implementação.

2.7 Classificação

O objetivo principal da criação de um modelo preditivo é identificar padrões e gerar uma previsão do que pode ocorrer, tendo em conta os dados a testar. De forma a obter o melhor modelo, é necessário garantir a melhor base de dados para a qual se conhece o resultado e, para tal, processos como a imputação e a seleção de *features* podem ser determinantes. Por outro lado, estas técnicas precedem o método de classificação e, portanto, devem ser escolhidas adequadamente para não afetar o desempenho do classificador. Estes são três pontos-chave para o objetivo final do estudo e portanto devem ser avaliados: a qualidade da imputação, a qualidade da seleção de *features* e o desempenho da classificação.

Métodos "tradicionais" de classificação estatística (por exemplo, LDA e *Logistic Regression* (LR)) têm sido amplamente utilizados em problemas de classificação médica para os quais a variável critério é dicotómica [104, 105]. Mais recentemente, a pesquisa tem-se baseado na precisão e eficiência de *data mining*, com classificadores como Redes Neurais, *Support Vector Machine*, Árvores de Classificação e *Random Forest* usadas para predição médica e tarefas de classificação.

Pesquisas sobre a precisão comparativa de classificadores tradicionais (LDA e LR) *versus* novos métodos de *data mining* (exigem grande poder de computação) algoritmos iterativos inovadores e com intervenção do usuário, têm crescido constantemente. Vários autores referem que os classificadores de *data mining* têm maior precisão e menores taxas de erro do que os métodos tradicionais de classificação [106]. Os resultados que apresentam os melhores valores de exatidão para métodos de classificação mais recentes em comparação com os métodos tradicionais, bem como a estabilidade das pesquisas, são ainda controversos. A maioria das comparações entre métodos baseia-se apenas na exatidão total da classificação e/ou margens de erro. Para além disso, em contextos médicos, a sensibilidade e a especificidade, bem como a AUC são das principais métricas que devem ser consideradas ao comparar os classificadores e os métodos de diagnóstico [107].

Para testar o desempenho de métodos de *data mining* e *machine learning* como Redes Neurais, SVM e RF, Maroco *et al.* [107] compararam-nos com três classificadores "tradicionais" (Análise Discriminante Linear, Análise Discriminante Quadrática e Regressão Logística) quanto à exatidão de classificação geral, especificidade, sensibilidade, AUC e Press'Q. Os preditores do modelo foram 10 testes neuropsicológicos usados atualmente no diagnóstico de demência.

O SVM apresentou a maior exatidão de classificação geral (mediana = 0,76) e área abaixo da curva ROC (mediana = 0,90). No entanto, este método mostrou alta especificidade (mediana = 1,00), mas baixa sensibilidade (mediana = 0,30). O *Random Forest* ficou em segundo lugar na precisão geral (mediana = 0,73) e obteve bons valores de área sob a ROC (mediana = 0,73), especificidade (mediana = 0,73) e sensibilidade (mediana = 0,64). Dos métodos ditos tradicionais, o LR obteve os melhores valores de especificidade, exatidão e Press'Q [107], embora que inferiores aos métodos de *data mining*.

Support Vector Machine

O *Support Vector Machine* é um classificador de aprendizagem supervisionada que define um hiperplano que separa os dois grupos em análise. Os dados de treino são fornecidos ao classificador para que, a partir desses, se crie o hiperplano que categorizará os dados de teste [108].

No caso em que não seja possível dividir os dados de treino através de uma linha num plano bidimensional, uma terceira dimensão é adicionada, a linha de separação é definida e depois transformada no plano original. Essas transformações são chamadas de *kernels*. Para *kernels* não-lineares, a fronteira de decisão do *Support Vector Machine* não é um hiperplano no espaço característico original, mas uma hipersuperfície não linear, cuja forma depende do tipo de *kernel*.

Para obter melhor precisão na previsão da classe certa para cada amostra, o algoritmo possui alguns parâmetros que ajudam a aumentar a exatidão num período de tempo razoável. Como mencionado anteriormente, um dos parâmetros corresponde ao *kernel* [109]. Existe também um parâmetro de regulação, conhecido como o parâmetro C , que define o quanto se pretende evitar a classificação errônea de cada amostra de treino. Para valores superiores de C , o hiperplano apresentará uma margem menor entre as classes, se puder classificar os pontos de treino corretamente. Por outro lado, ao escolher um valor menor para esse parâmetro, o hiperplano de separação terá uma margem maior podendo, no entanto, classificar erroneamente. Outro fator importante é o parâmetro *gamma*, usado para o *kernel* gaussiano, que define o que influencia a linha de separação entre as classes significado que, para valores baixos, pontos que estão mais distantes dessa linha serão tidos em consideração, enquanto que para valores mais altos, apenas os mais próximos são usados [108].

O SVM é uma poderosa técnica de classificação amplamente utilizada na área de

reconhecimento de padrões. Em *machine learning*, existem muitos modelos preditivos que permitem proceder à classificação, do qual o SVM é um exemplo, mas que tem sido extensamente apontado como um dos que obtém os melhores desempenhos [110].

Random Forest

O *Random Forest* [111] é um algoritmo de classificação que utiliza um conjunto de árvores de decisão, cada uma das quais construída numa amostra de dados de treino que correspondem a um subconjunto de *features* selecionado aleatoriamente.

Este modelo constrói várias árvores de decisão e combina-as para aumentar a exatidão da classificação. As árvores de decisão são algoritmos que dividem os dados de treino em subgrupos diferentes identificando os encadeamentos. Apesar disso, as árvores de decisão e o RF não são a mesma coisa: as árvores de decisão formulam algumas regras ao analisar o conjunto de treino, que será usado para previsões no conjunto de dados de teste. Por outro lado, o algoritmo de *Random Forest* seleciona *features* e amostras aleatoriamente, constrói as várias árvores de decisão para cada subconjunto e, em seguida, calcula a média dos resultados obtidos [111].

O algoritmo de *Random Forest* possui um número de propriedades interessantes, tornando-o adequado para a classificação de dados: é aplicável quando há mais *features* do que observações, incorpora interações entre *features* e baseia-se na teoria de *ensemble learning*, que permite ao algoritmo treinar com precisão tanto funções de classificação simples como complexas [112].

Esta técnica possui uma exatidão incomparável entre alguns dos algoritmos mais usados atualmente, apresentando um desempenho muito positivo relativamente a muitos classificadores, incluindo LDA, Redes Neurais e SVM. Além disso, este método possui apenas dois parâmetros que requerem definição por parte do utilizador: o número de árvores de decisão; e o número de *features* no subconjunto aleatório em cada nó da árvore. Esses parâmetros podem ser facilmente otimizados, embora o *Random Forest* não seja muito sensível à sua alteração [107].

O estudo desenvolvido por Díaz-Uriarte *et al.* [113] realizou uma avaliação empírica do modelo *Random Forest* no domínio da expressão génica de cancro, em 10 bases de dados diferentes, e concluiu que este classificador tem um desempenho preditivo ligeiramente superior ao SVM. Dois anos mais tarde, Statnikov *et al.* [112] vieram contrapor esses resultados. Num estudo de diagnóstico do cancro, o RF e o SVM

foram avaliados e comparados com recurso a 22 bases de dados de diagnósticos e prognósticos. Os resultados obtidos mostraram que o SVM supera o RF, na maioria dos conjuntos de dados, por uma grande margem.

Logistic Regression

O objetivo do método *Logistic Regression* é encontrar o modelo mais adequado para descrever a relação entre as classes (variável dependente ou resposta) e um conjunto de *features*. O método é relativamente robusto, flexível e fácil de usar. A LR é uma técnica de regressão não linear que assume que a probabilidade esperada de uma classe dicotómica é:

$$P = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots)}}, \quad (2.2)$$

onde X_i são *features* com valores numéricos e β_s são os coeficientes de regressão que quantificam a sua contribuição para a probabilidade.

A *Logistic Regression* tem sido testada e comparada com diversos modelos preditivos que têm vindo a ser criados. Song *et al.* [114] compararam a LR e *Artificial Neural Network* para deteção do cancro da mama usando dados médicos. Neste estudo, é mostrado que a LR e a *Artificial Neural Network* (ANN) têm quase a mesma eficácia, mas o segundo possui uma menor sensibilidade de deteção. Em [115], Rudolfer *et al.* compararam LR e *Decision Trees*, e concluíram que a eficácia de ambos os métodos é a mesma. Também Long *et al.* [116] compararam LR e *Decision Trees* na aplicação médica, considerando o critério da AUC. A comparação neste trabalho mostrou que, embora com uma pequena diferença, o LR possui uma melhor eficácia.

3

Metodologia

No presente trabalho, desenvolveu-se um modelo preditivo de ASD utilizando os métodos mais significativos da literatura, apresentados no capítulo anterior (capítulo 2). Para determinados modelos, foram criadas novas abordagens na tentativa de obter melhores resultados ou maior fiabilidade dos mesmos. A metodologia proposta para atingir esse fim é dividida em cinco etapas principais: (1) Análise da Base de Dados, (2) Imputação de Dados, (3) Seleção de *Features*, (4) Classificação e (5) Avaliação (figura 3.1). Na figura 3.2, são apresentados todos os métodos utilizados em cada uma das etapas executadas.

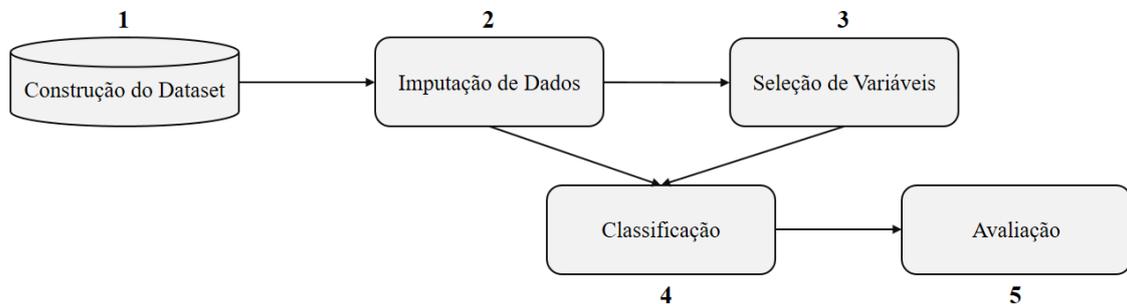


Figura 3.1: Resumo do procedimento experimental.

A aplicação destes procedimentos foi implementada usando o Python 2.7 por se tratar de uma fonte aberta, pela estrutura de dados de fácil utilização, pelos fortes recursos de integração e processamento de texto que contribuem para o aumento da velocidade, e pela fácil interação com a maioria das linguagens e plataformas.

3.1 Imputação de Dados

Inicialmente, analisou-se o número de *missings* presentes em cada *feature*. Tendo em conta que os métodos de imputação mais simples podem responder erroneamente a um valor elevado de *missings* e, dado que eliminar os registos com *missings* pode

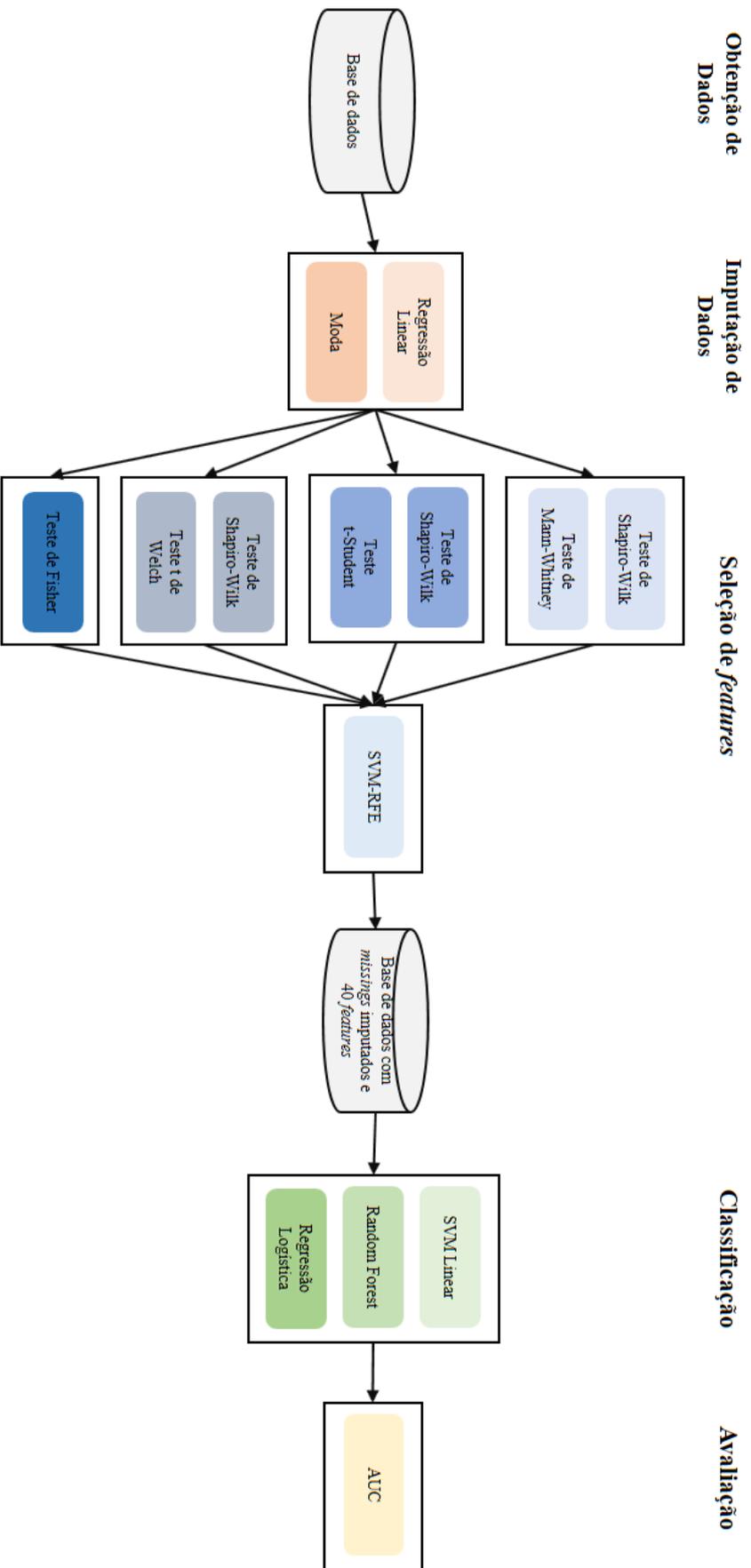


Figura 3.2: Representação esquemática detalhada do procedimento experimental.

prejudicar gravemente o tamanho da base de dados, considerou-se adequado eliminar as *features* com mais de 60 *missings*, o correspondente a 30% do número total de indivíduos.

Para as *features* com *missings*, mas em quantidade inferior aos 30% referidos, implementaram-se dois métodos de imputação simples bastante comuns: imputação recorrendo à tendência central e imputação pela regressão linear (consulte a secção 2.4).

Na análise da distribuição dos *missings* nas restantes *features*, deduziu-se que estes ocorriam por falta de resposta e de forma aleatória. Para uma correta imputação dos dados, foi necessário proceder à implementação de métodos diferentes de acordo com o tipo de *feature*. Uma vez que os *missings* se encontravam tanto em *features* quantitativas como qualitativas, foi necessário aplicar um método para cada um dos tipos: a imputação com recurso à moda para as *features* qualitativas, e a imputação com recurso à regressão linear para as *features* quantitativas.

Para execução do método de imputação recorrendo à regressão linear, utilizou-se a implementação de *LinearRegression()* em Python. Deste modo, a melhorar a imputação realizada, consideraram-se apenas as três *features* mais correlacionadas com a *feature* com *missings*. Por conseguinte, obtiveram-se os valores de correlação entre *features* implementando-se o método de *Spearman* através da função *corr()* e, posteriormente, procedeu-se à obtenção das estimativas para os valores em falta através da regressão linear. O coeficiente de *Spearman* foi utilizado como parâmetro de análise da correlação de *features* por ser bastante sensível ao erro e às discrepâncias dos dados e, para além disso, por ser apropriado tanto para *features* contínuas, como para *features* discretas, incluindo *features* ordinais.

Nos *missings* presentes em *features* qualitativas, procedeu-se à sua substituição pelos valores da moda. Embora esta abordagem não seja a mais robusta ou fidedigna, foi concretizada por ser de fácil implementação e execução e, porque, neste caso, apenas existiam três valores a substituir na base de dados total. Para casos como este, o valor da moda da *feature* obteve-se recorrendo à função *mode()*.

De forma a avaliar a eficácia e validade dos dados imputados, compararam-se os resultados obtidos pelos classificadores a partir de duas bases de dados distintas. A primeira, resultante do método de imputação apresentado, e a segunda correspondente à base de dados original. Dado que os classificadores não admitem *missings*, as amostras que não continham toda a informação foram eliminados.

