# Sistema de Predição de Diagnóstico do Transtorno do Espectro Autista Utilizando o Classificador Ingênuo de Bayes

Camila Barbosa Vieira
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
cbv2@cin.ufpe.br

Edson José Araújo Pereira Júnior
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
ejapj@cin.ufpe.br

Marcelo Cristian da Silva Brito
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
mcsb2@cin.ufpe.br

Maria Vitória Soares Muniz
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
mvsm3@cin.ufpe.br

***Abstract*—Este trabalho visa a criação de um sistema de predição da incidência de indivíduos dentro do Transtorno do Espectro Autista utilizando o Classificador Ingênuo de Bayes. Para isso, recorre a uma base de dados com traços comportamentais e características individuais que se mostraram eficazes na identificação do TEA.**

***Index Terms*—transtorno do espectro autista, predição, classificador ingênuo de bayes, Naive Bayes**

## I. Introdução

O transtorno do espectro autista (TEA) é responsável, em diferentes graus, por alterações no comportamento e dificuldades com comunicação e interação social. Segundo dados do Centro de Controle de Doenças dos Estados Unidos, a prevalência do TEA de 2014 para 2020 sofreu um aumento de 9%.

No entanto, o crescimento dessa taxa é algo positivo, uma vez que indica o aumento da compreensão sobre o transtorno e a adoção de métodos mais apropriados para o diagnóstico. Fatores essenciais para a melhora da qualidade de vida e da autonomia da pessoa.

Atualmente, para uma criança ou adolescente ser diagnosticada com autismo, ela deve ser avaliada por uma equipe que inclui pediatra, psicólogo, psiquiatra, fonoaudiólogo e neuropsicólogo. Por não existir um exame laboratorial, é necessário observar o comportamento, o histórico do paciente e realizar a exclusão de outras doenças. Quando em adultos, os sinais comumente são confundidos com timidez ou outros transtornos, por isso, a dificuldade em emitir um laudo correto é ainda maior.

Desse modo, o desenvolvimento de um método de classificação eficiente, o qual é capaz de prever os pacientes que devem ser diagnosticados com autismo, torna-se essencial. Tendo em vista a importância da identificação do transtorno e a complexidade de efetuar tal tarefa. Isso é possível devido à evolução do machine learning, da adoção do Classificador Ingênuo de Bayes e dos dados disponibilizados no repositório de aprendizagem de máquina da UCI.

## II. Objetivos

Esse projeto acadêmico visa à construção de um sistema eficaz, que consiga identificar de maneira rápida casos de autismo, a fim de dar o suporte necessário a uma parte significativa da população, visto que, segundo a OMS, uma em cada 160 crianças tem transtorno do espectro autista. Para tanto, esse projeto tem como objetivo principal a criação de um algoritmo de predição de casos do transtorno do espectro autista (TEA).

O sistema colocado em questão tem a finalidade de identificar características e padrões associados ao TEA utilizando uma base de dados do repositório UCI. Dessa forma, dados coletados de indivíduos acometidos desse transtorno serão analisados, utilizando o Classificador Ingênuo de Bayes, a partir de recursos e bibliotecas de aprendizagem de máquina, para identificar esses padrões e, assim, poder inferir um diagnóstico positivo ou negativo para tal caso.

Em vista disso, esse modelo será uma ferramenta capaz de ajudar profissionais da saúde a inferirem diagnósticos mais rápidos e precisos acerca do TEA, a fim de ajudar o paciente, dando à família desse a oportunidade de dar o tratamento adequado, visando o seu pleno desenvolvimento e inserção na sociedade.

## III. Justificativa

Em razão dos hodiernos avanços da ciência e da medicina, hoje temos o conhecimento de diversos transtornos e condições

erroneamente identificados há poucos anos atrás como, por exemplo, o transtorno do espectro autista. Porém, exatamente por ser tão recente a descoberta desse transtorno, seu diagnóstico ainda não é tão claro, gerando sobre o autismo uma grande incerteza prática e teórica por conta também da cada vez mais presente inconsistência de fatos sobre o assunto.

