Aplicação de Árvore de Decisão para Auxílio ao Diagnóstico do Transtorno do Espectro Autista

Matheus Frota[1], Manoel Vilela[1],Samuel Hericles[1], Gerônimo Aguiar[1], Pedro Renoir[1], Rayon Nunes[1],Denilson Gomes[1], Iális Cavalcante[1]

[1]Curso de Engenharia da Computação - Universidade Federal do Ceará (UFC) - Campus de Sobral - Caixa Postal 6.050 - 62.010-560 - Sobral - CE - Brazil

Abstract. *Machine Learning algorithms are being applied successfully to many areas of knowledge, in the health field, those algorithms allows professionals to be assisted to diagnose diseases and disorders more anticipatedly and more accurately, contributing to the effecttive treatment of their patients. In this context, the Decision Tree algorithm used in this paper proposes to build a model capable of simplify a complex set of decisions and produce a strategy based on an public and international dataset about the Autism Spectrum Disorder (ASD), allowing to develop a complement in it diagnosis and treatment.*

Resumo. *Algoritmos de aprendizado de máquina estão sendo aplicados com sucesso em diversas áreas do conhecimento. No campo da saúde, estes algoritmos abrem espaço para que profissionais possam ser auxiliados a diagnosticar doenças e transtornos antecipadamente e com maior precisão, contribuindo no tratamento eficaz de seus pacientes. Neste contexto, o algoritmo de Árvore de Decisão utilizado neste artigo se propõe a construir um modelo capaz de simplificar um conjunto complexo de decisões e produzir uma estratégia a partir de uma base de dados pública e internacional sobre o Transtorno do Espectro Autista (TEA), permitindo desenvolver um complemento no seu diagnóstico e possível tratamento.*

1. Introdução

Desde maio de 2013 os psicólogos e psiquiatras estão utilizando os critérios de avaliação do Manual de Diagnóstico e Estatístico de Transtorno Mentais (conhecido como DSM-5), elaborado pela *American Psychiatric Association* [Speaks 2019]. Os casos anteriormente analisados, utilizando o DSM-IV, que recebiam o diagnóstico de transtorno autista, transtorno de *Aspeger* ou transtorno de desenvolvimento penetrante, que não seja especificado de outra forma, passam a ser nomeados de transtorno do espectro autista, podendo ainda ser caracterizado em 3 (três) diferentes níveis de gravidade. Os sintomas desses transtornos representam um continuo único de prejuízos às pessoas portadoras, com intensidades variáveis, que vão de leve a grave, nos domínios de comunicação social e de comportamentos restritivos e repetitivos. Este estudo destaca que o transtorno do espectro do autismo (TEA), refere-se a uma ampla gama de condições caracterizadas por desafios com habilidades sociais, comportamentos repetitivos, fala e comunicação não-verbal [APA 2014].

O autismo afeta cerca de 1 em 59 crianças hoje nos Estados Unidos. Dessa forma, estima-se que o Brasil, com seus 200 milhões de habitantes, possua cerca de 3 milhões de

autistas. O autista possui um conjunto distinto de traços comportamentais. As maneiras pelas quais as pessoas com autismo aprendem, pensam e resolvem problemas podem variar de muito qualificadas a severamente desafiadas pela situação [One and Two 2019].

Tendo em vista a problemática de diagnosticação do autismo, o presente trabalho propõe um método para auxiliar especialistas na busca por um diagnóstico junto as informações necessárias cedidas pelos responsáveis, acelerando o procedimento.

O algoritmo é baseado em árvore de decisão, visando a classificação das características comportamentais em crianças, com um determinado grau de autismo ou não. É importante destacar a ideia principal, que visa apenas fornecer um auxílio com a previsão ao diagnóstico. Caso o resultado seja positivo significa que a criança pode apresentar algum grau de autismo e, portanto, o especialista pode fazer uso dos resultados desta aplicação como base para o laudo final.