3.2 Seleção de *features*

Nesta secção, o objetivo é reduzir o número de *features* de forma a melhorar ou tornar mais fiáveis os valores obtidos pelos modelos preditivos. Do ponto de vista prático e, até porque, esse não é o objetivo primordial deste trabalho, é impossível realizar uma análise abrangente sobre todos os métodos de seleção de *features* existentes. Para tal, foram considerados dois métodos de seleção de *features* comuns que indicam se as mais adequadas ao estudo, tanto pela robustez e eficácia, como pelo modo de implementação. Neste seguimento, escolheu-se um método *filter* univariado (independente do classificador utilizado) - Testes Estatísticos, e um método *embedded* multivariado - SVM-RFE. A escolha dos algoritmos não pretende indicar que os métodos excluídos deste estudo, por exemplo, os métodos *wrappers*, são inferiores aos utilizados; em vez disso, escolheram-se os que se acreditam serem mais adequados para a seleção de *features* da base de dados em análise. As descrições dos métodos de seleção de *features* referidos são fornecidas na secção 2.5.

Tendo em conta a base de dados em estudo, resolveu-se reduzir a dimensão da mesma dado que, existiam 74 *features* para 201 registos. Esta desproporcionalidade poderia provocar *overfitting* dos dados e os *outliers* poderiam tornar-se mais significativos. Para determinar o número de *features* a fornecer ao classificador, considerou-se adequada a proporção de uma *feature* para cada cinco amostras. Desta forma, através dos métodos implementados para a seleção de *features* identificaram-se as 40 *features* que melhor previam os grupos em estudo (ASD vs não-ASD).

Dado que o método SVM-RFE adota uma abordagem iterativa, e conseqüentemente, possui um custo computacional relativamente elevado, optou-se por um esquema de seleção de *features* de duas fases, para permitir que a simulação seja concluída num tempo razoável e de forma eficiente. Para além disso, e dado que os Testes Estatísticos não obtêm um número específicos de *features* mas sim as que são significativamente dependentes das classes, o número de *features* resultante pode, ainda assim, ser relativamente elevado e para tal é necessário proceder a uma nova seleção. Concluindo, realizou-se um método de seleção de *features* de duas fases distintas para reduzir o número de *features* e, possivelmente, encontrar um subconjunto mais preditivo, comparando com a aplicação direta de um algoritmo.

No primeiro método implementado, utilizou-se um algoritmo de seleção de *features* independente do classificador, para remover a maioria das *features* não informativas. No segundo método, foi aplicado um algoritmo de seleção de *features*, para

melhorar ainda mais o conjunto de *features* do primeiro método. Em suma, o grupo de *features* inicial é filtrado e, conseqüentemente, reduzido para um tamanho intermédio sendo que, após a aplicação do segundo método, surge um subconjunto de *features* mais criterioso de tamanho designado.

Testes Estatísticos

A aplicação dos Testes Estatísticos ao conjunto de *features* iniciais foi efetuada de acordo com o procedimento apresentado na figura 3.3. Em todos os testes realizados, o nível de significância adotado foi de 0,05.

as diferenças devem ser avaliadas entre os dois grupos que compõe as classes. Ou seja, verificar se a feateare A é estatisticamente diferente entre o grupo com autismo e sem autiso.

Inicialmente, na implementação dos Testes Estatísticos para avaliação da dependência das *features* com as classes, procedeu-se à avaliação da normalidade das *features* quantitativas. Essa análise foi realizada recorrendo ao teste de *Shapiro Wilk*. A avaliação da significância das diferenças entre os dois grupos que compõe as classes, ou seja, verificar se a *feature* A é estatisticamente diferente entre o grupo com ASD e não-ASD foi efetuada através do teste *t – Student* para *features* com distribuição normal e do teste *Mann – Whitney* para as *features* que não seguiam uma distribuição normal. O teste de *Levene* é implementado para avaliar as variâncias.

A associação entre as *features* qualitativas e os dois grupos foi avaliada com recurso ao teste de Fisher.

SVM-RFE

Antes de se proceder à implementação do SVM-RFE, foi necessário efetuar a normalização das *features*. O mesmo não foi necessário para realizar os Testes Estatísticos, dado que neste método a validação das *features* é individual, ao contrário do SVM-RFE que analisa o seu conjunto. Como algumas das *features* exibiam faixas de valores muito diferentes, um procedimento de normalização foi necessário para garantir a invariância da escala. Dessa forma, garantiu-se uma execução adequada dos classificadores, pois eliminou-se o risco de uma *feature* com maior faixa de valores ser dominante perante as outras. A normalização das *features* foi aplicada através da

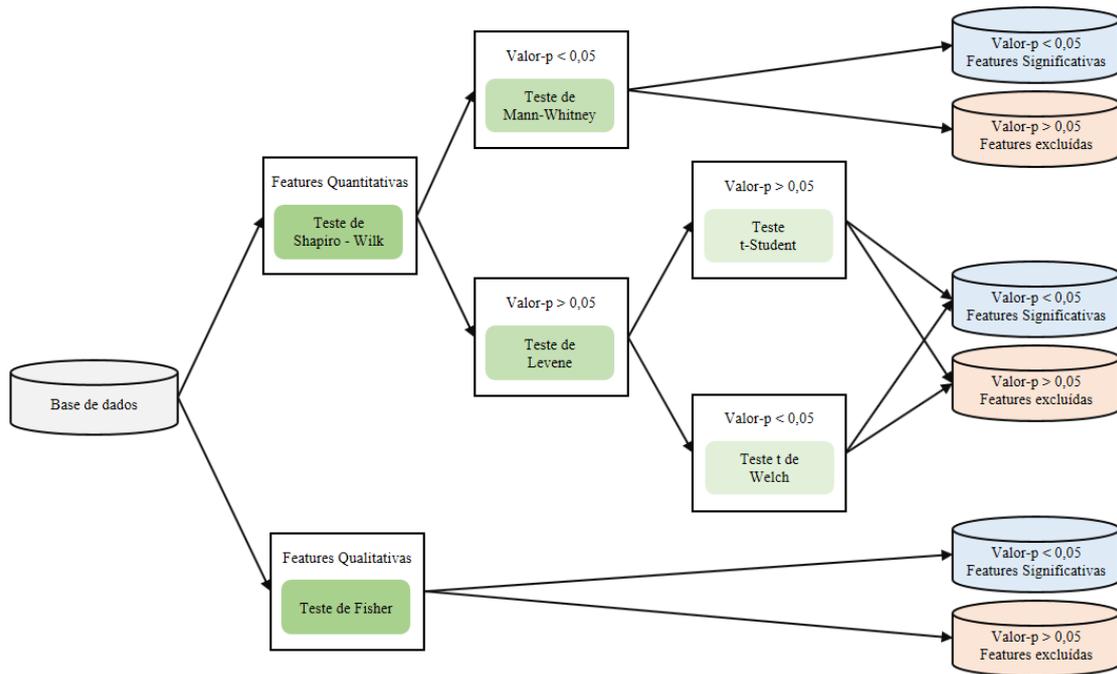


Figura 3.3: Procedimento executado para a implementação dos Testes Estatísticos à base de dados.

Equação 3.1.

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\gamma_j}, \quad (3.1)$$

onde x_{ij} representa o valor da j -ésima *feature* da base de dados, e μ_j e γ_j correspondem à média e desvio padrão da *feature* j , respetivamente. Depois de obter os recursos normalizados, procedeu-se à implementação do método SVM-RFE.

Após filtrar as *features* não informativas, o SVM-RFE foi utilizado para selecionar as *features* mais discriminativas. Primeiro, o modelo SVM foi construído no subconjunto de *features* atuais, obtendo-se o peso de cada *feature*, pelo que, as que apresentem menor peso são removidas. Em cada iteração, o subconjunto de *features* em questão é avaliado pela validação cruzada k -fold. Aquele com a maior precisão de validação cruzada é mantido como subconjunto final de *features* selecionadas. Computacionalmente, o modelo SVM foi implementado a partir da função $SVR(kernel = "linear")$, da biblioteca `sklearn.svm` enquanto o método de seleção RFE foi implementado pela função `RFE`, pertencente à biblioteca `sklearn.feature_selection`. O último apresentou o *ranking* de acordo com a importância de cada *feature*.

O método de seleção de *features* proposto foi testado e analisado fazendo a comparação dos classificadores obtidos através de diferentes bases de dados. Essas bases de dados foram obtidas por quatro metodologias diferentes, no que concerne à seleção de *features*: executando a seleção pelos Testes Estatísticos seguida de SVM-RFE; apenas pelos Testes Estatísticos; apenas pelo SVM-RFE, finalmente, não aplicando qualquer método de seleção de *features*, utilizando a base de dados original. Para que a implementação (ou não) da imputação não fosse um fator variável nesta avaliação, concatenaram-se os dados obtidos por cada método de seleção de *features* com e sem o método de imputação. Desta forma, compararam-se os valores de AUC obtidos para cada classificador e para cada modelo de seleção de *features*.

3.3 Balanceamento

Com a análise da base de dados utilizada, é evidente a existência de dados não balanceados, ou seja, a presença de uma classe majoritária com frequência muito maior que a outra classe. Num total de 201 observações, 135 correspondem à classe ASD e as restantes 66 apresentam um diagnóstico não-ASD. Portanto, é claramente um conjunto de dados desequilibrado e a proporção de registos da classe ASD para a classe não-ASD é de 3:1, aproximadamente.

As classes desequilibradas podem produzir um classificador com maior tendência para responder melhor à classe majoritária, em detrimento da minoritária.

A solução que melhor se enquadrava na base de dados em estudo, para diminuir os efeitos da desproporção de classes, foi o *oversampling* aleatório, ou seja, a replicação dos registos da classe minoritária (não-ASD). Esta replicação foi executada de forma aleatória nas amostras da classe não-ASD. Desta forma, a base de dados resultante apresentava 270 amostras, sendo 135 correspondentes a cada classe.

Também na classe majoritária procedeu-se ao *bootstrapping* das amostras, ou seja, de todos os casos correspondentes à classe em maior número, selecionou-se aleatoriamente um número igual de amostras. Esta técnica permitiu incrementar o peso de algumas das amostras da classe majoritária, tal como se sucede para a classe minoritária.

Embora esta técnica possa aumentar a probabilidade de *overfitting*, uma vez que replica os eventos da classe minoritária, não provoca a perda de informação e fornece mais dados ao classificador.

3.4 Classificação

Para avaliar o desempenho da classificação, foram considerados 70% dos dados para treino, e 30% para teste, perfazendo um total de 189 indivíduos para o primeiro grupo e 81 indivíduos para o segundo grupo.

A classificação definiu a última fase da metodologia, tendo a tarefa complexa de discriminar entre diagnóstico de ASD positivo ou diagnóstico de ASD negativo, usando as *features* obtidas pelos métodos de seleção. Dito isto, o principal objetivo é fornecer classificadores que permitam um bom diagnóstico da doença, de forma a tornar o processo mais rápido e menos dispendioso. Neste estudo, decidiu-se comparar três tipos de classificadores, *Support Vector Machine*, *Random Forest* e *Logistic Regression*. Uma breve descrição de cada método de classificação e a forma como foram implementados é apresentada abaixo.

Support Vector Machine

O *Support Vector Machine* baseia-se na construção de um hiperplano que define um limite de decisão. Por outras palavras, para um conjunto de dados identificados, o SVM gera um hiperplano ideal que maximiza a margem de separação e que classificará as novas amostras (dados do grupo de teste). Quando o SVM, através de uma separação linear, não permite a classificação, é possível implementar um *kernel* que mapeia as amostras do treino num espaço dimensional maior e, desta forma, aprende a separar as amostras nesse mesmo espaço. É esta adaptação do *kernel* que torna o SVM tão comum e preciso, uma vez que tem uma grande capacidade de generalização, independentemente da distribuição das amostras [117].

Para a criação do modelo preditivo, utilizou-se o SVM *classification* (*svm.SVC()*) com um *kernel* linear e considerando um valor de $C = 1$, implementado usando o *scikit-learn*.

Random Forest

Resumidamente, o algoritmo de *Random Forest* envolve a criação aleatória de subconjuntos de dados e a construção de uma árvore de decisão com base nessas amostras. Em cada nó da árvore, *mtry* (parâmetro correspondente ao número de *features*) é selecionado no conjunto de todas as *features*. A *feature* que fornece a melhor divisão (considerando quaisquer nós anteriores) é escolhida e, em seguida, o procedimento é repetido. Esse algoritmo é executado num grande número de

árvores, com base em diferentes subconjuntos de amostras, o que significa que esse método é menos propenso a *overfitting* do que outros [111].

O RFs foi implementado com recurso ao algoritmo *RandomForestClassifier()* da biblioteca *scikit-learn*, com o número de estimativas igual a 100.

Logistic Regression

A *Logistic Regression* é um classificador que utiliza o método da máxima verossimilhança e que permite estimar a probabilidade associada à ocorrência de um determinado evento, face a um conjunto de features explanatórias.

Utilizando o Python, neste estudo a regressão logística foi implementada através do algoritmo *LogisticRegression()*, pertencente à biblioteca *sklearn.linear_model*. Os parâmetros utilizados foram os *default*.

Os resultados dos modelos preditivos foram obtidos de acordo com o Teorema do Limite Central. Este teorema refere que a distribuição de médias amostrais pode ser aproximada a uma distribuição normal, à medida que o tamanho das amostras aumenta. Esta aproximação não depende da forma da distribuição da população.

Neste estudo, foram utilizadas 100 repetições, isto é, foram obtidos 100 grupos de teste e treino aleatórios, sendo que para cada grupo de treino os classificadores foram treinados e para cada grupo de teste respetivo foi obtido o valor de AUC. Seguidamente, procedeu-se à média dos valores obtidos para cada repetição (média amostral).

3.5 Métricas de avaliação

Area Under the ROC Curve (Área abaixo da curva ROC)

A análise da curva Receiver Operating Characteristic é uma ferramenta poderosa para medir e especificar problemas no desempenho do diagnóstico médico. Esta análise realizada através de um método gráfico simples e robusto, permite estudar a variação da sensibilidade e especificidade, para diferentes valores de corte [118].

Uma curva ROC é a representação gráfica dos pares sensibilidade e $1 - \text{especificidade}$, resultantes da variação do valor de corte ao longo de um eixo de decisão, x .

Como mencionado por vários autores [119, 120], a *Area Under the ROC Curve* (Área

abaixo da curva ROC) tem sido amplamente utilizada para medir o desempenho de modelos de classificação. A curva ROC é utilizada para fornecer uma ideia geral do potencial preditivo do classificador, dado que caracteriza a relação entre o número de verdadeiros positivos e o número de falsos positivos. Um valor elevado de AUC indica que o classificador, em todo o intervalo de decisão possível, tem um número de verdadeiros positivos mais alto. Foi demonstrado que a AUC tem menor variância e é mais confiável do que outras métricas de desempenho (como precisão e medida F) [121].

Diz-se que uma curva, ROC A, domina outra curva, ROC B, se A está sempre acima e à esquerda de B. Nos casos em que duas curvas ROC não se dominam uma à outra em toda a faixa, ou quando os custos de distribuição e erro de classe são desconhecidos, a área sob a curva ROC (AUC) é um bom “resumo” para comparação destes. A AUC toma valores entre 0 e 1, onde 0,5 denota a pior classificação possível e 1 ou 0 denotam a classificação perfeita.

A área abaixo da curva ROC está associada ao poder discriminante de um teste de diagnóstico.

Intervalo de Confiança

O intervalo de confiança é uma métrica de avaliação obtido para todos os resultados. Esta medida permite avaliar a probabilidade com a qual o intervalo observado contém o parâmetro real de interesse.

Neste estudo, o intervalo de confiança será aplicado à AUC e, portanto, um intervalo de confiança de 95% para a média da AUC significa, com 95% de probabilidade que o verdadeiro valor de AUC estar incluído na gama de valores que constituem esse mesmo intervalo. Desta forma, considerou-se pertinente utilizar este parâmetro para avaliação dos resultados obtidos.

Exatidão, sensibilidade e especificidade

Para realizar uma avaliação mais extensiva de cada classificador, foram aplicadas 3 métricas distintas, para além da AUC: sensibilidade, especificidade e exatidão. O objetivo foi analisar a matriz de confusão de cada classificador, compreendendo assim as discrepâncias entre os resultados da classificação e o resultado conhecido.

As definições das métricas usadas são apresentadas em baixo:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}, \quad (3.2)$$

$$\text{Especificidade} = \frac{VN}{FP + VN}, \quad (3.3)$$

$$\text{Exatidão} = \frac{VP + VN}{VP + FN + FP + VN}, \quad (3.4)$$

Onde VP (verdadeiro positivo) representa as instâncias positivas corretamente diagnosticadas, e VN (verdadeiro negativo) representa as instâncias negativas corretamente diagnosticadas. FP (falso positivo), representa as instâncias positivas classificadas incorretamente, enquanto FN (falso negativo) representa as instâncias negativas mal classificadas.