Portanto, torna-se evidente a grande importância da existência de um sistema que identifique uma pessoa de qualquer faixa etária com tal transtorno, levando em consideração os aspectos mais relevantes do TEA com o objetivo de tornar seu diagnóstico mais prático, eficaz e acessível.

## IV. Base de dados

A fim de desenvolver um sistema de predição utilizaremos três bases de dados, montadas a partir das respostas de pacientes ao questionário "Autism Quotient 10". Os dados foram coletados a partir de um estudo realizado por Tabtah, F [1] e também com a ajuda de um aplicativo chamado "ASD Test", utilizado para o diagnóstico do Transtorno do Espectro Autista (TEA).

TABLE I
Descrição das Bases de Dados

| Nome da Base de Dados | Atributos | Instâncias |
|---|---|---|
| Autistic Spectrum Disorder Screening Data for Children Data Set | 21 | 292 |
| Autistic Spectrum Disorder Screening Data for Adolescent Data Set | 21 | 104 |
| Autism Screening Adult Data Set | 21 | 704 |

As bases de dados possuem um conjuntos de 20 atributos que são utilizados para a previsão, na tabela abaixo podemos ver o que cada atributo significa:

TABLE II
Descrição dos Atributos da Base de Dados

| Id | Atributo | Tipo |
|---|---|---|
| 1 | Age | Intervalo da idade em anos (12-16 years, 12-15 years |
| 2 | Gender | Valor M: Masculino. Valor F: Feminino |
| 3 | Ethnicity | Etnia do paciente (Hispânico, Negro, Branco-Europeu, 'Oriente Médio', 'Sul Asiático', Outros, Latino, Asiático) |
| 4 | Born with jaundice | Valor 1: nasceu com icterícia. Valor 0: não nasceu com icterícia |
| 5 | Family member with PDD | Se algum membro imediato da família tem um PDD |
| 6 | Who is completing the test | Quem realizou o teste (Pai, Parente, Próprio, 'Profissional de Saúde', Outros) |
| 7 | Country of residence | O país em que o paciente mora em formato de texto |
| 8 | Used the screening app before | Valor 1: usou um aplicativo de triagem. Valor 0: não usou um aplicativo de triagem |
| 9 | Screening Method Type | O tipo de método de triagem escolhido com base na idade (Valor 0: criança, Valor 1: criança, Valor 2: adolescente, Valor 3: adulto) |
| 10-19 | Question Answer | Booleano de resposta da pergunta com base no método de triagem usado |
| 20 | Screening Score | A pontuação final obtida com base no algoritmo de pontuação do método de triagem utilizado |

## V. Análise Explorátoria dos Dados

Antes de utilizar os dados para a construção do Classificador Ingênuo de Bayes, iremos realizar uma Análise Explorátoria dos Dados. Para isso iremos utilizar algumas bibliotecas do Python e seguiremos os seguintes passos:

1) **Entendendo Base de Dados**: primeiramente, será necessário entender um pouco mais sobre as variáveis presentes no banco de dados e como elas estão dispostas.
    a) **Tipo das variáveis**: primeiramente, analisamos os tipos de cada variável e percebemos que todas são categóricas, algo que teremos que mudar, posteriormente, para construir o modelo com o classificador.

TABLE III
VARIÁVEIS E SEUS TIPOS

| Variável | Tipo |
|---|---|
| A1_Score | object |
| A2_Score | object |
| A3_Score | object |
| A4_Score | object |
| A5_Score | object |
| A6_Score | object |
| A7_Score | object |
| A8_Score | object |
| A9_Score | object |
| A10_Score | object |
| age | object |
| gender | object |
| ethnicity | object |
| jundice | object |
| austim | object |
| country | object |
| used_app_before | object |
| result_score | object |
| age_desc | object |
| who_answer | object |
| class | object |

b) **Presença de valores ausentes**: depois de utilizar alguns métodos da biblioteca pandas conseguimos visualizar a quantidade de valores ausentes em cada coluna da base de dados. Apenas três colunas tinham dados faltantes: ***age, ethnicity e who_answer***.

