Sistema de Predição de Diagnóstico do TEA

Usos do Classificador Ingênuo de Bayes

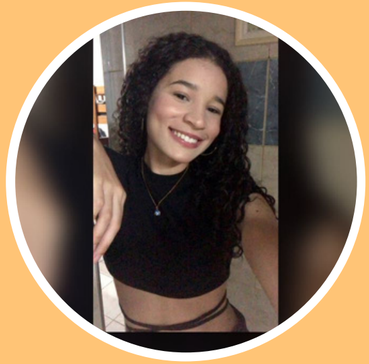
Equipe

Edson Júnior

ejapj

Marcelo Crístian

mcsb2



Maria Vitória

mvsm3

Camila Vieira

cbv2

Sumário

Contextualização

Problemas

O Transtorno do Espectro Autista (TEA) é responsável, em diferentes graus, por alterações no comportamento e por dificuldades com comunicação e interação social.

Grande equipe de médicos

Avaliação por uma equipe com pediatra, psicólogo, psiquiatra, fonoaudiólogo e neuropsicólogo para obter o diagnóstico.

Falta de exame laboratorial

Diagnóstico depende da observação do comportamento e do histórico do paciente.

Confusão com outros transtornos

A confirmação exige a exclusão de outras doenças. Sinais são comumente confundidos.

Objetivo

PRÓXIMOS PASSOS

1. Identificar características e padrões associados ao TEA

2. Utilizar o classificador ingênuo de Bayes para inferir um diagnóstico

3. Ajudar profissionais de saúde com resultados mais rápidos e precisos

Motivação

Reduzir a complexidade do diagnóstico

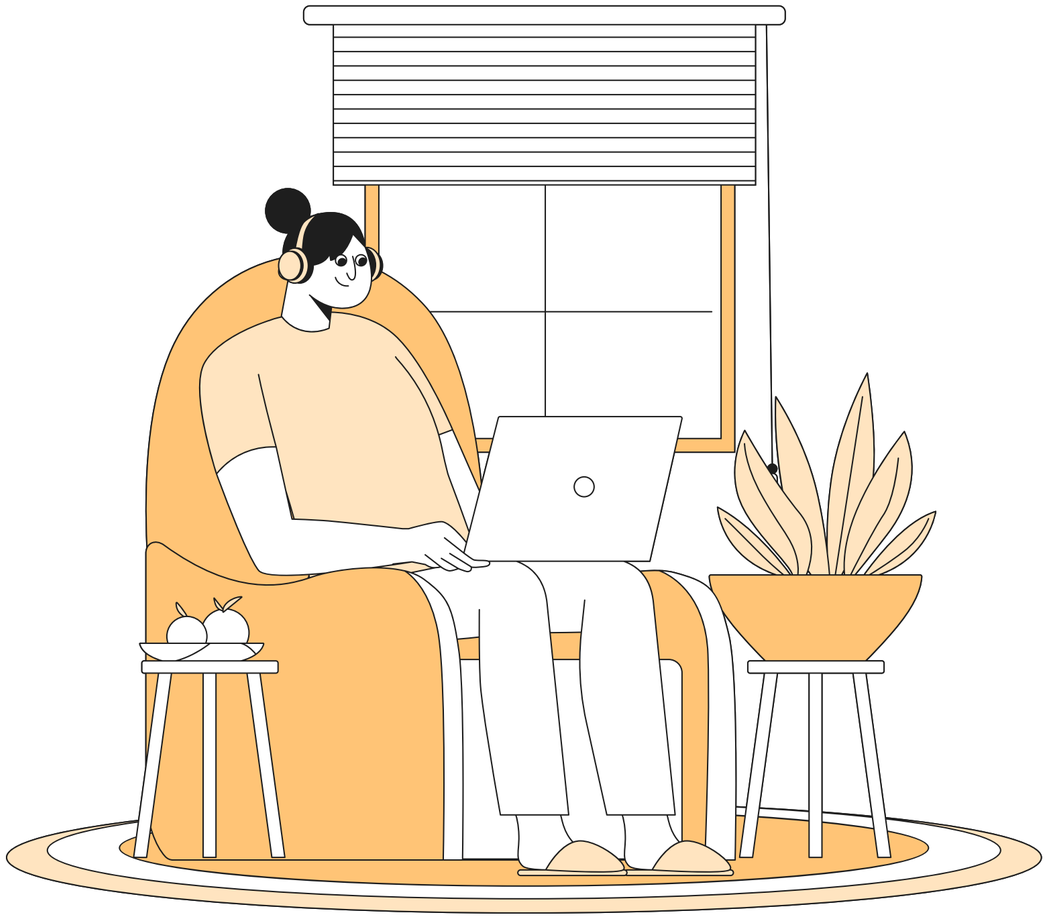
Graças a evolução do machine learning, podemos por meio do Classificador Ingênuo de Bayes fornecer um método simples, eficaz e acessível de classificação.

Oportunidade de tratamento adequado

O diagnóstico correto é essencial para a adoção dos tratamentos mais aconselhados para garantir o desenvolvimento do individuo.