A análise destes três parâmetros foi apenas executada para a melhor metodologia de imputação e de seleção de *features*.

3.6 Análise de amostras

Tendo em consideração a clara distinção entre as classes verificada anteriormente, considerou-se relevante analisar o motivo pelo qual os modelos preditivos aplicados não apresentam ainda melhor desempenho. Para tal, avaliou-se a taxa de erro associada à classificação de cada indivíduo, tendo sido obtido para cada um deles, o número de vezes em que o algoritmo efetuou um diagnóstico erróneo. Neste seguimento, procedeu-se à implementação do classificador para 500 grupos de teste e treino, gerados aleatoriamente, com o intuito de se comparar os resultados obtidos por este, com o diagnóstico clínico de cada indivíduo, sendo apenas contabilizados os resultados díspares.

Para as amostras com mais de 10% de falhas no diagnóstico realizou-se uma análise detalhada, por forma a justificar esse resultado, que pode ser dividida em duas partes. Primeiramente, por forma a avaliar o seu impacto analisou-se o desempenho do classificador *Random Forest* na sua ausência. Que se seguiu por uma verificação pormenorizada das *features* qualitativas, mais concretamente a falha ou não na resposta dos itens do M-CHAT e do SCQ por parte dos indivíduos com elevada taxa

de erro.

3.7 *Ranking* dos itens do M-CHAT

Tendo em conta a revisão bibliográfica efetuada, verificou-se que uma análise detalhada sobre os itens do teste M-CHAT é comum e genérica quando testado para populações de diferentes países, culturas ou faixas etárias. Dado que o estudo efetuado neste trabalho é exclusivo em Portugal considerou-se importante proceder à mesma análise. Por conseguinte, de acordo com os resultados obtidos do teste M-CHAT para todos os indivíduos participantes do estudo obteve-se o *ranking* dos respetivos itens.

O *ranking* dos itens do M-CHAT que melhor caracterizam os portadores de ASD foi adquirido através do método SVM-RFE e comparado com os *rankings* obtidos na pesquisa bibliográfica [41, 57, 54, 58].

No final, chegou-se à conclusão de quais os melhores procedimentos para discriminar crianças com Transtorno do Espectro do Autismo de forma mais eficiente.

Resultados e Discussão

Neste capítulo, será feita uma análise da base de dados utilizada, serão apresentados e discutidos os resultados obtidos nos modelos preditivos executados e, serão ainda apresentadas informações que se consideram pertinentes em cada uma das três secções descritas no capítulo 3 [92]:

Imputação de *missings* : analisar o impacto da imputação no modelo preditivo e do método selecionado para a sua execução (regressão linear e moda);

Seleção de *features*: analisar o seu impacto no modelo preditivo e a influência dos Testes Estatísticos e do SVM-RFE na sua implementação. Para além disso, pretende-se apresentar uma análise objetiva sobre a validade das *features* selecionadas;

Classificação: analisar o desempenho dos métodos preditivos gerados, comparativamente à literatura da área.

4.1 Base de dados

Os participantes foram 201 crianças (168, equivalente a 83,6%, do sexo masculino), com menos de quatro anos de idade, referenciadas devido à suspeita de problemas no desenvolvimento neurológico. Os participantes foram vistos como parte de um estudo em ambulatório especializado em transtornos do neurodesenvolvimento, especialmente ASD, entre 2009 e 2015.

Durante a avaliação, todas as crianças foram testadas para ASD usando os testes MCHAT e SCQ. Foi também aplicada uma extensa avaliação, clínica e funcional, por uma equipa multidisciplinar especializada coordenada por um pediatra do neurodesenvolvimento, para o diagnóstico de ASD e outros distúrbios do desenvolvimento neurológico. A avaliação incluiu uma entrevista abrangente aos pais, contendo

questões sobre informações básicas, como a idade em que a criança realizou pela primeira vez marcos relevantes do desenvolvimento neurológico (primeiras palavras, primeiras frases, idade em que começou a caminhar, controle diurno e noturno do esfíncter da bexiga) e áreas das primeiras e principais preocupações (descrições e idade do desenvolvimento e comportamento estranhos da criança). Além disso, foram ainda aplicados instrumentos projetados e realizados especificamente por especialistas para o diagnóstico de ASD, o *Autism Diagnostic Interview - Revised* e o *Autism Diagnostic Observation Schedule* (ADOS).

O diagnóstico final foi efetuado com base nos instrumentos padrão: o ADI-R, o ADOS e o exame clínico realizado pela equipa multidisciplinar, de acordo com os critérios expostos no *Diagnostic and Statistical Manual of Mental Disorders-V*. Tendo em conta este procedimento, cientificamente comprovado como sendo a forma mais estável e confiável de diagnosticar o ASD, subdividiram-se as crianças em dois grupos diagnósticos: ASD *versus* não-ASD.

No decorrer da avaliação efetuada, foram recolhidos alguns dados que poderiam complementar o estudo. Desta forma, na tabela 4.1 resumem-se os dados demográficos e as respostas a questões relevantes do neurodesenvolvimento das crianças participantes no estudo, para toda a amostra e para cada grupo de diagnóstico (ASD *versus* não-ASD). O grupo de diagnóstico ASD (N = 135) consistiu em 117 crianças do sexo masculino (86,7%) e 18 crianças do sexo feminino e a idade mediana da avaliação efetuada foi de 3 anos e 6 meses. O grupo não-ASD (N = 66) consistiu de 51 meninos (77,3%) e 15 meninas, com uma idade mediana da avaliação efetuada de 3 anos e 4 meses.

No âmbito dos dados demográficos, foram ainda registados os dados da literacia dos pais das crianças em estudo, tendo sido categorizados como apresentados na tabela 4.2.

Da base de dados utilizada constam ainda as respostas dos pais aos inquéritos de triagem de ASD, M-CHAT e SCQ. Os inquéritos fornecidos encontravam-se traduzidos para a língua portuguesa e são amplamente recomendados pelo Programa Nacional de Saúde Infantil e Juventude, em Portugal (Anexos A e B).

Desta forma, a base de dados utilizada, é composta por 201 amostras e 75 *features*. Cada amostra corresponde aos dados relativos a uma criança suspeita de possuir ASD. Das 75 *features*, 23 delas correspondem à avaliação das respostas dos pais, falha ou não falha aos itens do inquérito M-CHAT, 40 avaliação das respostas dos

Tabela 4.1: Dados demográficos e ocorrência dos principais marcos do neurodesenvolvimento dos participantes do estudo, para toda a amostra e para cada grupo de diagnóstico (ASD *versus* não-ASD).

Características	Total	ASD	Não-ASD
	(N=201)	(N=135)	(N=66)
Género			
n°. masculino; n°. feminino (%)	168; 33 (84; 16)	117; 18 (87; 13)	51; 15 (77; 23)
Idade na avaliação* (dias)	1285 (1031; 1537)	1350 (1141; 1608)	1250 (1010; 1457)
História familiar positiva			
n°. sim; n°. não (%)	75; 119 (37; 59)	46; 83 (36; 64)	29; 36 (45; 55)
Primeiras palavras* (meses)	24 (18; 30)	24 (18; 30)	18 (18; 24)
Primeiras frases* (meses)	36 (36; 48)	36 (30; 36)	40 (36; 48)
Primeiros passos* (meses)	14 (12; 16)	14 (12; 16)	14 (12; 15)
Idade materna* (anos)	31 (28; 34)	32 (28; 34)	31 (28; 34)
Idade paterna* (anos)	33 (30; 36)	33 (31; 36)	32 (30; 37)
Primeiros passos* (meses)	14 (12; 16)	14 (12; 16)	14 (12; 15)
Controlo diurno do esfíncter da bexiga* (meses)	36 (31; 48)	36 (36; 48)	36 (24; 36)
Controlo noturno do esfíncter da bexiga* (meses)	41 (36; 48)	48 (36; 48)	36 (36; 42)

*Dados apresentados da seguinte forma: mediana (percentil-25; percentil-75).
Determinadas *features* apresentadas podem possuir alguns dados em falta.

Tabela 4.2: Literacia dos pais para cada grupo de diagnóstico (ASD *versus* não-ASD).

ID	Escolaridade	Literacia materna		Literacia paterna	
		ASD (N=135)	Não-ASD (N=66)	ASD (N=135)	Não-ASD (N=66)
1	Até ao 9ºano	34 (25)	31 (47)	51 (38)	41 (62)
2	Entre o 10ºe o 12ºano	55 (41)	21 (32)	47 (35)	13 (20)
3	Licenciatura	41 (31)	12 (18)	27 (20)	11 (17)
4	Mestrado	1 (1)	0 (0)	2 (2)	0 (0)
5	Doutoramento e Bacharelato	2 (1)	2 (3)	2 (2)	1 (1)
6	Curso técnico/profissional e Pós-graduação	1 (1)	0 (0)	4 (3)	0 (0)

*Dados apresentados da seguinte forma: frequência absoluta (frequência relativa em %).
Os valores relativos à literacia parental apresentavam alguns dados em falta.

país, falha ou não falha aos itens do inquérito SCQ sendo as restantes 12 *features* correspondentes aos dados apresentados nas tabelas 4.1 e 4.2.

4.2 Avaliação

Os resultados de AUC apresentados na tabela 4.3 foram obtidos após a implementação dos classificadores, de forma a avaliar o seu desempenho. Para cada um dos classificadores utilizados, fizeram-se alterações à base de dados introduzida por forma a avaliar o seu efeito.

Na imputação de *missings*, testou-se a influência que teria a sua implementação, sendo esta realizada recorrendo ao método de regressão linear para as *features* quantitativas e à moda para as *features* qualitativas. Para termo de comparação, foi obtido o valor de AUC para a base de dados primária, sem valores imputados.

De forma a testar os métodos de seleção de *features*, os dois métodos selecionados foram conjugados de diferentes formas: implementação dos Testes Estatísticos seguido do método SVM-RFE, implementação isolada dos Testes Estatísticos, implementação isolada do SVM-RFE e, ainda, teste dos modelos preditivos para a base de dados com as *features* originais (apenas se retiraram as *features* com mais de 30% de *missings*).

As duas alternativas analisadas na imputação de *missings*, associadas com os quatro grupos de *features* obtidos na seleção de *features* e tendo em conta os três classificadores implementados, resultaram num conjunto de 24 procedimentos diferentes. Para cada procedimento, obteve-se o valor de AUC após a implementação do classificador em 100 grupos de teste e treino diferentes, gerados aleatoriamente. Seguidamente, calculou-se o valor médio e o respetivo intervalo de confiança a 95%.

4.3 Imputação de *missings*

Inicialmente, analisou-se a quantidade de *missings* em cada *feature*. As *features* em que este valor ultrapassava os 30% foram eliminadas.

Como observado no gráfico da figura 4.1, a base de dados original possuía 14 *features* com *missings* . Ao analisar a quantidade de *missings* correspondente a cada uma delas, constata-se que 6 *features* (SCQ1, SCQ2, SCQ3, SCQ4, SCQ5, SCQ6 e SCQ7) têm mais de metade dos dados omissos e, portanto, um valor superior aos 30% mencionados. Por consequência, essas mesmas *features* foram eliminadas

Tabela 4.3: Valores médios de AUC e respetivos intervalos de confiança para os classificadores *Support Vector Machine*, *Random Forest* e *Logistic Regression*, tendo em conta os diversos métodos de seleção de *features* e o método de imputação de *missings*. Os dados são apresentados da seguinte forma: média AUC (intervalo de confiança).

	Método de Seleção de <i>Features</i>	Classificador		
		SVM	RF	LR
Imputação: Regressão Linear + Moda	Não aplicado	0,945 (0,940; 0,950)	0,952 (0,947; 0,957)	0,936 (0,930; 0,942)
	Teste-t	0,946 (0,940; 0,952)	0,955 (0,950; 0,960)	0,933 (0,926; 0,940)
	SVM-RFE	0,934 (0,928; 0,940)	0,958 (0,953; 0,963)	0,929 (0,923; 0,935)
	Teste-t + SVM-RFE	0,932 (0,926; 0,938)	0,956 (0,951; 0,961)	0,924 (0,918; 0,93)
Imputação não aplicada	Não aplicado	0,919 (0,910; 0,928)	0,941 (0,934; 0,948)	0,920 (0,912; 0,928)
	Teste-t	0,926 (0,917; 0,935)	0,942 (0,934; 0,950)	0,916 (0,907; 0,925)
	SVM-RFE	0,909 (0,900; 0,918)	0,936 (0,928; 0,944)	0,909 (0,900; 0,918)
	Teste-t + SVM-RFE	0,929 (0,921; 0,937)	0,947 (0,940; 0,954)	0,930 (0,921; 0,939)

da base de dados e em situação alguma foram consideradas, nomeadamente, nos resultados obtidos para os casos sem implementação de métodos de imputação.

As *features* referidas anteriormente correspondem às questões do teste SCQ relativas à verbalização dos indivíduos em estudo, ou seja, caso a criança não falasse os dados destas *features* eram considerados *missings*. Neste seguimento, podem-se considerar *missings* não aleatórios e que justificam a sua causa.

Desta forma, os métodos de imputação foram aplicados em 8 *features*, num total de 113 *missings*. Destes, 103 apresentam-se em *features* quantitativas (idade paterna, primeiras palavras e frases e controlo do esfíncter da bexiga noturno e diurno) e 10 em *features* qualitativas (literacia do pai e da mãe e história familiar).

Tal como referido na secção 3.1, foram aplicados métodos de imputação diferentes de

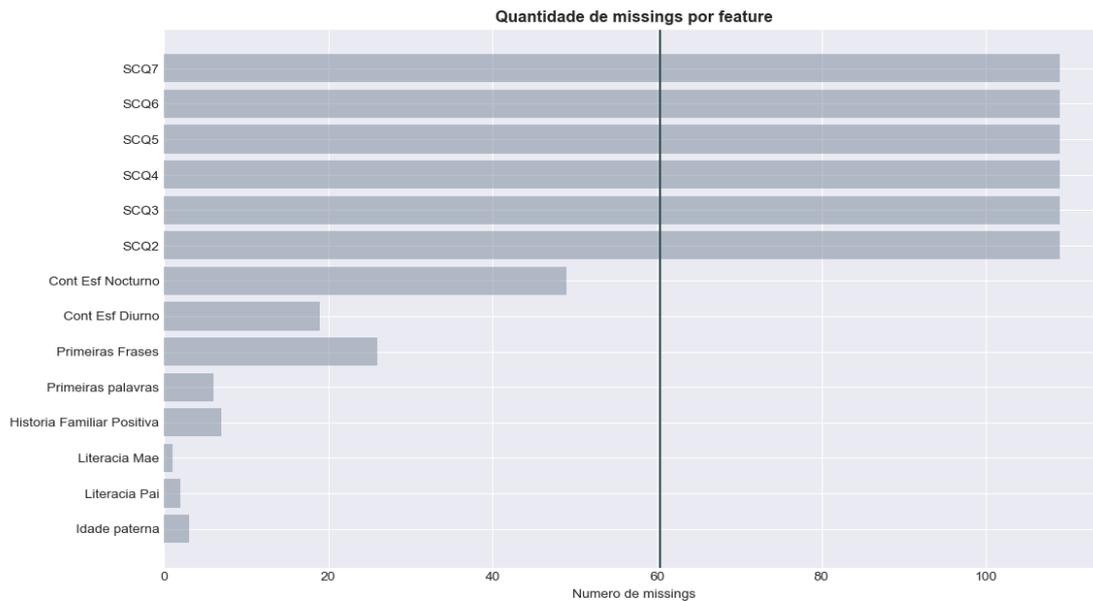


Figura 4.1: Número de *missings* por *feature* presentes na base de dados original. A linha vertical traçada identifica o valor correspondente a 30% de *missings* .

acordo com o tipo de *feature* a que se destinavam. Para as *features* quantitativas, os *missings* foram substituídos com base num modelo de Regressão Linear criado a partir das três *features* mais correlacionadas. No caso dos *missings* pertencerem a *features* qualitativas, o valor da moda foi utilizado na substituição.

Features quantitativas - Regressão Linear

Primeiramente, para definir as *features* a utilizar na implementação da Regressão Linear, procedeu-se à criação da matriz de correlação de *features* (figura 4.2) e identificaram-se as três *features* com maior valor de correlação relativamente à *feature* em que se pretende efetuar a imputação. Seguidamente, criou-se o modelo de regressão linear a partir das *features* identificadas, treinou-se com os dados existentes na *feature* em análise e aplicou-se aos dados em falta. Desta forma, geraram-se valores preditivos para substituição dos *missings* .