O artigo está organizado como segue. Na seção 2 descreve-se os trabalhos relacionados à classificação e TEA. Na seção 3 é descrita toda estrutura da base de dados utilizada, apontando os atributos utilizados. A seção 4 aborda toda a fundamentação teórica do trabalho. Na seção 5 são apresentados os resultados obtidos, mostrando o gráfico de importância das variáveis utilizadas no modelo. Por fim, a seção 6 conclui o trabalho e sugere aplicações futuras.

2. Trabalhos relacionados

Em Brito & Fernandes (2019) foi proposto o uso de redes neurais para classificar crianças com presença ou não de TEA. Utilizaram uma base de dados com 259 amostras e com as estratégias de *cross validation* e *k-fold* para auxílio no treinamento. Os dois modelos comportaram-se bem mas os autores deram como inconclusivo os seus resultados [One and Two 2019].

Processamento de imagens e algoritmos de aprendizado de máquina foram empregados por Liu et al. 2016 para classificar crianças com autismo. Os dados foram de rosto de crianças junto com um conjunto de dados do movimento ocular dessas, por fim obtiveram um acurácia de 88,51% nos resultados [Liu et al. 2016].

3. Base de dados

A base de dados utilizada neste trabalho, disponível em [Thabtah 2017], conta com um total de 292 observações e 21 atributos. Dentre esses atributos, tem 10 variáveis comportamentais que são descritas na tabela um.

4. Fundamentação teórica

Nesta seção é mostrado os principais métodos que foram utilizados para a caracterização das variáveis em estudo. Destacam-se aqui o método de classificação de árvore de decisão e a função de avaliação *feature importance*.

4.1. Árvore de decisão

Pelo fato das variáveis utilizadas no modelo serem binárias (sim ou não), o modelo de árvore de decisão foi o mais adequado para o trabalho. Pode ser definido como uma estrutura de dados que determina uma classe como um nó folha, um nó decisão que

Variável	Descrição
A1_Score	Alta percepção em baixos ruídos que geralmente outros não percebem
A2_Score	Maior concentração na visão do todo em comparação a pequenos detalhes
A3_Score	Facilidade de comunicação com várias pessoas diferentes ao mesmo tempo
A4_Score	Facilidade de fazer múltiplas tarefas simultaneamente
A5_Score	Dificuldade de manter uma conversa com seus colegas
A6_Score	Facilidade em manter conversas informais
A7_Score	Dificuldade de percepção de intenções e sentimentos em histórias
A8_Score	Dificuldade de brincar com a imaginação fugindo da realidade
A9_Score	Facilidade de reconhecer sentimentos a partir de expressões faciais alheias
A10_Score	Dificuldade de fazer novas amizades
result	Soma das caracteristicas comportamentais avaliadas relacionados ao TEA.
Class/ASD	Decisão do algoritmo sobre a criança pertencer ou não ao espectro autista

Tabela 1. Variáveis comportamentais

contém algum teste sobre um atributo e, a cada resultado, uma aresta para uma subárvore [Rezende et al. 1999].

Este algoritmo tem algumas vantagens, como reduzir a complexidade do problema em regiões de decisão. É possível aproximar espaços de alta-dimensionalidade em vários níveis da árvore e conjunto de dados com muitas características. Permite-se ainda estimar um agrupamento menor desses em cada nó da árvore para testar com outros subconjuntos para melhorar a performance do algoritmo. Porém, com grandes bases de dados,os erros podem se acumular ao longo dos níveis da árvore de decisão e não podendo otimizar precisão e eficiência ao mesmo tempo [Safavian and Landgrebe 1991].

4.2. *Feature importance*

O método *feature importance* trata-se do processo de busca das melhores variáveis que descrevem o modelo. A Equação 1 destaca o comportamento dessa medida, com base nas técnicas de *impureza de Gini* ou *redução média das impurezas* que calcula a importância de cada característica de entrada no resultado do modelo [Safavian and Landgrebe 1991]. Com base nisso, é calculado pela soma das frequências das classes sobre o número total delas, como segue:

$$G = \sum_{j=1}^{J} P_j(1 - P_j), \tag{1}$$

em que *G* varia de (0,1] e representa o grau de importância da variável *j* no modelo , *J* é a quantidade de classes e *P* representa a frequência de uma categoria.