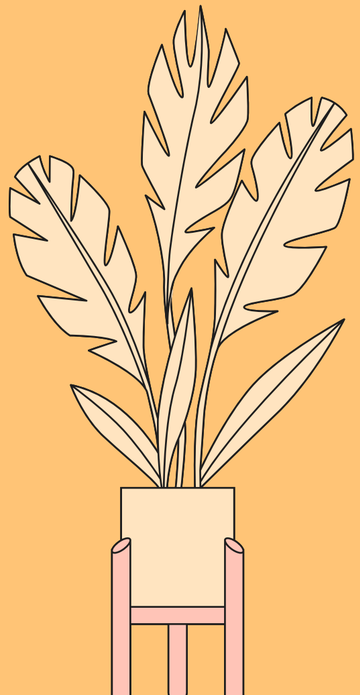
Melhorar a qualidade de vida dos pacientes

O aumento da compreensão sobre o transtorno e o aprimoramento do diagnóstico contribuem no aumento da autonomia e da qualidade de vida.



Análise Exploratória dos Dados

Análise Exploratória dos Dados



1. Base de dados

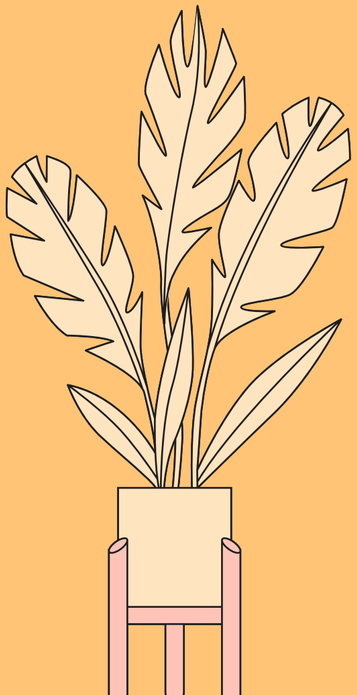
2. Entendendo base de dados

3. Limpeza de dados

4. Representação gráfica

Base de Dados

Utilizamos três bases de dados, montadas a partir das respostas de pacientes ao questionário “Autism Quotient 10”

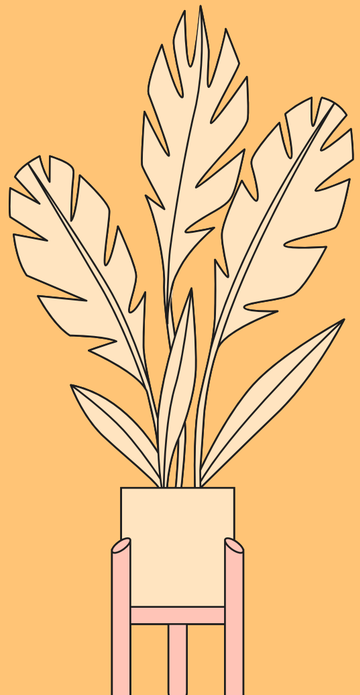


Descrição dos atributos

Atributo	Tipo
Age	Intervalo da idade em anos (12-16 years, 12-15 years)
Gender	Valor M: Masculino. Valor F: Feminino
Ethnicity	Etnia do paciente
Jundice	Valor 1: nasceu com icterícia. Valor 0: não nasceu com icterícia
Family member with PDD	Se algum membro imediato da família com PDD
Question Answer (1-10)	Booleano de resposta da pergunta com base no método de triagem usado

Atributo	Tipo
Who_answer	Quem realizou o teste
Country	O país em que o paciente mora
Used_app_before	Valor 1: usou um aplicativo de triagem. Valor 0: não usou um aplicativo de triagem
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)
Screening Score	A pontuação final obtida com base no algoritmo de pontuação do método de triagem utilizado

Entendendo Base de Dados



1. Tipos das variáveis

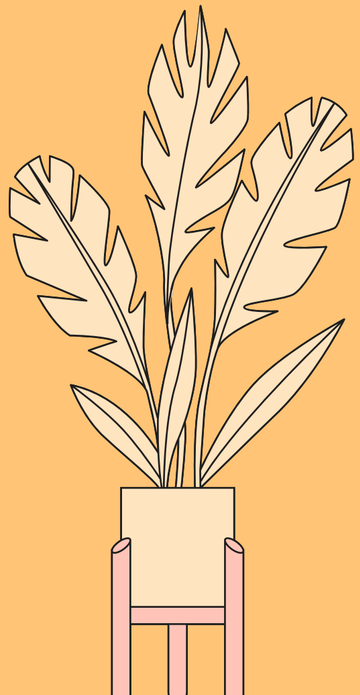
2. Presença de valores ausentes

3. Presença de Outliers

4. Verificando o balanceamento da classe

Entendendo Base de Dados

Primeiramente, analisamos os tipos de cada variáveis e percebemos que todas são categóricas.