Features qualitativas - Moda

No caso das *features* qualitativas, o processo implementado é bastante simples mas, tendo em conta que apenas seriam imputados 10 *missings* em três *features*, considerou-se que não seria necessário investir num método computacionalmente mais exigente e em que o resultado final não sofreria uma alteração significativa.

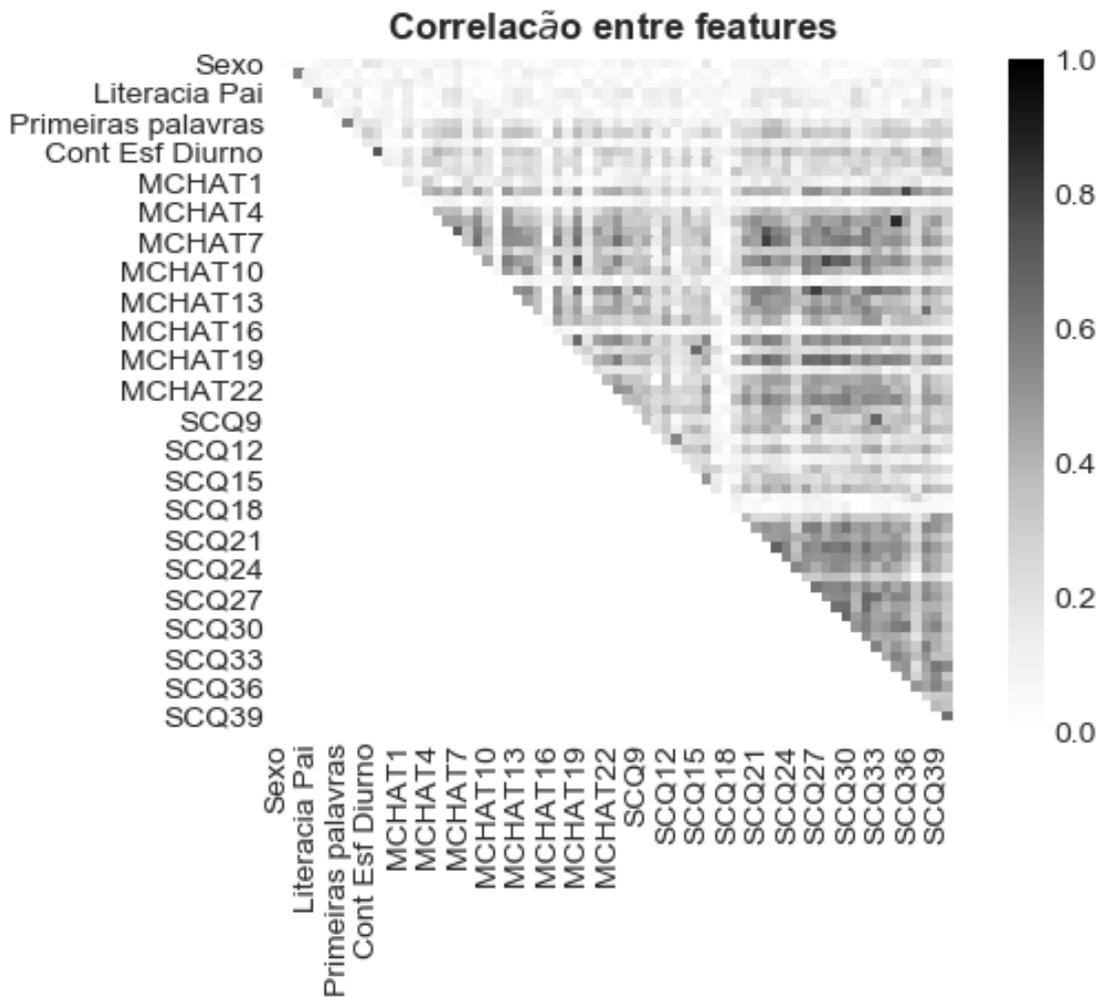


Figura 4.2: *Heatmap* da correlação entre *features*. De referir que os eixos não apresentam os nomes de todas as *features*.

Neste processo de imputação, isolaram-se os valores da *feature* com *missings* e obteve-se o respetivo valor da moda. Os valores omissos dessa *feature* são então substituídos pelo valor da moda.

4.3.1 Avaliação

Para avaliar o modelo de imputação apresentado, os resultados dos classificadores obtidos pela base de dados resultante foram comparados com os resultados da base de dados original. Pelos motivos apresentados em 3.1, para esta última eliminaram-se as amostras que continham *missings*. Assim sendo, a base de dados resultante da imputação continha 201 amostras e 68 *features*, com 113 valores preditos, enquanto a base de dados comparativa continha 164 amostras e 68 *features*.

Tendo em conta os resultados apresentados na tabela 4.3 e, de forma a avaliar exclusivamente a imputação executada, foram utilizados todos os valores de AUC obtidos para cada classificador e para todos os métodos de seleção de *features*. Assim, gerou-se cada *boxplot* comparativo utilizando 400 valores de AUC provenientes das 100 repetições executadas para cada método de seleção de *features* em cada classificador.

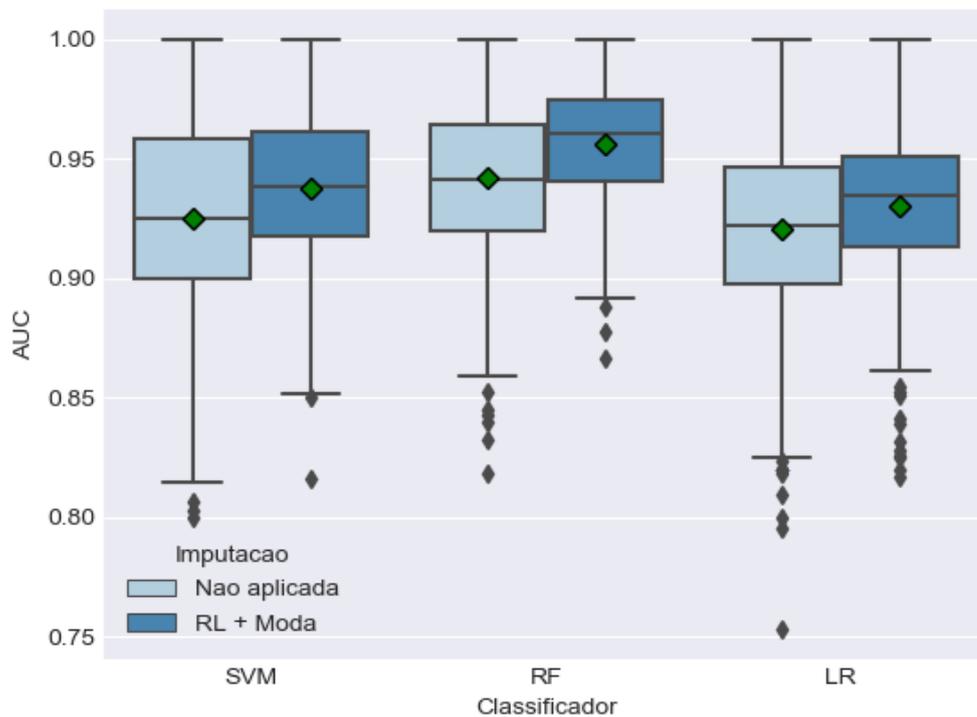


Figura 4.3: Comparação entre os classificadores gerados a partir de dados não imputados e de dados imputados através do modelo apresentado. Nota: os pontos verdes assinalados correspondem ao valor da média para cada conjunto de dados.

Observando os *boxplots* de cada classificador na figura 4.3, são evidentes os aumentos dos valores médios quando aplicado o método de imputação, comparativamente aos casos em que as amostras com *missings* foram eliminadas. Este facto pode ser sustentado pelo intervalo entre quartis já que, para todos os classificadores este intervalo diminui cerca de 3%. Para além disso, os valores medianos também corroboram as conclusões anteriores, dado que os valores da mediana aumentaram e os intervalos entre quartis diminuíram. Assim, é possível afirmar que existe um aumento de potência, como é de esperar na imputação.

Concluindo, é seguro afirmar que houve diferenças entre a implementação ou não do

método de imputação proposto — Regressão Linear para as *features* quantitativas e Moda para as *features* qualitativas.

4.4 Seleção de *features*

Tendo em conta o método de seleção de *features* proposto — implementação de Testes Estatísticos seguidos do SVM-RFE — pretendeu-se fazer uma avaliação abrangente para garantir a fiabilidade dos resultados. Neste seguimento, avaliaram-se os classificadores gerados pelas seguintes bases de dados:

- Totalidade das *features*;
- *Features* selecionadas através dos Testes Estatísticos;
- *Features* selecionadas através do SVM-RFE, considerando as 40 mais relevantes;
- *Features* selecionadas pelos Testes Estatísticos precedidas de seleção pelo SVM-RFE, considerando também as 40 mais relevantes.

Para cada uma delas, obtiveram-se os respetivos valores de AUC nos 100 diferentes conjuntos de treino e teste e com implementação ou não de imputação de *missings*. Para avaliação específica de cada base de dados e, conseqüentemente, dos métodos de seleção, criou-se o gráfico *boxplot* com os dados de AUC. Para cada um, utilizaram-se os dados obtidos com imputação concatenados com os dados dos indivíduos sem *missings* (perfazendo um total de 200 valores de AUC).

Testes Estatísticos (Teste-t)

Tendo em conta o procedimento descrito em 3.2, obtiveram-se os valores de p para cada uma das *features* (tabela 4.4). O valor-p serviu como *cutoff* e, desta forma, as *features* com valores superiores a 0,05 foram eliminadas das análises subsequentes. Neste seguimento, as *features*: sexo, historia familiar, idade paterna, idade materna, marcha, idade, MCHAT3, MCHAT16, SCQ17 e SCQ18 mostraram não ser significativamente discriminativas para os grupos em análise (ASD *versus* não-ASD) e, portanto, foram excluídas.

SVM-RFE

O método de SVM-RFE foi executado considerando duas bases de dados distintas tendo em conta o número de *features* que as compunham. Quando este método foi

4. Resultados e Discussão

aplicado isoladamente, as *features* analisadas correspondiam às originais. No caso em que a seleção de *features* foi executada primeiramente pelos Testes Estatísticos, a base de dados analisada pelo método SVM-RFE foi o resultado da seleção prévia. Por conseguinte, para esta situação apenas se obteve o *ranking* de 58 *features*, dado que as restantes tinham sido eliminadas. Para além disso, ambas foram também implementadas tanto nos dados imputados como nos dados sem imputação.

Na tabela 4.4 são apresentados os *rankings* da importância das *features* para as duas situações referidas anteriormente, e ambas com os dados imputados. Como se pode concluir na secção anterior, o método de imputação utilizado permitiu uma melhoria interessante do modelo preditivo e, portanto, na tabela 4.4 apenas foi apresentado o *ranking* do método SVM-RFE para esse caso.

Tabela 4.4: Valor-p de cada *feature* e *ranking* do SVM-RFE com e sem implementação prévia dos Testes Estatísticos. * Controlo do esfíncter da bexiga diurno/noturno.

Feature	Valor-p	Ranking SVM-RFE	Ranking SVM-RFE (após Teste-t)	Feature	Valor-p	Ranking SVM-RFE	Ranking SVM-RFE (após Teste-t)
Sexo	0,13	63	NA	MCHAT23	<0,05	7	9
Literacia Pai	<0,05	2	43	SCQ8	<0,05	42	24
Literacia Mae	<0,05	10	22	SCQ9	<0,05	23	32
Historia Familiar	0,37	60	NA	SCQ10	<0,05	5	1
Idade paterna	0,79	47	NA	SCQ11	<0,05	18	3
Idade materna	0,59	59	NA	SCQ12	<0,05	38	27
Primeiras palavras	<0,05	57	46	SCQ13	<0,05	43	36
Primeiras Frases	<0,05	1	57	SCQ14	<0,05	44	47
Marcha	0,16	4	NA	SCQ15	<0,05	27	56
Cont Esf Diurno*	<0,05	3	52	SCQ16	<0,05	22	14
Cont Esf Nocturno*	<0,05	58	48	SCQ17	0,66	67	NA
Idade	0,08	66	NA	SCQ18	1,00	65	NA
MCHAT1	<0,05	62	33	SCQ19	<0,05	52	16
MCHAT2	<0,05	33	13	SCQ20	<0,05	9	6
MCHAT3	0,16	64	NA	SCQ21	<0,05	45	34
MCHAT4	<0,05	39	49	SCQ22	<0,05	13	40
MCHAT5	<0,05	40	37	SCQ23	<0,05	16	50
MCHAT6	<0,05	46	5	SCQ24	<0,05	54	10
MCHAT7	<0,05	30	19	SCQ25	<0,05	56	25
MCHAT8	<0,05	50	11	SCQ26	<0,05	11	8
MCHAT9	<0,05	19	35	SCQ27	<0,05	25	17
MCHAT10	<0,05	32	15	SCQ28	<0,05	37	4
MCHAT11	<0,05	41	12	SCQ29	<0,05	31	39
MCHAT12	<0,05	28	53	SCQ30	<0,05	21	20
MCHAT13	<0,05	51	28	SCQ31	<0,05	26	44
MCHAT14	<0,05	24	21	SCQ32	<0,05	14	54
MCHAT15	<0,05	35	23	SCQ33	<0,05	29	18
MCHAT16	1,00	68	NA	SCQ34	<0,05	17	2
MCHAT17	<0,05	6	30	SCQ35	<0,05	34	38
MCHAT18	<0,05	48	55	SCQ36	<0,05	20	51
MCHAT19	<0,05	8	45	SCQ37	<0,05	55	42
MCHAT20	<0,05	49	41	SCQ38	<0,05	15	29
MCHAT21	<0,05	61	31	SCQ39	<0,05	36	26
MCHAT22	<0,05	12	58	SCQ40	<0,05	53	7

4.4.1 Avaliação

Tendo por base os resultados apresentados pela figura 4.4, é possível retirar algumas conclusões relativamente aos métodos utilizados para a seleção de *features*. Fazendo inicialmente uma análise dos métodos por classificador, é perceptível que para o SVM o método proposto (teste-t + SVM-RFE) não produziu os resultados expectáveis, tendo obtido o pior valor de média e mediana, bem como o maior intervalo entre quartis. Para este classificador, embora que com uma diferença mínima, os Testes Estatísticos demonstram ser o melhor método de seleção de *features*.

Para o modelo preditivo *Random Forest*, todos os métodos apresentam valores de média e mediana muito próximos. De qualquer forma, é possível afirmar que o melhor valor médio corresponde ao SVM-RFE, e a melhor mediana aos Testes Estatísticos. Comparando os intervalos entre quartis, o SVM-RFE volta a comprovar a sua validade. Todos os restantes são muito semelhantes entre si.

A utilização das *features* originais no modelo *Logistic Regression* foi um fator que se distinguiu pela obtenção de melhores resultados em todos os parâmetros (média, mediana e intervalo entre quartis).

De uma forma geral, para todos os modelos preditivos, os métodos de seleção obtêm valores de AUC médios e medianos muito próximos e com intervalos de confiança muito semelhantes. Ainda assim, é possível distinguir os Testes Estatísticos como sendo o melhor método de seleção. Mesmo que com pequenas alterações, obteve os melhores resultados no conjunto dos três classificadores. Relativamente ao método de seleção sugerido na metodologia (Testes-t + SVM-RFE), é possível aferir que foi a alternativa, de entre as quatro analisadas com menor desempenho.

4.5 Classificação

Tendo em consideração o objetivo final deste estudo, é essencial proceder-se a uma análise detalhada e rigorosa dos vários tipos de classificadores testados. Desta forma, pretende-se constatar qual o classificador que melhor prevê o diagnóstico de ASD em crianças.