Para uma árvore de decisão, as classes são dividas em nós que são características que classificam as amostras do conjunto de dados. Desse modo, seja $\hat{\phi}_j(t)$ a frequência de uma classe *j* em um nó *t*. A impureza de Gini aplicada ao nó *t* é definida como [Ishwaran 2015]:

$$\hat{\Gamma}(t) = \sum_{j=1}^{J} \hat{\phi}_j(t)(1 - \hat{\phi}_j(t)). \tag{2}$$

Logo, cada nó da árvore é avaliado a partir dessa métrica. Com isso, depois de obtido os valores das impurezas de todas as classes, a árvore é rearranjada com base na categoria que possuir o menor resultado, isso irá se repetir até a última classe da base de dados.

5. Resultados e testes

Ao aplicar o algoritmo de Árvore de decisão à base de dados, foi possível observar a taxa de erros, a performance e extrair as características mais influentes na decisão do modelo.

5.1. Matriz de Correlação

Para a construção de um modelo consistente, faz-se necessário a seleção de variáveis descorrelacionadas entre si, pois dadas duas variáveis análogas, há uma redundância de informações que poderá afetar o desempenho do modelo. Dessa forma, para analisar a correspondência entre as variáveis, desenvolveu-se a matriz de correlação como visto na Figura 1.

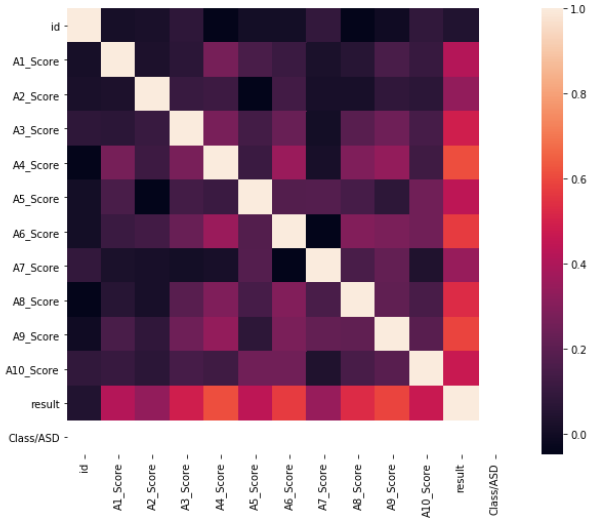

Figura 1. Matriz de Correlação

Em seguida, pode-se observar uma correlação de no máximo que 0,35 em um índice no intervalo absoluto [0, 1], das variáveis de entrada entre si, e um índice mais alto na correlação entre as variáveis de entrada e a variável resposta. Isso indica que as variáveis de entrada não são correlacionadas entre si, e possuem correlação relevante com a variável resposta.

5.2. Avaliação das características

De modo a entender como o modelo interpreta a decisão dos dados, deve-se atentar à *feature importance* ilustrada na Figura 2. Esta descreve as variáveis *A4-Score* e *A10-Score* como tendo maior importância em relação às demais. A variável *A6-Score* possui ao final uma baixa influência no resultado do teste.

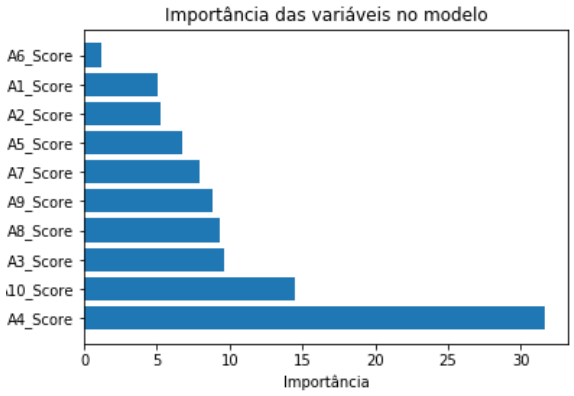

Figura 2. Importância das variáveis no modelo

5.3. Modelo e Análise dos dados

Ao submeter a base de dados ao algoritmo de Árvore de Decisão na intenção de prever a variável *Class/ASD*, obtém-se a descrição do modelo de dados como definido na Figura 3. O modelo ilustrado relaciona a classificação obtida pela variável *result* distribuída pela frequência de casos de forma a indicar que os indivíduos que atendem à 7 ou mais características esperadas da TEA necessariamente foram clinicamente diagnosticados com o transtorno.