TABLE IV
VARIÁVEIS E VALORES AUSENTES

| Variável | Valores ausentes |
|---|---|
| age | 6 |
| ethnicity | 144 |
| who_answer | 144 |

c) **Presença de Outliers**: ao visualizar a dispersão dos dados através de box-plots, percebemos a presença de alguns dados muito diferentes na coluna de idade. Alguns não eram muito discrepantes, mas tinha uma instância com mais de 300 anos. Podemos visualizar melhor atráves da imagem abaixo [Fig.1]:

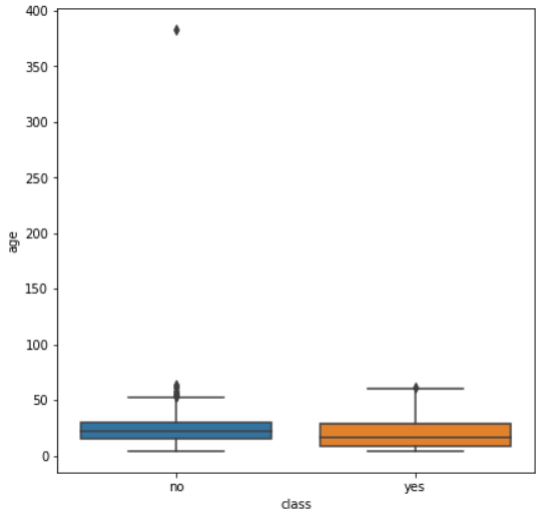



Fig. 1. Distribuição da coluna age

d) **Verificando balanceamento da classe**: por últimos tentamos entender como está a disposição dos dados na coluna ***class***, um atributo bem importante utilizado para a classificação. Percebmos que ela é bem desbalanceada, temos mais instância com diagnóstico negativo para o TEA do que positivo.

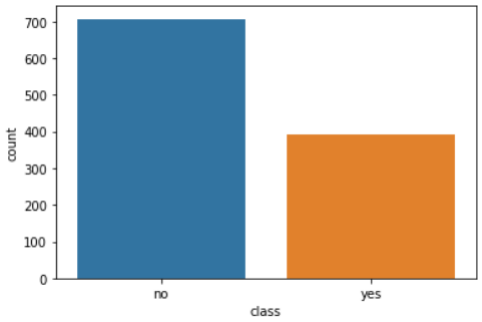



Fig. 2. Distribuição da classe

2) **Limpeza de Dados**: iremos processar e tratar os dados, procurando por inconsistências, faltas, valores nulos e possíveis outliers. Neste ponto, decidimos o que fazer com entradas faltantes, ou discrepantes. Para isso, iremos utilizar a Biblioteca Pandas do Python.
   a) **Valores ausentes**: Como temos muitos dados categóricos decidimos resolver esses problema da seguinte forma:
      - **Coluna age**: substituímos os valores ausentes pelos mais próximos. A base de dados atual foi resultado da concatenação de três bases que eram divididas por idade. Com isso, chegamos a conclusão que era mais lógico fazer esse tratamento.
      - **Coluna ethnicity e who_answer**: para esses atributos utilizamos a moda de cada coluna e substituímos os valores ausentes por esse valor.

b) **Mudança dos tipos das variáveis**: como foi falado anterior, todos os nossos dados são categóricos e isso dificulta bastante e impossibilita algumas análises. Com isso, decidimos fazer um mapeamento para transformar os dados para o tipo inteiro.

c) **Outliers**: encontramos Outliers apenas na coluna de age e para fazer o tratamento utilizamos o intervalo interquartil e removemos as instâncias que tinham esses dados discrepantes.

3) **Análise Estatística**: Com a finalidade de definir o perfil do paciente com diagnóstico positivo, filtramos o dataframe e analisamos algumas características. Verificamos a distribuição destes pacientes por idade, gênero, nascidos com icterícia, ter familiar com diagnóstico positivo e etnia. Por exemplo, na figura 3 podemos perceber que temos um grande número de crianças e adolescentes com diagnóstico positivo para o TEA. Na figura 4 podemos notar que temos mais de 18% dos pacientes que nasceram com icterícia estão dentro do espectro autista.