Após a avaliação dos melhores métodos de imputação e de seleção de *features*, utilizaram-se os resultados obtidos por esses procedimentos para proceder à análise dos classificadores. Neste seguimento, os modelos preditivos foram criados e testados com recurso a dados imputados pela Regressão Linear e pela Moda e com *features*

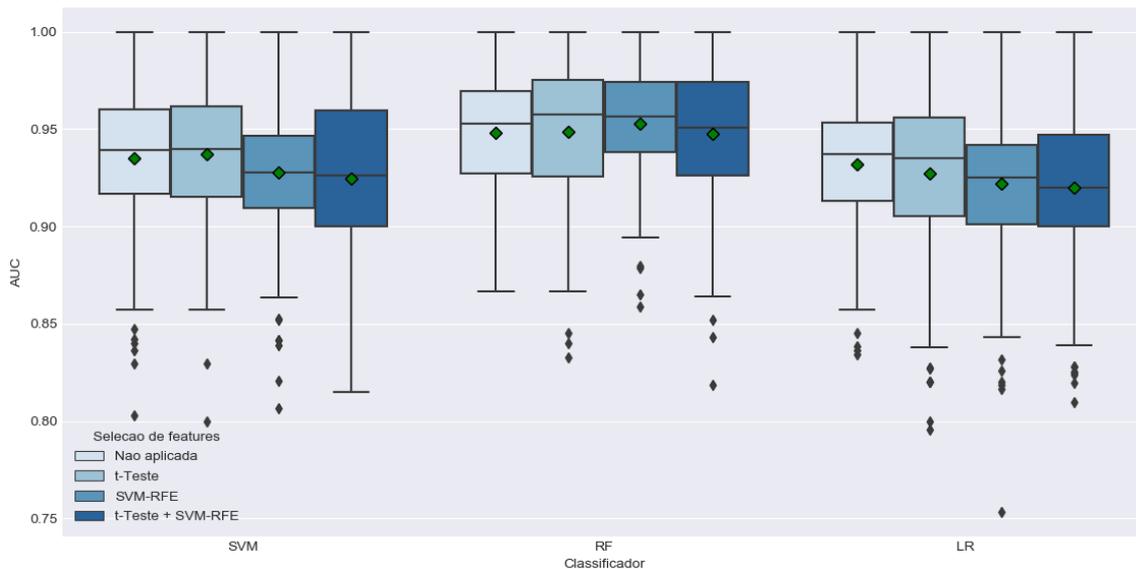


Figura 4.4: Comparação entre os modelos preditivos gerados. Nota: os pontos verdes assinalados correspondem ao valor da média para cada conjunto de dados.

selecionadas através dos Testes Estatísticos.

Para comparar de forma mais intuitiva o desempenho dos três classificadores considerados, obteve-se o gráfico apresentado na figura 4.5. Este gráfico permite aferir que o método que obteve os melhores resultados de classificação foi o *Random Forest*, comprovado tanto pelos valores da mediana como também pelos valores da média.

Após uma apreciação dos resultados de cada classificador tendo em conta todos os métodos de seleção de *features* e a execução ou não de imputação de *missings*, considerou-se necessário proceder-se a uma análise mais criteriosa dos desempenho dos classificadores para a melhor metodologia realizada (imputação de *missings* através da Regressão Linear + Moda e seleção de *features* através do SVM-RFE). Deste modo, na tabela 4.5 são apresentados os valores da média da exatidão, da sensibilidade, da especificidade e da AUC com os respetivos intervalos de confiança, para a implementação dos classificadores em 100 grupos de teste e treino diferentes.

Pela avaliação objetiva das métricas referidas identifica-se claramente que o modelo preditivo *Random Forest* obteve o melhor desempenho a distinguir a classe ASD da classe não-ASD. Com os valores aproximados de 95,7% de exatidão, 97,4% de sensibilidade, uma especificidade de 94,2% e uma AUC de 95,8% com um intervalo de confiança com uma amplitude aproximada de 0,5%, RF destaca-se evidentemente

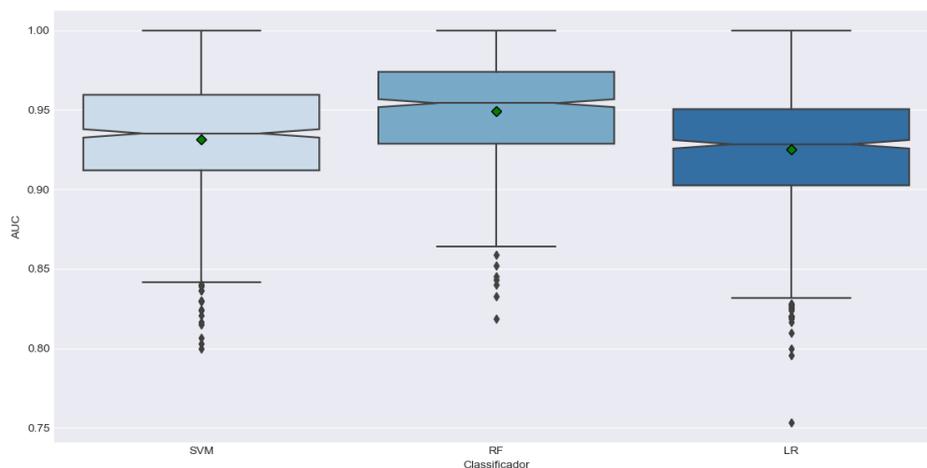


Figura 4.5: Comparação entre os classificadores gerados a partir de dados: sem seleção de *features*, com seleção de *features* efetuada pelos Testes Estatísticos, com seleção de *features* efetuada pelo SVM-RFE e com seleção de *features* efetuada com a conjugação dos Testes Estatísticos e SVM-RFE. Nota: os pontos verdes assinalados correspondem ao valor da média para cada conjunto de dados.

dos restantes classificadores.

O método SVM obteve o segundo melhor valor de AUC (93,1%) e a segunda menor amplitude do intervalo de confiança de AUC (0,6%), tal como se constata para os valores de exatidão (93,0%), sensibilidade (94,9%) e especificidade (91,2%).

O algoritmo LR registou os piores resultados de classificação de ASD em crianças. Em todos os parâmetros de avaliação obteve uma redução da percentagem entre 2,4% e os 4,5% comparativamente aos resultados do modelo RF e, no que concerne ao intervalo de confiança de AUC, um aumento de 0,2% passando a 0,7%.

Tabela 4.5: Valores das métricas de avaliação dos classificadores implementados para o conjunto de dados que sofreu imputação e seleção de *features* recorrendo ao SVM-RFE.

Métrica de avaliação	SVM	RF	LR
Exatidão	0,930	0,957	0,923
Sensibilidade	0,949	0,974	0,950
Especificidade	0,912	0,942	0,897
AUC	0,931	0,958	0,924
Intervalo de Confiança (AUC)	0,925 - 0,937	0,953 - 0,963	0,917 - 0,931

Por conseguinte, realizando uma análise mais clínica dos resultados é possível afirmar que o modelo preditivo *Random Forest* garante uma classificação correta das crianças que efetivamente são diagnosticadas com ASD em 97,4% dos casos. Por outro lado, em 94,2% dos indivíduos que não são diagnosticadas com ASD o classificador confirma esse mesmo resultado. O RF possui ainda uma garantia de 95,7% de previsões corretas, sejam eles indivíduos classificados com ASD e que efetivamente possuem ASD ou indivíduos classificados como não possuindo ASD e que efetivamente não apresentam o transtorno.

Tendo em conta os ótimos resultados de classificação alcançados, procurou-se validar os mesmos através da análise da separabilidade das classes. Assim sendo, verificou-se a distribuição de dados para cada uma das classes. Na figura 4.6, para a *label* "Grupo", os valores de 0 e 1 correspondem ao diagnóstico real de cada indivíduo (ASD e não-ASD, respetivamente) e, para as *features*, os mesmos correspondem à falha (ou não) às questões dos testes em estudo, respetivamente.

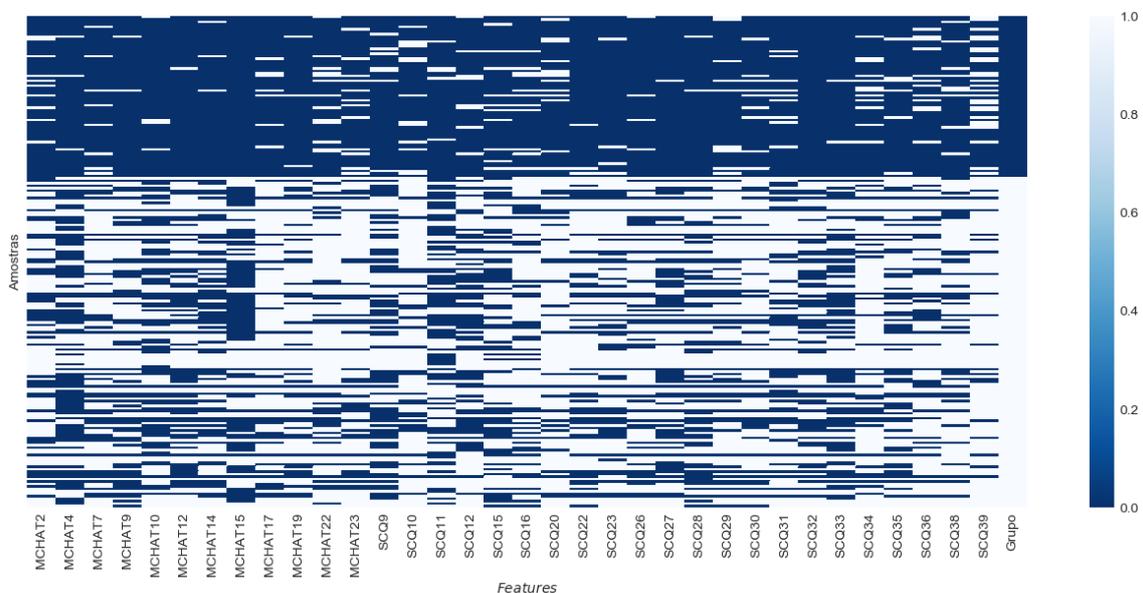


Figura 4.6: Heatmap da distribuição das respostas aos testes M-CHAT e SCQ em função do diagnóstico real de cada indivíduo (representado na última coluna).

Pela observação da figura 4.6, é possível distinguir claramente as duas classes através da distribuição de valores das *features* em análise. Isto significa que, sem qualquer tipo de classificação, e apenas pela observação direta da base de dados seria possível distinguir as classes em estudo. Desta forma, a avaliação efetuada pode justificar, em parte, os ótimos resultados alcançados pelos classificadores desenvolvidos.

4.6 Análise de indivíduos

Para que se pudesse avaliar o motivo pelo qual não se obteve um melhor desempenho do modelo preditivo procedeu-se à análise dos indivíduos em estudo. Acreditava-se que poderiam estar a existir crianças que repetidamente obtinham resultados negativos e, portanto, influenciavam negativamente os resultados gerais do modelo. Assim sendo, de forma a perceber se as crianças em que ocorriam falhas de diagnóstico era aleatória ou não, obteve-se o gráfico das percentagens relativas correspondentes à taxa de erro dos indivíduos em 500 grupos de teste e treino diferentes, ou seja, em cada iteração verificaram-se as crianças que obtinham um diagnóstico errado quando comparado com o real. Por fim, somaram-se e obtiveram-se os valores das frequências relativas de cada criança (figura 4.7).

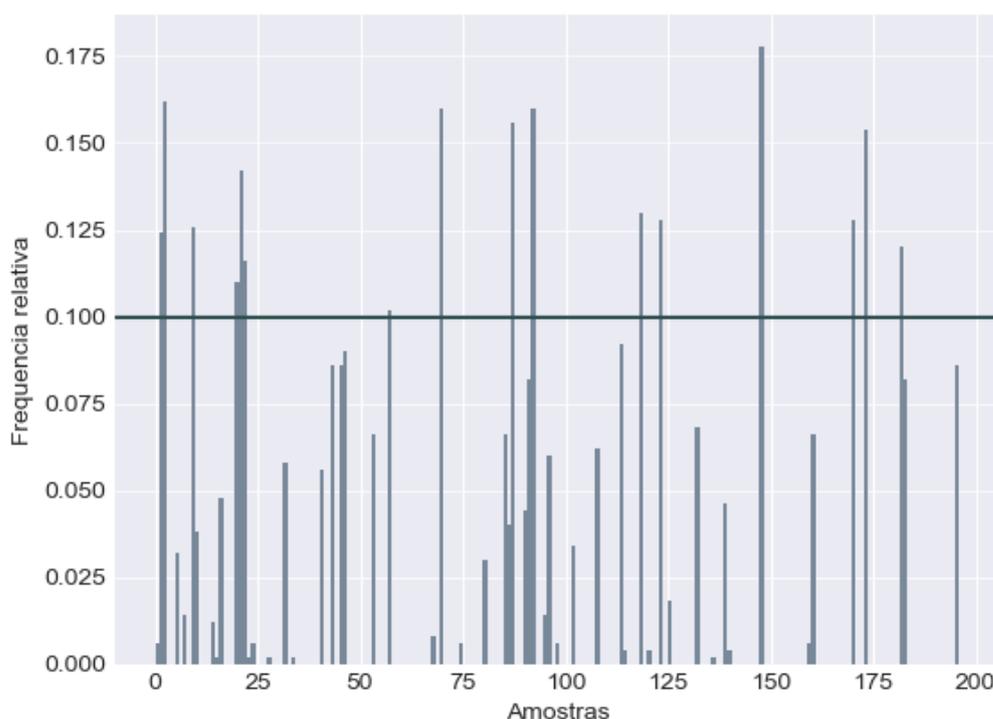


Figura 4.7: Percentagem relativa correspondente ao número de vezes que o modelo preditivo falha o diagnóstico de cada indivíduo (com 500 repetições).

Para melhor elucidação dos resultados fornecidos pelo gráfico 4.7: o indivíduo identificado pelo número 148, quando são efetuadas 500 iterações do modelo preditivo, obteve um diagnóstico errado 88 vezes, alcançando assim uma percentagem de 17,5% de ocorrência de falha de diagnóstico.

Pela observação da figura 4.7, partindo do pressuposto que esta corresponde a uma generalização dos possíveis resultados, pode-se inferir que, das 201 crianças em estudo, 16 delas obtêm um diagnóstico errado em mais de 10% dos casos. Percentagem esta que pode ser considerada significativa tendo em conta que os grupos de treino e teste são gerados aleatoriamente, o que implica que determinada criança nem sempre seja testada pelo algoritmo, podendo fazer parte do treino do classificador. Desta forma, é possível verificar que a falha de diagnóstico por parte do modelo não é aleatória e ocorre recorrentemente em determinadas crianças.

Posteriormente, seguiu-se a obtenção do valor de AUC do modelo preditivo para 100 repetições do conjunto de dados em que os indivíduos supracitados não se encontravam presentes, bem como a respetiva média e intervalo de confiança. Desta forma, contava-se avaliar o impacto que a classificação errónea destes indivíduos teria nos resultados gerais.

$$\mathbf{AUC}_{\mathbf{RF}} = 0,980$$

$$\mathbf{IC}_{\mathbf{RF}} = 0,976 - 0,984$$

Pelos valores obtidos e, comparando com os resultantes da abordagem apresentada na secção 4.5 (0,958 e [0,953; 0,963], respetivamente) pode-se concluir que os indivíduos retirados influenciavam negativamente o desempenho do classificador. O valor médio de AUC aumentou mais de 2% e a amplitude do intervalo de confiança diminuiu cerca de 0,2%, para esta nova análise. Concluindo-se que as crianças identificadas, por corresponderem a menos de 8% dos indivíduos totais, provocam um enorme peso no erro de classificação.

Para os indivíduos referidas anteriormente, tentou-se perceber se existia algum padrão que justificasse a falha no diagnóstico destes indivíduos. Assim sendo, investigaram-se as respostas que os indivíduos davam em cada pergunta/item dos testes M-CHAT e SCQ (*features*) e, conseqüentemente, se estas correspondiam a erros de resposta tendo em conta o ASD. Desta forma, procurou-se alguma conformidade entre a soma desses resultados e o diagnóstico real dos indivíduos.

Neste sentido, após conhecer quais as crianças que favoreciam um diagnóstico errado, sentiu-se necessidade de perceber se essas falhas estariam relacionadas com alguma *feature* específica. Para os itens dos testes M-CHAT e SCQ considerados relevantes analisaram-se as respostas geradas para as crianças que obtinham uma maior taxa de diagnóstico errado. Procurou-se assim, identificar um padrão nas respostas aos

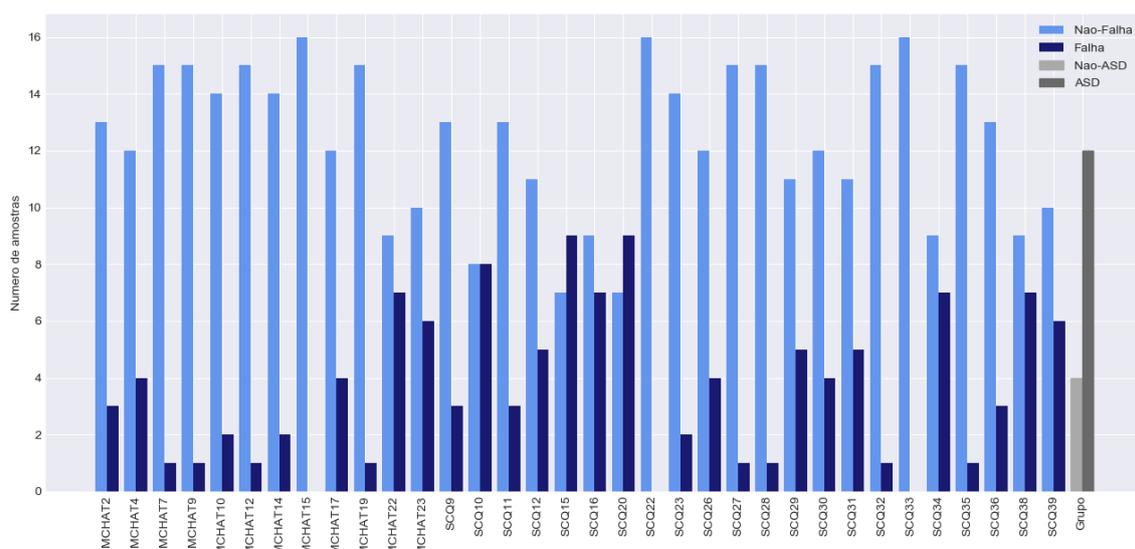


Figura 4.8: Relação entre o número de indivíduos com mais de 10% de falhas no diagnóstico, o resultado de cada item dos testes M-CHAT e SCQ e o diagnóstico real.

inquéritos de cada criança que permita afirmar a existência de uma diagnóstico errado por parte do classificador.