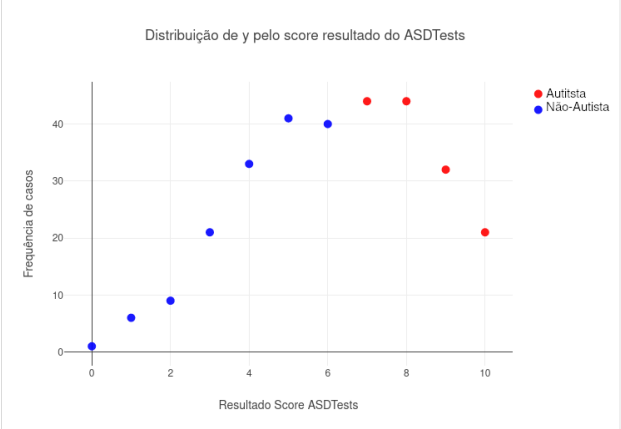

Figura 3. Modelo de análise dos dados

5.4. Avaliação da Performance do Modelo

Para evitar que o modelo crie um viés às condições iniciais de implementação, utilizou-se uma validação cruzada do tipo *k-fold* com $k = 5$. Além deste tratamento, é necessário avaliar a quantidade de falso-positivos e falso-negativos em uma amostra não observada no treinamento do modelo. Para tal utilizou-se a Matriz de Confusão descrita na Tabela 2. Tomando uma amostragem de 20% dos indivíduos separados para teste do modelo, repetindo o experimento k vezes de forma com que toda a população possa ser testada. A média de falso-positivos foi de 7,692% enquanto a média de falso-negativos obtida foi de 3,704%.

Rótulo Real	Autismo	26	1
	Não-Autismo	2	27
		Autismo	Não-Autismo
		Rótulo Previsto	

Tabela 2. Matriz de Confusão

6. Conclusão

O algoritmo proposto atingiu uma boa análise da base de dados sobre TEA para auxiliar o diagnóstico por parte dos especialistas. Com este trabalho é possível medir que as variáveis mais discriminantes que apontam ao diagnóstico de TEA consistem na facilidade de fazer múltiplas tarefas simultaneamente (A4) e na dificuldade de fazer novas amizades (A10), sendo relacionadas pela ausência e presença destes traços comportamentais, respectivamente. Ainda é possível concluir a partir do modelo de árvore de decisão que o indivíduo que apresenta correspondência em sete ou mais dos dez traços comportamentais propostos pelo questionário possui fortes indícios de possuir o transtorno em algum espectro do autismo.

Futuramente este trabalho se propõe a aplicar o modelo em outras bases de dados, adequando seus hiperparâmetros para que estes se tornem mais otimizados. Além de comparar este modelo com outros algoritmos de aprendizagem de máquina para avaliar sua performance.

Referências

APA, A. P. A. (2014). *DSM-5 - Manual Diagnóstico e Estatístico de Transtornos Mentais*. Artmed.

Ishwaran, H. (2015). The effect of splitting on random forests. *Machine Learning*, 99(1):75–118.

Liu, W., Li, M., and Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8):888–898.

One, A. and Two, A. (2019). Análise de desempenho com redes neurais artificiais, arquiteturas MLP e RBF para um problema de classificação de crianças com autismo. *iSys-Brazilian Journal of Information Systems*, 12(1).

Rezende, S. O., Monard, M. C., and Carvalho, A. C. P. d. L. (1999). Sistemas inteligentes para engenharia: pesquisa e desenvolvimento. *Anais III Workshop de Sistemas Inteligentes para Engenharia*.

Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674.

Speaks, A. (2019). Autism speaks. `https://www.autismspeaks.org/`. Acesso em: 05/04/2019.

Thabtah, F. F. (2017). Autistic spectrum disorder screening data for children data set. `https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++`. Acesso em: 08/07/2019.