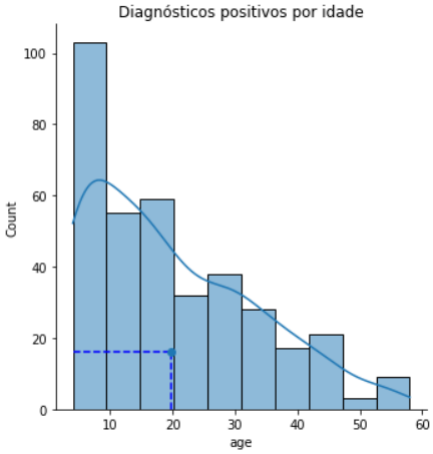



Fig. 3.


Fig. 4.


Fig. 5.

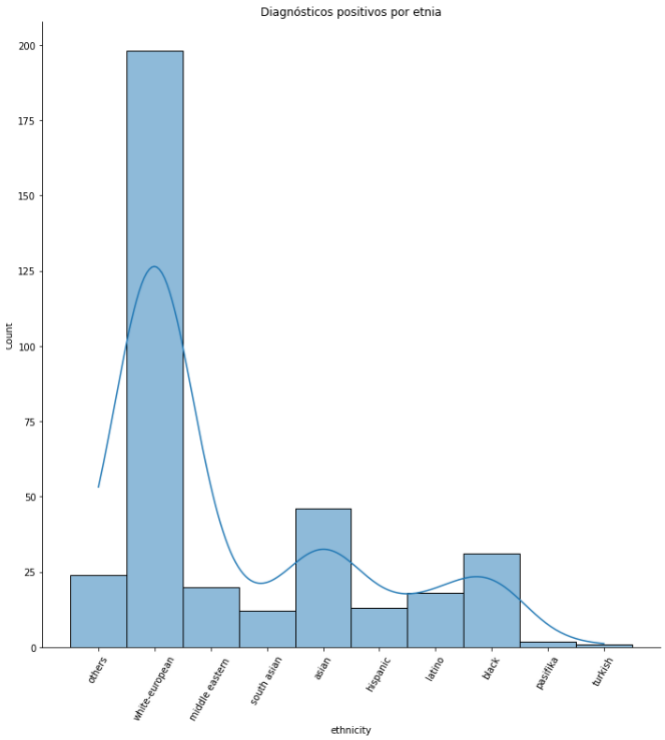



Fig. 6.

4) **Representação Gráfica**: com o auxílio da biblioteca matplotlib iremos plotar gráficos para entender melhor como está a distribuição e dispersão dos dados. Por exemplo, a disposição de pacientes com TEA em diferentes países.

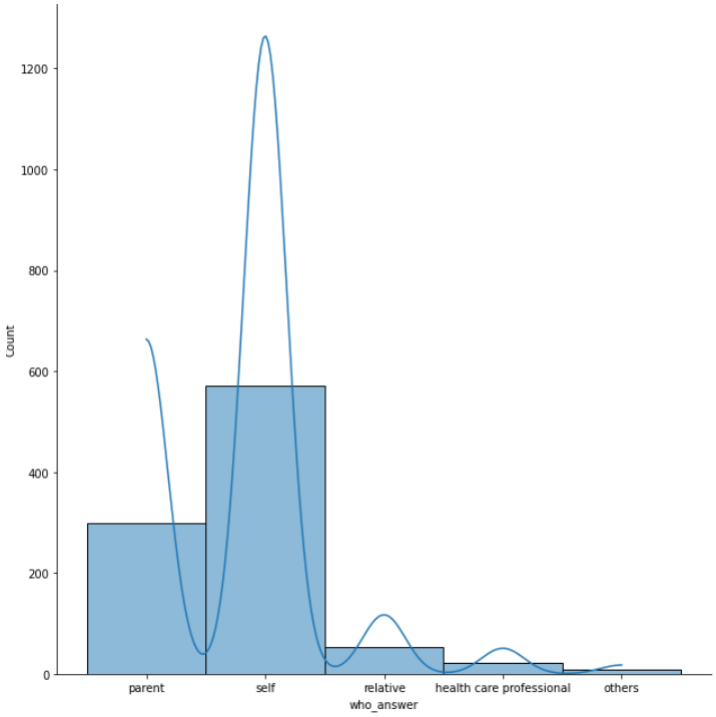
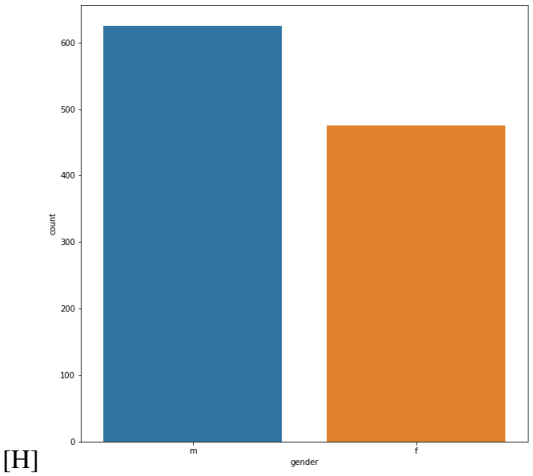