A figura 4.8 ilustra as respostas fornecidas para os indivíduos que obtiveram um diagnóstico erróneo e que são consideradas falhas (ou não) para a predição do ASD. Neste gráfico é também possível perceber qual das classes é mais propensa ao erro de diagnóstico.

Pela observação da figura 4.8 podemos concluir que o modelo preditivo criado tem maior propensão de falha para os indivíduos com ASD, dado que, das 16 crianças analisadas, 75% foram declaradas como não-ASD, quando na verdade possuíam o transtorno. Neste seguimento, o facto dos itens 15 do M-CHAT, 22 e 33 do SCQ não possuírem nenhuma falha nas respostas, ou seja, que dos indivíduos considerados nenhum demonstra as características implícitas nestas perguntas, pode significar que muito provavelmente estes itens estão a influenciar negativamente a classificação para os casos de ASD. Dito de outra forma, tendo em conta que os indivíduos considerados lhe têm associado um diagnóstico diferente do real, o facto de não falharem os itens referidos pode contribuir para "enganar" o classificador e produzir um diagnóstico errado.

Mais concretamente, quando a criança é portadora de ASD, ao ser analisada por este modelo preditivo pode, muito provavelmente provocar um falso diagnóstico se:

- acompanha com o olhar, caso se aponte para um brinquedo do outro lado da sala (M-CHAT15);
- aponta espontaneamente para as coisas à sua volta para as mostrar ou ainda quando a criança usa gestos (SCQ22);
- apresenta uma variedade normal de expressões faciais (SCQ33).

Neste seguimento, aquando do diagnóstico de ASD é essencial analisar detalhadamente as crianças que demonstrem os comportamentos implícitos nos itens M-CHAT15, SCQ22 e SCQ23. Para esses casos, o risco de se tratarem de falsos positivos é considerável e, portanto, aconselha-se um diagnóstico mais pormenorizado.

Para além disso, o facto de não falharem as questões M-CHAT7, M-CHAT9, M-CHAT12, M-CHAT19, SCQ27, SCQ28, SCQ32 e SCQ35 quando apresentam um diagnóstico errado para o ASD podem também ser determinantes para o erro de classificação, dado que apenas 1 em cada 16 indivíduos falha estes itens.

De seguida, efetuou-se uma análise do comportamento dos classificadores para diferentes conjuntos de dados, de acordo com a idade dos indivíduos aquando da realização da avaliação. Os modelos preditivos foram implementados a 4 grupos de indivíduos: um correspondente às idades inferiores ao 1º quartil da *feature* das idades, um correspondente ao intervalo entre o 1º quartil e a mediana, um entre a mediana e o 3º quartil e finalmente um quarto grupo correspondente às idades superiores ao 3º quartil. No entanto, é relevante referir que não se verificaram alterações significativas entre os resultados obtidos na classificação dos grupos referidos (dados não apresentados).

4.7 *Ranking* dos itens do M-CHAT

Para determinar quais os itens de M-CHAT diferenciam melhor os portadores de ASD dos portadores de PDD-NOS e explorar o impacto das características dos indivíduos nas pontuações deste teste procedeu-se à aquisição do *ranking* dos itens do M-CHAT. Para os dados utilizados no presente estudo o *ranking* foi adquirido através do método SVM-RFE. Dado que esta técnica permite obter a lista de importância relativa das features tendo em conta as classes a distinguir e como obteve ótimos resultados de classificação considerou-se um bom procedimento a executar para a análise do teste M-CHAT em particular.

Na tabela 4.6 é possível comparar os resultados obtidos na investigação efetuada

Tabela 4.6: *Ranking* dos itens mais relevantes do teste M-CHAT para a classificação do ASD, obtidos no estudo efetuado e na literatura consultada.

Resultados obtidos	Robins <i>et al.</i> [41]	Wong <i>et al.</i> [57]	Kamio <i>et al.</i> [54]	Srisinghasongkram <i>et al.</i> [58]
MCHAT17	MCHAT7	MCHAT13	MCHAT6	MCHAT2
MCHAT23	MCHAT14	MCHAT5	MCHAT13	MCHAT6
MCHAT19	MCHAT2	MCHAT7	MCHAT5	MCHAT7
MCHAT22	MCHAT9	MCHAT23	MCHAT15	MCHAT8
MCHAT9	MCHAT15	MCHAT9	MCHAT21	MCHAT9
MCHAT14	MCHAT13	MCHAT15	MCHAT9	MCHAT13

com os resultados analisados na revisão bibliográfica (secção 2.2.2).

Tendo em conta os dados dos indivíduos participantes no presente estudo, os itens do teste M-CHAT mais relevantes (por ordem decrescente de importância) em distinguir crianças com ASD de crianças suspeitas de possuir problemas de neurodesenvolvimento correspondem às perguntas: "Olha para as coisas para as quais o adulto está a olhar?" (MCHAT17), "Procura a sua reação facial quando se vê confrontada com situações desconhecidas?" (MCHAT23), "Tenta chamar a sua atenção para o que está a fazer?" (MCHAT19), "Por vezes fica a olhar para o vazio ou deambula ao acaso pelos espaços?" (MCHAT22), "Alguma vez lhe trouxe objetos (brinquedos) para lhe mostrar alguma coisa?" (MCHAT9) e "Responde/olha quando o(a) chamam pelo nome?" (MCHAT14).

Ao proceder à comparação dos itens relevantes em cada estudo é possível constatar que o item 9 do teste M-CHAT é comum a todos. Relativamente aos restantes itens dados como relevantes pelo método SVM-RFE verifica-se ainda que o item 14 e 23 são também considerados importantes na pesquisa de Robins *et al.* [41] e Wong *et al.* [57], respetivamente. Desta forma, o facto da criança trazer objetos para mostrar parece ser determinante e transversal a todas as culturas e países avaliados através das respetivas traduções do teste M-CHAT.

Embora não exista concordância com os restantes itens críticos obtidos, isto pode justificar-se devido às diferenças culturais de cada população dado que os estudos comparativos dizem respeito a diferentes países. Neste caso comparam-se as versões: americana [41], chinesa [57], japonesa [54] e, ainda, a versão tailandesa [58] do M-CHAT com a versão portuguesa. Para além disso, as pesquisas em causa tentaram distinguir crianças com ASD de crianças com desenvolvimento normal, enquanto

que o presente estudo procura diferenciar crianças com ASD de crianças suspeitas de possuir problemas de neurodesenvolvimento. Este fator pode também ele afetar os itens que se identificam como críticos.

Com esta análise é possível afirmar que é necessário ajustar os testes e métodos de triagem para os contextos culturais em que eles são usados. A diferença entre os itens críticos relevantes para cada contexto cultural é interessante e potencialmente relevante ao nível científico, mas este tópico vai além do objetivo principal deste estudo.

Conclusão

Neste momento, a intervenção precoce surge como a melhor resposta para o Transtorno do Espectro Autista. Torna-se assim imperativo o desenvolvimento de ferramentas confiáveis para a triagem de crianças menores de três anos, que permitirão um diagnóstico numa fase mais inicial da vida do indivíduo e conseqüentemente uma intervenção antecipada, levando ao objetivo final — diminuição dos sintomas associados ao ASD.

Tendo em consideração o objetivo principal deste estudo — a criação de um modelo preditivo capaz de distinguir o ASD em crianças de alto risco — pode-se concluir que foi alcançado com sucesso, dado que os resultados da avaliação do classificador foram muito satisfatórios. De qualquer forma, é importante salientar que os dados utilizados neste estudo apresentavam à partida uma elevada separabilidade entre classes, o que poderá ter contribuído para um resultado enganador ou ilusório da predição do ASD.

É importante também evidenciar que 67% dos indivíduos que participaram no estudo possuíam um diagnóstico clínico de ASD, comprovando que os dados utilizados correspondiam a uma amostra de alto risco e não a indivíduos geralmente avaliados em consultas comuns de cuidados primários em crianças.

Pode-se considerar que os resultados obtidos para o modelo preditivo fornecem um valioso contributo para o diagnóstico diferencial e a tomada de decisão clínica, permitindo uma rápida e correta atualização das famílias e conseqüente planeamento do tratamento, particularmente ao avaliar as crianças em situação de risco. Este fator é de uma importância crucial, pois os encaminhamentos antecipados aumentam as intervenções precoces, melhorando o prognóstico ao longo da vida e atenuando, de forma substancial, o impacto que o ASD tem no quotidiano das crianças e das respetivas famílias.

Foi também relevante avaliar o desempenho do M-CHAT e do SCQ em crianças com suspeitas de ASD ou de outra patologia do neurodesenvolvimento, de acordo com a aplicabilidade clínica atual e real da triagem de ASD. Embora a faixa etária da nossa amostra seja diferente das originalmente recomendadas para estes testes, os resultados obtidos indicam que ambos os questionários são muito úteis na avaliação de crianças entre os três e os cinco anos de idade.

Para além disso, tendo em consideração que este trabalho corresponde ao primeiro estudo de avaliação da capacidade de diagnóstico da versão portuguesa do M-CHAT e do SCQ, seria interessante o investimento numa análise mais aprofundada ao nível clínico, com o objetivo de se perceber se existe alguma correlação entre os resultados obtidos e os efeitos reais do ASD.

No que toca a futuros trabalhos nesta área, e tendo em consideração os resultados aqui obtidos, uma das possíveis direções a seguir seria a reprodução deste estudo com crianças ainda mais jovens. Essa análise seria de grande interesse para identificar a idade mais precoce em que uma triagem confiável e preditiva para o ASD é possível e, por meio de programas de intervenção adequados, promover resultados de desenvolvimento ainda melhores.

Por fim, como observado anteriormente, é importante destacar que o modelo preditivo criado não foi projetado para fins de diagnóstico nem substitui a necessidade de avaliações adicionais e abrangentes, pode sim contribuir como um método de apoio à decisão. Portanto, apesar dos resultados encorajadores, a importância da avaliação clínica não pode ser subestimada, e deve ser combinada com testes e avaliações apropriados para discriminar com precisão e fiabilidade se uma criança apresenta ASD, se tem outro transtorno do desenvolvimento neurológico ou se possui um desenvolvimento neurológico típico.

Bibliografia

- [1] National Institute of Neurological Disorders and Stroke, “Autism Spectrum Disorder Fact Sheet — National Institute of Neurological Disorders and Stroke,” 2017.
- [2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*. 2013.
- [3] The National Institute of Mental Health, “NIMH » Autism Spectrum Disorder,” 2018.
- [4] Baron-Cohen, “The extreme male brain theory of autism,” *Trends in cognitive sciences*, vol. 6, no. 6, pp. 248–254, 2002.
- [5] V. B. Gupta, *Autistic spectrum disorders in children*. Marcel Dekker, 2004.
- [6] S. Ozonoff, G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Dobkins, T. Hutman, J. M. Iverson, R. Landa, S. J. Rogers, M. Sigman, and W. L. Stone, “Recurrence Risk for Autism Spectrum Disorders: A Baby Siblings Research Consortium Study,” *Pediatrics*, 2011.
- [7] S. Sumi, H. Taniai, T. Miyachi, and M. Tanemura, “Sibling risk of pervasive developmental disorder estimated by means of an epidemiologic survey in Nagoya, Japan,” *Journal of Human Genetics*, vol. 51, no. 6, pp. 518–522, 2006.
- [8] R. E. Rosenberg, J. K. Law, G. Yenokyan, J. McGready, W. E. Kaufmann, and P. A. Law, “Characteristics and Concordance of Autism Spectrum Disorders Among 277 Twin Pairs,” vol. 163, no. 10, pp. 907–914, 2009.
- [9] J. Hallmayer, S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe,

- J. Miller, A. Fedele, J. Collins, K. Smith, L. Lotspeich, L. A. Croen, S. Ozonoff, J. K. Grether, and N. Risch, "Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism," vol. 68, no. 11, pp. 1095–1102, 2015.
- [10] A. Ronald, F. Happé, P. Bolton, L. M. Butcher, T. S. Price, S. Wheelwright, S. Baron-Cohen, and R. Plomin, "Genetic heterogeneity between the three components of the autism spectrum: A twin study," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 45, no. 6, pp. 691–699, 2006.
- [11] H. Taniyai, T. Nishiyama, T. Miyachi, M. Imaeda, and S. Sumi, "Genetic influences on the broad spectrum of autism: Study of proband-ascertained twins," *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, vol. 147, no. 6, pp. 844–849, 2008.
- [12] L. Croen, D. Najjar, B. Fireman, and E. Al, "Maternal and paternal age and risk of autism Spectrum Disorders," *Archives of Pediatrics and Adolescent Medicine*, vol. 161, pp. 334–340, 2007.
- [13] M. S. Durkin, M. J. Maenner, C. J. Newschaffer, L.-C. Lee, C. M. Cunniff, J. L. Daniels, R. S. Kirby, L. Leavitt, L. Miller, W. Zahorodny, and L. A. Schieve, "Advanced Parental Age and the Risk of Autism Spectrum Disorder," *American Journal of Epidemiology*, vol. 168, no. 11, pp. 1268–1276, 2008.
- [14] M. B. Lauritsen, C. B. Pedersen, and P. B. Mortensen, "Effects of familial risk factors and place of birth on the risk of autism: A nationwide register-based study," *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 46, no. 9, pp. 963–971, 2005.
- [15] S. Tordjman, C. Baumann, L. Burglen, D. Cohen, E. Excoffier, G. Lazar, P. Mazet, A. Verloes, and D. He, "Specific Genetic Disorders and Autism : Clinical Contribution Towards their Identification," vol. 35, no. 1, 2005.
- [16] D. Schendel and T. K. Bhasin, "Birth Weight and Gestational Age Characteristics of Children With Autism, Including a Comparison With Other Developmental Disabilities," *Pediatrics*, vol. 121, no. 6, pp. 1155–1164, 2008.
- [17] A. Society, "Diagnosis - Autism Society."
- [18] P. Filipek, P. Accardo, S. Ashwal, G. Baranek, E. Cook, G. Dawson, B. Gor-

- don, J. Gravel, C. Johnson, R. Kallen, S. Levy, N. Minshew, S. Ozonoff, B. Prizant, I. Rapin, S. Rogers, W. Stone, S. Teplin, R. Tuchman, and F. Volkmar, "Practice parameter: Screening and diagnosis of autism: Report of the Quality Standards Subcommittee of the American Academy of Neurology and the Child Neurology Society," *Neurology*, vol. 55, no. 4, pp. 468–479, 2000.
- [19] "Healthcare Providers — Autism Spectrum Disorder (ASD) — NCBDDD — CDC."
- [20] O. Tadevosyan-leyfer, M. Dowd, R. Mankoski, B. Winklosky, S. Putnam, L. McGrath, H. Tager-flusberg, and S. E. Folstein, "A Principal Components Analysis of the Autism Diagnostic Interview-Revised," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 42, pp. 864–872, 7 2003.
- [21] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, Jr., B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism," *Journal of Autism and Developmental Disorders*, vol. 30, no. 3, pp. 205–223, 2000.
- [22] M. E. Van Bourgondien, L. M. Marcus, and E. Schopler, "Comparison of DSM-III-R and childhood autism rating scale diagnoses of autism," *Journal of Autism and Developmental Disorders*, vol. 22, pp. 493–506, 12 1992.
- [23] J. Gilliam, "Gilliam autism rating scale: Examiner's manual," 1995.
- [24] S. Sutera, J. Pandey, E. L. Esser, M. A. Rosenthal, L. B. Wilson, M. Barton, J. Green, S. Hodgson, D. L. Robins, T. Dumont-Mathieu, and D. Fein, "Predictors of optimal outcome in toddlers diagnosed with autism spectrum disorders," *Journal of Autism and Developmental Disorders*, vol. 37, no. 1, pp. 98–107, 2007.
- [25] T. Charman, E. Taylor, A. Drew, H. Cockerill, J. A. Brown, and G. Baird, "Outcome at 7 years of children diagnosed with autism at age 2: Predictive validity of assessments conducted at 2 and 3 years of age and pattern of symptom change over time," *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 46, no. 5, pp. 500–513, 2005.
- [26] C. Lord, S. Risi, P. S. DiLavore, C. Shulman, A. Thurm, and A. Pickles, "Autism From 2 to 9 Years of Age," *Archives of General Psychiatry*, vol. 63,

p. 694, 6 2006.