Fig. 7. Distribuição de quem realizou o teste




[H]

Fig. 10. Distribuição por faixa etária

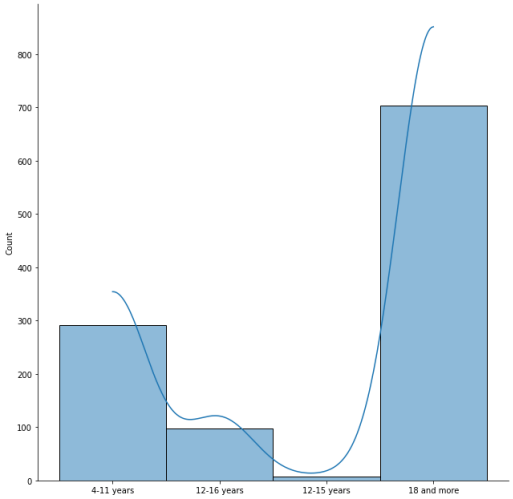



Fig. 9. Distribuição por gênero

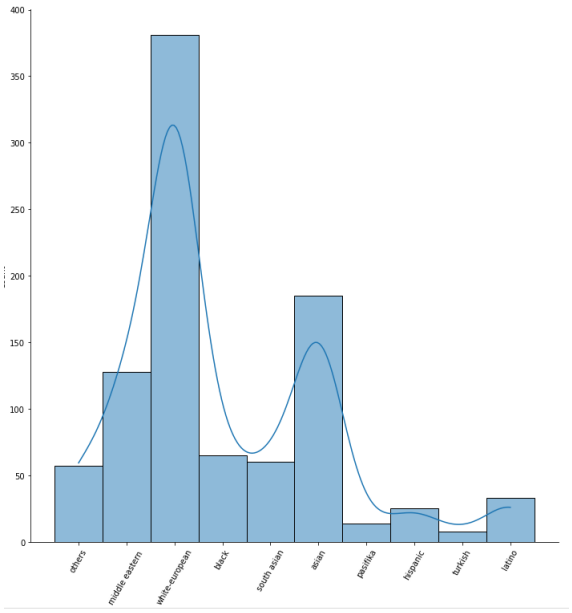



Fig. 8. Distribuição das etnias

### *A. Implemnetação do Classificador Ingênuo de Bayes*

O classificador Naive Bayes é um algoritmo que se baseia nas descobertas de Thomas Bayes para realizar predições em aprendizagem de máquina. A principal característica do algoritmo, e também o motivo de receber "naive" (ingênuo) no nome, é que ele desconsidera completamente a correlação entre as variáveis (features), ou seja, que elas são independentes entre si.

o Naive Bayes é um dos modelos mais conhecidos a aplicar o conceito de probabilidade, como o nome indica, faz uso do teorema de Bayes como princípio fundamental, sendo definido como:

$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)} \tag{1}$$

Temos que:

- P(A/B): probabilidade de um evento A acontecer dado que um evento B ocorreu;

- P(B/A): probabilidade de um evento B acontecer dado que um evento A ocorreu;
- P(A): probabilidade de um evento A ocorrer;
- P(B): probabilidade de um evento A ocorrer.

Esse algoritmo de classificação é robusto para previsões em tempo real, ainda mais por precisar de poucos dados para realizar a classificação. Além disso, tem enumeras aplicações, como na área de análise de textos e também no diagnóstico médico.

Vale ressaltar que, na análise de grande parte dos fenômenos da natureza, percebe-se um comportamento normal, ou seja, fenômenos, os quais possuem distribuições de probabilidades que podem ser muito bem descritas por uma Distribuição Gaussiana. Dessa forma, esse trabalho visa a tratar as características, representadas como variáveis aleatórias, como sendo distribuídas normalmente. Sendo assim, a utilização do Classificador Naive Bayes em Distribuições Gaussianas é descrita de acordo com a equação (2).

### *B. Treinamento do Modelo*

Para medir o desempenho real do modelo criado, é necessário que realizemos testes com ele, utilizando dados diferentes dos que foram apresentados em sua criação.

Com esta finalidade, após a realização do pré-processamento, iremos separar a totalidade dos dados históricos existentes em dois grupos, sendo o primeiro responsável pelo treino do modelo, e o segundo por realizar os testes.

Para isso utilizaremos a técnica de Cross-validation com método K-Fold em que consiste em dividir a base de dados de forma aleatória em K subconjuntos (em que K é definido previamente) com aproximadamente a mesma quantidade de amostras em cada um deles. A cada iteração, treino e teste, um conjunto formado por K-1 subconjuntos são utilizados para treinamento e o subconjunto restante será utilizado para teste gerando um resultado de métrica para avaliação.

### *C. Definição de Métricas*

Durante o processo de construção de um modelo precisamos medir seu desempenho de acordo com o objetivo pela o qual ele foi criado. Existem funções matemáticas que nos ajudam a avaliar a capacidade de erro e acerto do nosso modelo, com isso iremos utilizar:

1) **Acurácia**: é uma indicador que nos fornece a performace geral do modelo, ele indica a quantidade de acertos do nosso modelo divido pelo total da amostra, independente das classes. Ele pode ser calculado da seguinte forma:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Onde:

- **Verdadeiro Positivo (TP)**: são as observações que são positiva e o modelo consegue classificar corretamente;
- **Verdadeiro Negativo (TN)**: são as observações que o modelo previu como negativas e realmente eram negativas, ou seja, o modelo classificou corretamente;
- **Falso Positivo (FP)**: são as observações que o modelo previu como positivas, mas na realidade eram negativas;
- **Falso Negativo (FN)**: são as observações que o modelo identificou como negativas, mas eram positivas. Ou seja, as observações que o modelo estimou errado.

2) **Recall ou Sensibilidade**: Essa métrica avalia a capacidade do método de detectar com sucesso resultados classificados como positivos, ela pode ser calculada pela fórmula:

$$Sensibilidade = \frac{TP}{TP + FN} \tag{3}$$

3) **Especificidade**:avalia a capacidade do método de detectar resultados negativos, ela pode ser calculada pela fórmula:

$$Especificidade = \frac{TN}{TN + FP} \tag{4}$$

4) **Precisão**: também conhecida como Valor Preditivo Positivo (VPP), é a métrica que traz a informação da quantidade de observações classificadas como positiva que realmente são positiva, sendo calculada como:

$$Precisão = \frac{TP}{TP + FP} \tag{5}$$

5) **Matriz de confusão**: é a matriz quadrada em que se compara os verdadeiros valores de uma classificação com os valores preditos através de algum modelo. Sua diagonal é composta pelos acertos do modelo e os demais valores são os erros cometidos. Com isso, podemos entender se o modelo está favorecendo uma classe em detrimento da outra.