- [27] C. Dietz, S. Swinkels, E. Van Daalen, H. Van Engeland, and J. K. Buitelaar, “Screening for autistic spectrum disorder in children aged 14-15 months. II: Population screening with the Early Screening of Autistic Traits Questionnaire (ESAT). Design and general findings,” *Journal of Autism and Developmental Disorders*, vol. 36, no. 6, pp. 713–722, 2006.
- [28] K. Chawarska, A. Klin, R. Paul, and F. Volkmar, “Autism spectrum disorder in the second year: Stability and change in syndrome expression,” *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 48, no. 2, pp. 128–138, 2007.
- [29] B. S.E., R. S.J., and F. E., “Autism Spectrum Disorders: Early Detection, Intervention, Education, and Psychopharmacological Management,” *Canadian Journal of Psychiatry*, vol. 48, no. 8, pp. 506–516, 2003.
- [30] G. Dawson, S. B. Ashman, and L. J. Carver, “The role of early experience in shaping behavioral and brain development and its implications for social policy,” *Dev Psychopathol*, vol. 12, no. 4, pp. 695–712, 2000.
- [31] S. L. Harris and J. S. Handleman, “Age and IQ at Intake as Predictors of Placement for Young Children with Autism : A Four- to Six- Year Follow-Up Age and IQ at Intake as Predictors of Placement for Young Children with Autism : A Four- to Six-Year Follow-Up,” vol. 30, no. January, 2000.
- [32] C. W. McGovern and M. Sigman, “Continuity and change from early childhood to adolescence in autism,” *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 46, no. 4, pp. 401–408, 2005.
- [33] J. D. Osofsky and A. F. Lieberman, “A call for integrating a mental health perspective into systems of care for abused and neglected infants and young children,” *American Psychologist*, vol. 66, no. 2, pp. 120–128, 2011.
- [34] P. F. Bolton, J. Golding, A. Emond, and C. D. Steer, “Autism spectrum disorder and autistic traits in the avon longitudinal study of parents and children: Precursors and early signs,” *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 51, no. 3, pp. 249–260, 2012.
- [35] P. R. F., “State differences in the identification of autistic spectrum disorders,”

-
- Archives of Pediatrics and Adolescent Medicine*, vol. 159, no. 3, pp. 266–269, 2005.
- [36] M. Elsabbagh, G. Divan, Y. J. Koh, Y. S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C. S. Paula, C. Wang, M. T. Yasamy, and E. Fombonne, “Global Prevalence of Autism and Other Pervasive Developmental Disorders,” *Autism Research*, vol. 5, no. 3, pp. 160–179, 2012.
- [37] G. Oliveira, A. Ataíde, C. Marques, T. S. Miguel, A. M. Coutinho, L. Mota-Vieira, E. Gonçalves, N. M. Lopes, V. Rodrigues, H. Carmona da Mota, and A. M. Vicente, “Epidemiology of autism spectrum disorder in Portugal: prevalence, clinical characterization, and medical conditions,” *Developmental Medicine & Child Neurology*, vol. 49, no. 10, pp. 726–733, 2007.
- [38] CDC, “Treatment — Autism Spectrum Disorder (ASD) — NCBDDD — CDC.”
- [39] E. Fombonne, “Epidemiological surveys of autism and other pervasive developmental disorders: an update.,” *Journal of autism and developmental disorders*, vol. 33, pp. 365–82, 8 2003.
- [40] J. Pandey, A. Verbalis, D. L. Robins, H. Boorstein, A. Klin, T. Babitz, K. Chawarska, F. Volkmar, J. Green, M. Barton, and D. Fein, “Screening for autism in older and younger toddlers with the Modified Checklist for Autism in Toddlers,” *Autism*, vol. 12, no. 5, pp. 513–535, 2008.
- [41] B. M. & G. J. D. Robins Fein, D., “The modified-checklist for autism in toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders.,” *Journal of Autism and Developmental Disorders*, vol. 31, no. 2, pp. 131–144, 2001.
- [42] T. Charman and K. Gotham, “Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders - lessons from research and practice,” *Child and Adolescent Mental Health*, vol. 18, no. 1, pp. 52–63, 2013.
- [43] & L. C. M. Rutter Bailey, A., “Social Communication Questionnaire,” *Los Angeles, CA*, 2003.
- [44] T. Charman, G. Baird, E. Simonoff, S. Chandler, A. Davison-Jenkins, A. Sharma, T. O’Sullivan, and A. Pickles, “Testing two screening instruments

- for autism spectrum disorder in UK community child health services,” *Developmental Medicine and Child Neurology*, vol. 58, no. 4, pp. 369–375, 2016.
- [45] S. Baron-Cohen, J. Allen, and C. Gillberg, “Brief Reports Can Autism be Detected at 18 Months?,” *British Journal of Psychiatry*, vol. 161, pp. 839–843, 1991.
- [46] DGS, “Programa Nacional de Saúde Infantil e Juvenil,” *Direção Geral de Saúde*, pp. 1–107, 2012.
- [47] D. L. Robins, “Screening for autism spectrum disorders in primary care settings,” *Autism*, vol. 12, no. 5, pp. 537–556, 2008.
- [48] J. M. Kleinman, D. L. Robins, P. E. Ventola, J. Pandey, H. C. Boorstein, E. L. Esser, L. B. Wilson, M. A. Rosenthal, S. Sutera, A. D. Verbalis, M. Barton, S. Hodgson, J. Green, T. Dumont-Mathieu, F. Volkmar, K. Chawarska, A. Klin, and D. Fein, “The modified checklist for autism in toddlers: A follow-up study investigating the early detection of autism spectrum disorders,” *Journal of Autism and Developmental Disorders*, vol. 38, no. 5, pp. 827–839, 2008.
- [49] C. Chlebowski, D. L. Robins, M. L. Barton, and D. Fein, “Large-Scale Use of the Modified Checklist for Autism in Low-Risk Toddlers,” *Pediatrics*, vol. 131, no. 4, pp. e1121–e1127, 2013.
- [50] R. Canal-Bedia, P. García-Primo, M. V. Martín-Cilleros, J. Santos-Borbujo, Z. Guisuraga-Fernández, L. Herráez-García, M. Del Mar Herráez-García, L. Boada-Muñoz, J. Fuentes-Biggi, and M. Posada-De La Paz, “Modified checklist for autism in toddlers: Cross-cultural adaptation and validation in Spain,” *Journal of Autism and Developmental Disorders*, vol. 41, no. 10, pp. 1342–1351, 2011.
- [51] L. C. Eaves, H. Wingert, and H. H. Ho, “Screening for autism: Agreement with diagnosis,” *Autism*, vol. 10, no. 3, pp. 229–242, 2006.
- [52] A. V. Snow and L. Lecavalier, “Sensitivity and specificity of the modified checklist for autism in toddlers and the social communication questionnaire in preschoolers suspected of having pervasive developmental disorders,” *Autism*, vol. 12, no. 6, pp. 627–644, 2008.
- [53] Y. Kamio, N. Inada, T. Koyama, E. Inokuchi, K. Tsuchiya, and M. Kuroda,

- “Effectiveness of using the modified checklist for autism in toddlers in two-stage screening of autism spectrum disorder at the 18-month health check-up in Japan,” *Journal of Autism and Developmental Disorders*, vol. 44, no. 1, pp. 194–203, 2014.
- [54] Y. Kamio, H. Haraguchi, A. Stickley, K. Ogino, M. Ishitobi, and H. Takahashi, “Brief Report: Best Discriminators for Identifying Children with Autism Spectrum Disorder at an 18-Month Health Check-Up in Japan,” *Journal of Autism and Developmental Disorders*, vol. 45, no. 12, pp. 4147–4153, 2015.
- [55] D. Fein, D. Robins, and M. Barton, “Testing two screening instruments for autism spectrum disorder,” *Developmental Medicine and Child Neurology*, vol. 58, no. 3, pp. 314–315, 2016.
- [56] S. A. Samadi and R. McConkey, “Screening for Autism in Iranian Preschoolers: Contrasting M-CHAT and a Scale Developed in Iran,” *Journal of Autism and Developmental Disorders*, vol. 45, no. 9, pp. 2908–2916, 2015.
- [57] V. Wong, L.-H. S. Hui, W.-C. Lee, L.-S. J. Leung, P.-K. P. Ho, W.-L. C. Lau, C.-W. Fung, and B. Chung, “A Modified Screening Tool for Autism (Checklist for Autism in Toddlers [CHAT-23]) for Chinese Children,” *Pediatrics*, vol. 114, no. 2, pp. e166–e176, 2004.
- [58] P. Srisinghasongkram, C. Pruksananonda, and W. Chonchaiya, “Two-Step Screening of the Modified Checklist for Autism in Toddlers in Thai Children with Language Delay and Typically Developing Children,” *Journal of Autism and Developmental Disorders*, vol. 46, no. 10, pp. 3317–3329, 2016.
- [59] S. K. Berument, M. Rutter, C. Lord, A. Pickles, A. Bailey, S. Kazak, B. Michael, C. Lord, A. Pickles, and A. Bailey, “Autism screening questionnaire : diagnostic validity Autism screening questionnaire : diagnostic validity,” *British Journal of Psychiatry*, vol. 175, pp. 444–451, 1999.
- [60] C. Lord, M. Rutter, and A. LeCouteur, “Autism diagnostic interview: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders,” *J Autism Dev Disord*, vol. 24, no. 5, 1994.
- [61] C. Corsello, V. Hus, A. Pickles, S. Risi, E. H. Cook, B. L. Leventhal, and C. Lord, “Between a ROC and a hard place: Decision making and making

- decisions about using the SCQ,” *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 48, no. 9, pp. 932–940, 2007.
- [62] L. D. Wiggins, R. Bakeman, L. B. Adamson, and D. L. Robins, “The Utility of the Social Communication Questionnaire in Screening for Autism in Children Referred for Early Intervention,” *Focus on Autism and Other Developmental Disabilities*, vol. 22, no. 1, pp. 33–38, 2007.
- [63] A. N. Witwer and L. Lecavalier, “Autism screening tools: An evaluation of the Social Communication Questionnaire and the Developmental Behaviour Checklist-Autism Screening Algorithm,” *Journal of Intellectual and Developmental Disability*, vol. 32, no. 3, pp. 179–187, 2007.
- [64] S. Chandler, T. Charman, G. Baird, E. Simonoff, T. Loucas, D. Meldrum, M. Scott, and A. Pickles, “Validation of the Social Communication Questionnaire in a population cohort of children with autism spectrum disorders,” *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 46, no. 10, pp. 1324–1332, 2007.
- [65] T. Charman, G. Baird, E. Simonoff, T. Loucas, S. Chandler, D. Meldrum, T. O. M. Loucas, and A. W. Pickles, “Efficacy of three screening instruments in the identification of autistic-spectrum disorders service AUTHOR ’ S PROOF Efficacy of three screening instruments in the identification of autistic-spectrum disorders,” *British journal of psychiatry*, vol. 39, pp. 554–559, 2007.
- [66] L. C. Eaves, H. D. Wingert, H. H. Ho, and E. C. R. Mickelson, “Screening for autism spectrum disorders with the social communication questionnaire,” *Journal of developmental and behavioral pediatrics : JDBP*, vol. 27, no. 2 Suppl, pp. S95–S103, 2006.
- [67] C. W. Allen, N. Silove, K. Williams, and P. Hutchins, “Validity of the social communication questionnaire in assessing risk of autism in preschool children with developmental problems,” *Journal of Autism and Developmental Disorders*, vol. 37, no. 7, pp. 1272–1278, 2007.
- [68] L. C. Lee, A. B. David, J. Rusyniak, R. Landa, and C. J. Newschaffer, “Performance of the Social Communication Questionnaire in children receiving preschool special education services,” *Research in Autism Spectrum Disorders*, vol. 1, no. 2, pp. 126–138, 2007.

-
- [69] E. L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M. D. Cubiles-de-la Vega, “Missing value imputation on missing completely at random data using multilayer perceptrons,” *Neural Networks*, vol. 24, no. 1, pp. 121–129, 2011.
- [70] N. M. Noor, “Part of Chapter 2 - thesis,” no. November, 2014.
- [71] A. Plaia and A. L. Bondi, “Single imputation method of missing values in environmental pollution data sets,” *Atmospheric Environment*, vol. 40, no. 38, pp. 7316–7330, 2006.
- [72] I. Eekhout, R. M. de Boer, J. W. R. Twisk, H. C. W. de Vet, and M. W. Heymans, “Missing Data,” *Epidemiology*, vol. 23, no. 5, pp. 729–732, 2012.
- [73] C. K. Enders, “Applied Missing Data Analysis,” 2010.
- [74] K. Widaman, “Missing data: What to do with or without them,” *Monographs of the Society for Research in Child Development*, vol. 71, no. 3, pp. 42–64, 2006.
- [75] Z. Zhang, “Missing data imputation: focusing on single imputation.,” *Annals of translational medicine*, vol. 4, no. 1, p. 9, 2016.
- [76] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, “Feature Selection : An Ever Evolving Frontier in Data Mining,” *JMLR: Workshop and Conference Proceedings*, vol. 4, pp. 4–13, 2010.
- [77] M. Dash and H. Liu, “Consistency-based search in feature selection,” *Artificial Intelligence*, vol. 151, pp. 155–176, 2003.
- [78] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, “A comparative evaluation of feature ranking methods for high dimensional bioinformatics data,” *Proceedings of the 2011 IEEE International Conference on Information Reuse and Integration, IRI 2011*, pp. 315–320, 2011.
- [79] S. Nogueira, “Quantifying the Stability of Feature Selection,” 2018.
- [80] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” tech. rep.
- [81] J. Hua, W. D. Tembe, and E. R. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” *Pattern Recognition*,

- vol. 42, no. 3, pp. 409–424, 2009.
- [82] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, “A review of feature selection methods on synthetic data,” *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, 2013.
- [83] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [84] H. Liu, L. Liu, and H. Zhang, “Feature selection using mutual information: An experimental study,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5351 LNAI, pp. 235–246, 2008.
- [85] I. Guyon, *Feature extraction : foundations and applications*. Springer-Verlag, 2006.
- [86] I. Guyon, J. Weston, S. Barnhill, V. V. M. Learning, and U. 2002, “Gene selection for cancer classification using support vector machines,” *Springer*, 2002.
- [87] Y. Ding and D. Wilkins, “Improving the performance of SVM-RFE to select genes in microarray data,” *BMC Bioinformatics*, vol. 7, no. SUPPL.2, pp. 1–8, 2006.
- [88] K. Duan and J. C. Rajapakse, “SVM-RFE peak selection for cancer classification with mass spectrometry data,” *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference*, pp. 191–200, 2004.
- [89] P. S. Biology and M. Genetics, “Robust Feature Selection Using Ensemble Feature Selection Techniques.pdf,” pp. 313–325, 2008.
- [90] C. Hu, J. Wang, C. Zheng, S. Xu, H. Zhang, Y. Liang, L. Bi, Z. Fan, B. Han, and W. Xu, “Raman spectra exploring breast tissues: Comparison of principal component analysis and support vector machine-recursive feature elimination,” *Medical Physics*, vol. 40, no. 6, 2013.
- [91] Y. Tang, Y. Q. Zhang, and Z. Huang, “Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 365–381, 2007.