## VI. Experimentos e Resultados

### *A. Considerações Iniciais*

O modelo em questão visa predizer se um determinado paciente está ou não no espectro do TEA. Por se tratar de um espectro, existindo níveis mais graves e outros mais leves do transtorno, o sistema de predição poderia ter dificuldade de interpretar essas nuances. Outro fator relevante a ser considerado, é a seriedade das consequências de um falso negativo em comparação a um falso positivo. Devido a isso, a prioridade é a redução dos falso negativos.

### *B. Primeiro Experimento*

O primeiro experimento consiste no treinamento e avaliação da performance do modelo sem uso da seleção de feature. Os seguintes resultados foram obtidos:

TABLE V
SAÍDA DA MATRIZ CONFUSÃO: PRIMEIRO EXPERIMENTO

| | **Previsão de Diagnóstico Negativo** | **Previsão de Diagnóstico Positivo** |
|---|---|---|
| **Verdadeiro Negativo** | 273 | 6 |
| **Verdadeiro Positvo** | 8 | 151 |

TABLE VI
INDICADORES: PRIMEIRO EXPERIMENTO

| | **Precisão** | **Recall** | **f1-score** | **Suporte** |
|---|---|---|---|---|
| **Negativo (0)** | 0.97 | 0.98 | 0.97 | 279 |
| **Positvo (1)** | 0.96 | 0.95 | 0.96 | 159 |
| **Acurácia** | - | - | 0.97 | 438 |
| **Média Macro** | 0.97 | 0.96 | 0.97 | 438 |
| **Média Ponderada** | 0.97 | 0.97 | 0.97 | 438 |

### *C. Segundo Experimento*

Já o segundo experimento fez a seleção de features como meio de aumentar a performance. Isso ocorreu através da utilização do Random Forest - Algoritmo de Boruta, o qual retorna as melhores variáveis. Neste caso, foram selecionadas as colunas A1_Score a A10_Score, age e result_score. Os resultados obtidos podem ser vistos abaixo.

TABLE VII
SAÍDA DA MATRIZ CONFUSÃO: SEGUNDO EXPERIMENTO

| | **Previsão de Diagnóstico Negativo** | **Previsão de Diagnóstico Positivo** |
|---|---|---|
| **Verdadeiro Negativo** | 276 | 3 |
| **Verdadeiro Positvo** | 3 | 156 |

TABLE VIII
INDICADORES: SEGUNDO EXPERIMENTO

| | **Precisão** | **Recall** | **f1-score** | **Suporte** |
|---|---|---|---|---|
| **Negativo (0)** | 0.99 | 0.99 | 0.99 | 279 |
| **Positvo (1)** | 0.98 | 0.98 | 0.98 | 159 |
| **Acurácia** | - | - | 0.99 | 438 |
| **Média Macro** | 0.99 | 0.99 | 0.99 | 438 |
| **Média Ponderada** | 0.99 | 0.99 | 0.99 | 438 |

### *D. Terceiro Experimento*

Enquanto isso, o terceiro experimento busca o balanceamento de classes pelo método SMOTE, aumentando as instâncias de classes minoritárias por meio do uso de dados sintéticos. A seguir os resultados obtidos.

TABLE IX
SAÍDA DA MATRIZ CONFUSÃO: TERCEIRO EXPERIMENTO

| | **Previsão de Diagnóstico Negativo** | **Previsão de Diagnóstico Positivo** |
|---|---|---|
| **Verdadeiro Negativo** | 268 | 10 |
| **Verdadeiro Positvo** | 19 | 267 |

TABLE X
INDICADORES: TERCEIRO EXPERIMENTO

| | **Precisão** | **Recall** | **f1-score** | **Suporte** |
|---|---|---|---|---|
| **Negativo (0)** | 0.93 | 0.96 | 0.95 | 278 |
| **Positvo (1)** | 0.96 | 0.93 | 0.95 | 286 |
| **Acurácia** | - | - | 0.95 | 564 |
| **Média Macro** | 0.95 | 0.95 | 0.95 | 564 |
| **Média Ponderada** | 0.95 | 0.95 | 0.95 | 564 |

## VII. CONCLUSÕES E DISCUSSÕES

Analisando os resultados obtidos por este modelo, percebe-se que, mesmo sem o uso da seleção de features, a acurácia é de quase 97%. Como esperado, a base de dados escolhida conta com traços comportamentais e características individuais excepcionais para a identificação do Transtorno do Espectro Autista.

Dado o problema inicial, predizer a incidência de um indivíduo dentro do TEA, é de extrema importância reduzir o número de falsos negativos para que se possa haver melhora da qualidade de vida e da autonomia da pessoa por meio da adoção do tratamento adequado. Pelo o experimento um, houveram oito falsos negativos entre as 281 previsões de diagnóstico dos pacientes como fora do espectro.

A partir do momento que opta-se por realizar a seleção das características que realizam melhor performance por meio do Random Forest (algoritmo Boruta), a acurácia se aproxima dos 99%. E os casos de falso negativo ficam três entre 279 previsões de diagnóstico contrário.

Já o terceiro experimento, utilizando a técnica de SMOTE e dados sintéticos, apresenta uma acurácia de 95%. Uma queda considerável em comparação com os modelos anteriores. Chama ainda mais atenção o aumento dos casos de falso negativos, com um aumento para 19 casos entre as 287 previsões negativas.

Sendo assim, o objetivo de construir um sistema de predição que auxilie na identificação de casos de pacientes dentro do espectro do TEA foi alcançado. O modelo mais adequado para se adotar é o do segundo experimento, não só por ter a maior acurácia (quase 99%) e precisão, mas principalmente por ter a menor taxa de falsos negativos, medida através do recall (superior a 98%). Algo de extrema relevância dada a natureza da problemática.

## REFERENCES


[1] Thabtah, F. (2017, May). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. In Proceedings of the 1st International Conference on Medical and Health Informatics, 2017 (pp. 1-6). ACM

[2] Fadi Fayez Thabtah (2017), "Autistic Spectrum Disorder Screening Data for children", UCI — Machine Learning Reporitory. Disponível: http://encurtador.com.br/ghlrz

[3] Fadi Fayez Thabtah (2017), "Autistic Spectrum Disorder Screening Data for Adolescent Data Set", UCI — Machine Learning Reporitory. Disponível: http://encurtador.com.br/fgrN0

[4] Fadi Fayez Thabtah (2017), "Autism Screening Adult Data Set", UCI — Machine Learning Reporitory. Disponível: https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

[5] Mariano, Diego. Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e F-score, 2021. Disponível em: https://l1nq.com/N5q2c

[6] Scudilio, Juliana. Qual a melhor métrica para avaliar os modelos de Machine Learning?, 2020. Disponível em: https://www.flai.com.br/juscudilio/qual-a-melhor-metrica-para-avaliar-os-modelos-de-machine-learning/

[7] Raja, S. Masoodb, S , "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques", Procedia Computer Science, Volume 167, 2020, Pages 994-1004

[8] Braz, Eduardo. Cross Validation: Avaliando seu modelo de Machine Learning, 2019. Disponível em: https://medium.com/@edubrazrabello/cross-validation-avaliando-seu-modelo-de-machine-learning-1fb70df15b78

[9] Center for Disease Control and Prevention. Data Statistics on Autism Spectrum Disorder. Disponível em: https://www.cdc.gov/ncbddd/autism/data.html

[10] Organização Pan-Americana da Saúde. Transtorno do espectro autista. Disponível em: https://www.paho.org/pt/topicos/transtorno-do-espectro-autista

[11] The National Institute Of Mental Health. Autism Spectrum Disorder. Disponível em: https://encurtador.com.br/vxGV3

[12] Rosal, Iury. Engenharia de Features: Transformando dados categóricos em dados numéricos, 2021. Disponível em: https://medium.com/data-hackers/engenharia-de-features-transformando-dados-categ%C3%B3ricos-em-dados-num%C3%A9ricos-e5d3991df715

[13] Azank, Felipe. Dados Desbalanceados — O que são e como lidar com eles, 2020. Disponível em: https://medium.com/turing-talks/dados-desbalanceados-o-que-s%C3%A3o-e-como-evit%C3%A1-los-43df4f49732b