-
- [92] M. West-Nielsen, E. V. Høgdall, E. Marchiori, C. K. Høgdall, C. Schou, and N. H. H. Heegaard, “Sample Handling for Mass Spectrometric Proteomic Investigations of Human Sera,” *Analytical Chemistry*, vol. 77, pp. 5114–5123, 8 2005.
- [93] S. S. Shapiro and M. B. Wilk, “An Analysis of Variance Test for Normality (Complete Samples),” *Biometrika*, vol. 52, p. 591, 12 1965.
- [94] H. Levene, “On a Matching Problem Arising in Genetics,” *The Annals of Mathematical Statistics*, vol. 20, pp. 91–94, 3 1949.
- [95] Student, “THE PROBABLE ERROR OF A MEAN,” *Biometrika*, vol. 6, pp. 1–25, 3 1908.
- [96] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, vol. 1, p. 80, 12 1945.
- [97] J. B. Carlin and L. W. Doyle, “Statistics for Clinicians 4 : Basic concepts of statistical reasoning : Hypothesis tests and the t-test,” *Group*, pp. 72–77, 2001.
- [98] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, “Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data,” *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.
- [99] J. Winter, “Using the Student ’ s t -test with extremely small sample sizes,” *Practical Assessment, Research & Evaluation*, vol. 18, no. 10, pp. 1–12, 2013.
- [100] A. C. Haury, P. Gestraud, and J. P. Vert, “The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures,” *PLoS ONE*, vol. 6, no. 12, pp. 1–12, 2011.
- [101] V. Ganganwar, “An overview of classification algorithms for imbalanced datasets,” Tech. Rep. 4, 2012.
- [102] N. V. Chawla, N. Japkowicz, and A. Kotcz, “Editorial: Special Issue on Learning from Imbalanced Data Sets,” *ACM SIGKDD Explorations Newsletter*, vol. 6, p. 1, 6 2004.
- [103] G. Batista, R. Prati, M. M. A. S. e. Newsletter, and U. 2004, “A study of the

- behavior of several methods for balancing machine learning training data,” *dl.acm.org*, 2004.
- [104] P. C. Austin, “A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality,” *Statistics in Medicine*, vol. 26, pp. 2937–2957, 7 2007.
- [105] M. Pohar, M. Blas, and S. Turk, “Comparison of Logistic Regression and Linear Discriminant Analysis : A Simulation Study,” *Metodološki zvezki*, vol. 1, no. 1, pp. 143–161, 2004.
- [106] I. Kurt, M. Ture, and A. T. Kurum, “Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease,” *Expert Systems with Applications*, vol. 34, no. 1, pp. 366–374, 2008.
- [107] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. De Mendonça, “Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests,” *BMC Research Notes*, vol. 4, 2011.
- [108] “Chapter 2 : SVM (Support Vector Machine) — Theory – Machine Learning 101 – Medium.”
- [109] S. Amari and S. Wu, “Improving support vector machine classifiers by modifying kernel functions,” *Neural Networks*, vol. 12, pp. 783–789, 7 1999.
- [110] C. Zhang, C. Liu, X. Zhang, and G. Alpanidis, “An up-to-date comparison of state-of-the-art classification algorithms,” *Expert Systems with Applications*, vol. 82, pp. 128–150, 2017.
- [111] L. Breiman, “Random Forests,” tech. rep., 2001.
- [112] A. Statnikov, L. Wang, and C. F. Aliferis, “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,” *BMC Bioinformatics*, vol. 9, pp. 1–10, 2008.
- [113] R. Díaz-Uriarte and S. Alvarez de Andrés, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, p. 3, 1 2006.

-
- [114] J. H. Song, S. S. Venkatesh, E. A. Conant, P. H. Arger, and C. M. Sehgal, “Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses,” *Academic Radiology*, vol. 12, no. 4, pp. 487–495, 2005.
- [115] S. Rudolfer, G. Paliouras, I. P. C. research, biomedical, and u. 1999, “A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome,” *leg.ufpr.br*.
- [116] W. J. Long, J. L. Griith, H. P. S er, and R. B. D, “A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain,” tech. rep., 1993.
- [117] J. D. Sa, “Pattern recognition: concepts, methods and applications,” 2012.
- [118] A. C. Braga, “Curvas ROC: aspectos funcionais e aplicações,” 2001.
- [119] J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. Routledge, 5 1996.
- [120] J. A. Hanley and B. J. McNeil, “The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve1,” tech. rep.
- [121] Y. Jiang, J. Lin, B. Cukic, T. M. S. o. S. . . . , and U. 2009, “Variance analysis in software fault prediction models,” *computer.org*, 2009.

Anexos

A

**Modified Checklist for Autism in
Toddlers (M-CHAT) — Versão
Portuguesa**

Nome: _____ Preenchido por: _____

Data de Nascimento: _____

Parentesco do informador: _____

Data: _____

Modified Checklist for Autism in Toddlers (M-CHAT)

Diana Robins, Deborah Fein & Marianne Barton, 1999

Por favor, preencha este questionário sobre o comportamento usual da criança. Responda a todas as questões. Se o comportamento descrito for raro (ex. foi observado uma ou duas vezes), responda como se a criança não o apresente. Faça um círculo à volta da resposta "Sim" ou "Não".

1	Gosta de brincar ao colo fazendo de "cavalinho", etc.?	Sim	Não
2	Interessa-se pelas outras crianças?	Sim	Não
3	Gosta de subir objectos, como por exemplo, cadeiras, mesas?	Sim	Não
4	Gosta de jogar às escondidas?	Sim	Não
5	Brinca ao faz-de-conta, por exemplo, falar ao telefone ou dar de comer a uma boneca, etc.?	Sim	Não
6	Aponta com o indicador para pedir alguma coisa?	Sim	Não
7	Aponta com o indicador para mostrar interesse em alguma coisa?	Sim	Não
8	Brinca apropriadamente com brinquedos (carros ou Legos) sem levá-los à boca, abanar ou deitá-los ao chão?	Sim	Não
9	Alguma vez lhe trouxe objectos (brinquedos) para lhe mostrar alguma coisa?	Sim	Não
10	A criança mantém contacto visual por mais de um ou dois segundos?	Sim	Não
11	É muito sensível aos ruídos (ex. tapa os ouvidos)?	Sim	Não
12	Sorri como resposta às suas expressões faciais ou ao seu sorriso?	Sim	Não
13	Imita o adulto (ex. faz uma careta e ela imita)?	Sim	Não
14	Responde/olha quando o(a) chamam pelo nome?	Sim	Não
15	Se apontar para um brinquedo do outro lado da sala, a criança acompanha com o olhar?	Sim	Não
16	Já anda?	Sim	Não
17	Olha para as coisas para as quais o adulto está a olhar?	Sim	Não
18	Faz movimentos estranhos com as mãos/dedos próximo da cara?	Sim	Não
19	Tenta chamar a sua atenção para o que está a fazer?	Sim	Não
20	Alguma vez se preocupou quanto à sua audição?	Sim	Não
21	Compreende o que as pessoas lhe dizem?	Sim	Não
22	Por vezes fica a olhar para o vazio ou deambula ao acaso pelos espaços?	Sim	Não
23	Procura a sua reacção facial quando se vê confrontada com situações desconhecidas?	Sim	Não

Traduzido pela Unidade de Autismo

Centro de Desenvolvimento da Criança – Hospital Pediátrico de Coimbra
Autorização Diana Robins

Modified Checklist for Autism in Toddlers (M-CHAT)

Diana Robins, Deborah Fein & Marianne Barton, 1999

O (M-CHAT) é um breve questionário referente ao desenvolvimento e comportamento utilizado em crianças dos 16 aos 30 meses, com o objectivo de rastrear as perturbações do espectro do autismo (PEA). Pode ser aplicado tanto numa avaliação periódica de rotina (cuidados primários de saúde), como por profissionais especializados em casos de suspeita. Como na maioria dos testes de rastreio poderá existir um grande número de falsos positivos, indicando que nem todas as crianças que cotam neste questionário irão ser diagnosticadas com esta perturbação. No entanto estes resultados podem apontar para a existência de outras anomalias do desenvolvimento, sendo por isso necessária a avaliação por profissionais desta área.

Cotação:

A cotação do **M-CHAT** leva menos de dois minutos. Resultados superiores a **3 (falha em 3 itens no total)** ou em **2 dos itens considerados críticos (2,7,9,13,14,15)**, após confirmação, justificam uma avaliação formal por técnicos de neurodesenvolvimento.

As respostas Sim/Não são convertidas em passa/falha. A tabela que se segue, regista as repostas consideradas **Falha** para cada um dos itens do M-CHAT. As questões a “**Negrito**” representam os **itens CRITICOS**.

1. Não	6. Não	11. Sim	16. Não	21. Não
2. Não	7. Não	12. Não	17. Não	22. Sim
3. Não	8. Não	13. Não	18. Sim	23. Não
4. Não	9. Não	14. Não	19. Não	
5. Não	10. Não	15. Não	20. Sim	

Referências bibliográficas:

- <http://www2.gsu.edu/~psydlr>
- Kleinman et al. (2008) ‘The Modified Checklist for Autism in Toddlers: a Follow-up Study Investigating the Early Detection of Autism Spectrum Disorders’, *Journal of Autism and Developmental Disorders*, 38:827-839.
- Robins, D. (2008) ‘Screening for autism spectrum disorders in primary care settings’, *Autism*, Vol 12(5) 481-500.

B

**Social Communication
Questionnaire (SCQ) — Versão
Portuguesa**

Social Communication Questionnaire

AutoScore™ Form

Michael Rutter, M.D., F.R.S., Anthony Bailey, M.D.,
Sibel Kazak Berument, Ph.D., Catherine Lord, Ph.D. and Andrew Pickles, Ph.D.

Publicado por WESTERN PSYCHOLOGICAL SERVICES

Data: ____ / ____ / ____	
Nome: _____	
Data de Nascimento: ____ / ____ / ____	
Idade Cronológica: ____ Y ____ M	Género <input type="checkbox"/> F <input type="checkbox"/> M
Preenchido por: _____	
Grau de Parentesco: _____	
Local da Avaliação: _____	
Examinador: _____	

Instruções

Obrigada por ter disponibilizado o seu tempo para responder a este questionário. Por favor responda a cada uma das questões fazendo um círculo à volta do **SIM** ou do **NÃO**.

Algumas das questões colocadas descrevem vários padrões de comportamento que se relacionam entre si. Faça um círculo à volta do **SIM** nestas questões sempre que **QUALQUER UM** dos comportamentos referidos se encontrou presente **NÃO** seu/vosso filho/a nos **ÚLTIMOS 3 MESES**.

Mesmo que não tenha a certeza se determinado comportamento se encontrou presente ou não no período de tempo referido, não deixe de responder **SIM** ou **NÃO** a todas as questões, de acordo com o que pensa.

- | | | | |
|--|---------------------------------------|-----|-----|
| 1. Tem frases? | SE NÃO avance para a questão 8 | SIM | NÃO |
| 2. É capaz de manter um diálogo adequado (i.é, uma conversa que envolva uma discussão e uma troca de ideias)? | | SIM | NÃO |
| 3. Alguma vez utilizou ou repetiu palavras/frases estranhas de uma forma exaustiva e repetitiva (sempre da mesma maneira), fossem palavras/frases que tivesse ouvido dizer ou que tivesse inventado? | | SIM | NÃO |
| 4. Alguma vez fez perguntas ou afirmações socialmente inapropriadas (ex. fazer perguntas de foro pessoal/íntimo ou comentários constrangedores a/sobre alguém)? | | SIM | NÃO |
| 5. Alguma vez inverteu a utilização dos pronomes numa frase (i.é, dizer <i>tu</i> ou <i>ele/ela</i> em vez de <i>eu</i>)? | | SIM | NÃO |
| 6. Alguma vez utilizou palavras aparentemente inventadas por si ou utilizou uma forma metafórica de dizer as coisas (ex. referir-se a <i>vapor</i> como <i>chuva quente</i>)? | | SIM | NÃO |
| 7. Alguma vez repetiu a mesma palavra/frase de modo exaustivo e repetitivo (sempre da mesma maneira) ou lhe pediu para o fazer? | | SIM | NÃO |

8 Alguma vez realizou uma actividade ou tarefa de um modo particular ou numa determinada ordem ou insistiu que outras pessoas realizassem determinados rituais (ex. insistir que alguém abrisse e fechasse a porta repetidamente)?	SIM	NÃO
9 Considera que ele/ela revela uma expressão facial adequada a determinada situação?	SIM	NÃO
10 Alguma vez utilizou a mão de outras pessoas para segurar ou alcançar objectos ou como se fosse um prolongamento do seu próprio corpo (ex. utilizar o dedo de outra pessoa para apontar ou colocar a mão de outra pessoa NÃO puxador da porta para a abrir)?	SIM	NÃO
11 Alguma vez manifestou interesse por algo ou por alguma actividade que considerasse estranha ou que parecesse estranha aos olhos de outras pessoas (ex. marcas de automóveis, canos de esgoto ou horários)?	SIM	NÃO
12 Alguma vez lhe pareceu estar mais interessado/a em partes de um brinquedo ou objecto do que propriamente em utilizá-lo ou em brincar com ele do modo que seria esperado (ex. girar as rodas de um carro em vez de brincar com ele)?	SIM	NÃO
13 Alguma vez manifestou um interesse que tivesse considerado ser de uma intensidade invulgar, embora fosse algo apropriado à sua idade e aos gostos do seu grupo de pares (ex. dinossauros, comboios)?	SIM	NÃO
14 Alguma vez revelou um interesse invulgar em fixar o olhar, tocar, ouvir, lambeu ou cheirar objectos ou pessoas?	SIM	NÃO
15 Alguma vez manifestou maneirismos motores ou fez movimentos estranhos com as mãos ou dedos (ex. fazer <i>flapping</i> (“abandar as mãos”) ou abandar os dedos ou as mãos à frente dos olhos)?	SIM	NÃO
16 Alguma vez manifestou movimentos estranhos como, por exemplo, andar às voltas ou pular repetidamente?	SIM	NÃO
17 Alguma vez se auto-agrediu de forma deliberada (ex. morder-se, bater com a cabeça na parede)?	SIM	NÃO
18 Tem um objecto que insiste em levar sempre consigo, que considere invulgar?	SIM	NÃO
19 Tem algum amigo/a preferido/a ou um/a melhor amigo/a?	SIM	NÃO

Nos comportamentos seguintes, tenha em conta o período de tempo entre os 4 e os 5 anos de idade. Pode facilitar recordar-se de eventos chave, como o início de frequência no jardim de infância, mudanças de casa, época natalícia ou outros eventos que são particularmente marcantes para a sua/vossa família.

Se o seu/vosso filho/a ainda não tiver 4 anos de idade cronológica, refira-se ao seu comportamento nos últimos 12 meses.

20. Quando tinha 4-5 anos, alguma vez falou consigo só para ser amável (em vez de ser para pedir algo)?	SIM	NÃO
21. Quando tinha 4-5 anos, alguma vez o/a imitou espontaneamente ou a outras pessoas ou as vossas acções (ex. aspirar, jardinagem, consertar coisas)?	SIM	NÃO
22. Quando tinha 4-5 anos, alguma vez apontou espontaneamente para as coisas à sua volta para as mostrar (ou apontava apenas como forma de pedir algo)?	SIM	NÃO
23. Quando tinha 4-5 anos, alguma vez usou gestos, para além do apontar ou puxar pela sua mão, de forma a demonstrar o que queria?	SIM	NÃO
24. Quando tinha 4-5 anos, acenava com a cabeça para dizer que “sim”?	SIM	NÃO
25. Quando tinha 4-5 anos, acenava com a cabeça para dizer que “não”?	SIM	NÃO
26. Quando tinha 4-5 anos, costumava olhar directamente para sua cara enquanto fazia coisas ou falava para si?	SIM	NÃO
27. Quando tinha 4-5 anos, sorria em resposta ao sorriso de outra pessoa?	SIM	NÃO
28. Quando tinha 4-5 anos, alguma vez ele/ela lhe mostrava coisas de forma a obter a sua atenção?	SIM	NÃO
29. Quando tinha 4-5 anos, costumava partilhar coisas consigo, sem ser comida?	SIM	NÃO
30. Quando tinha 4-5 anos, alguma vez procurava partilhar consigo a alegria de estar a fazer algo que lhe provocava prazer?	SIM	NÃO
31. Quando tinha 4-5 anos, alguma vez procurou confortá-lo/a quando estava triste, magoado/a ou doente?	SIM	NÃO
32. Quando ele/ela tinha 4-5 anos, quando queria alguma coisa ou precisava de algo, olhava para si e usava sons ou palavras acompanhadas por gestos para obter a sua atenção?	SIM	NÃO
33. Quando tinha 4-5 anos, apresentava uma variedade normal de expressões faciais?	SIM	NÃO
34. Quando tinha 4-5 anos, participava espontaneamente e tentava imitar acções em jogos sociais (ex. brincar à roda ou brincar à apanhada)?	SIM	NÃO
35. Quando tinha 4-5 anos, tinha jogo simbólico, ou seja, brincava ao <i>faz-de-conta</i> ?	SIM	NÃO
36. Quando tinha 4-5 anos, interessava-se por outras crianças aproximadamente da mesma idade, que não conhecia?	SIM	NÃO

- | | | |
|---|-----|-----|
| 37. Quando tinha 4-5 anos, reagia positivamente quando outras crianças se aproximavam dele/dela? | SIM | NÃO |
| 38. Quando tinha 4-5 anos, se chegasse a uma divisão da casa e começasse a falar para ele/ela, sem o/a chamar pelo nome, costumava olhar para si ou prestar atenção ao que lhe estava a dizer? | SIM | NÃO |
| 39. Quando tinha 4-5 anos, alguma vez brincou de forma imaginativa com outra criança de uma maneira que o/a pai/mãe conseguia perceber que as crianças compreendiam o que cada uma estava a imaginar? | SIM | NÃO |
| 40. Quando tinha 4-5 anos, participava cooperativamente em jogos de grupo (ex. escondidas, mata)? | SIM | NÃO |