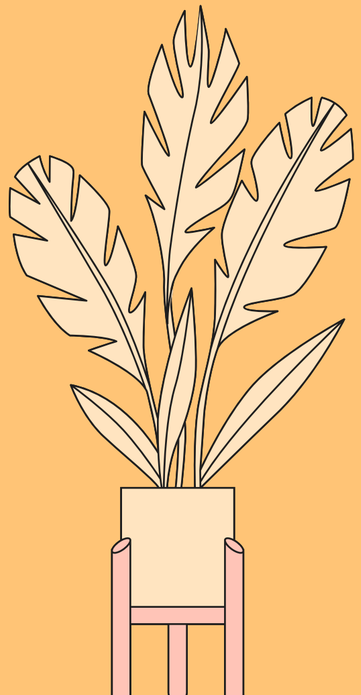


Tipo das variáveis

0	id	1100 non-null	object
1	A1_Score	1100 non-null	object
2	A2_Score	1100 non-null	object
3	A3_Score	1100 non-null	object
4	A4_Score	1100 non-null	object
5	A5_Score	1100 non-null	object
6	A6_Score	1100 non-null	object
7	A7_Score	1100 non-null	object
8	A8_Score	1100 non-null	object
9	A9_Score	1100 non-null	object
10	A10_Score	1100 non-null	object
11	age	1094 non-null	object
12	gender	1100 non-null	object
13	ethnicity	956 non-null	object
14	jundice	1100 non-null	object
15	austim	1100 non-null	object
16	country	1100 non-null	object
17	used_app_before	1100 non-null	object
18	result_score	1100 non-null	object
19	age_desc	1100 non-null	object
20	who_answer	956 non-null	object
21	class	1100 non-null	object

Entendendo Base de Dados

Após utilizar alguns métodos do pandas, conseguimos visualizar a quantidade de valores ausentes em cada coluna.



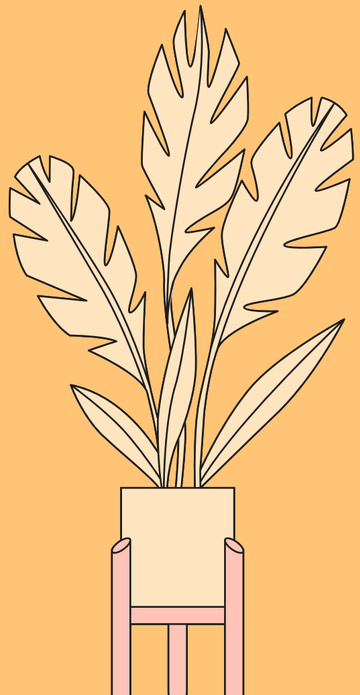
Presença de valores ausentes

id	0	A10_Score	0
A1_Score	0	age	6
A2_Score	0	gender	0
A3_Score	0	ethnicity	144
A4_Score	0	jundice	0
A5_Score	0	austim	0
A6_Score	0	country	0
A7_Score	0	used_app_before	0
A8_Score	0	result_score	0
A9_Score	0	age_desc	0
A10_Score	0	who_answer	144
		class	0

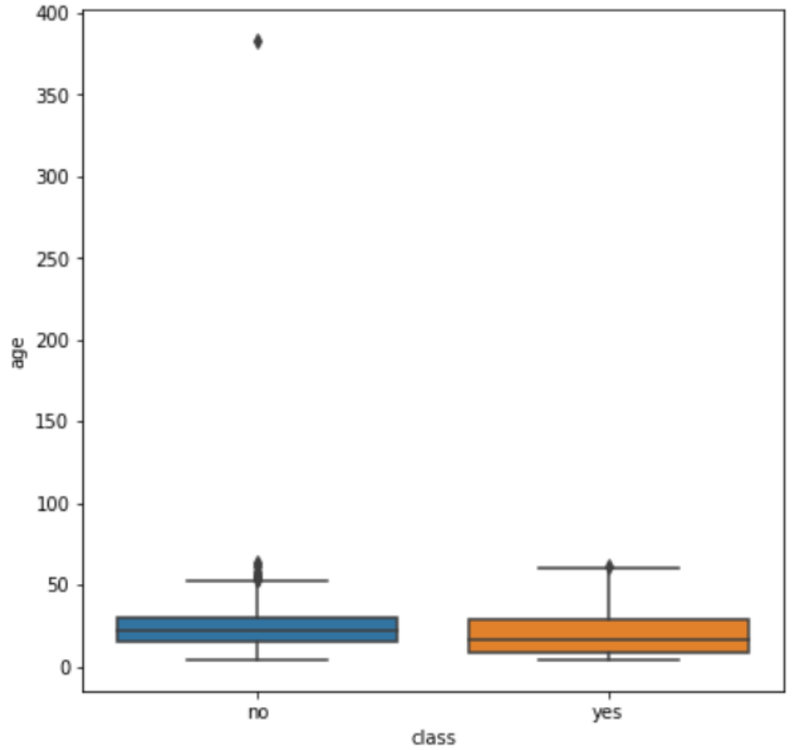
Apenas as colunas de age, ethnicity e who_answer apresenta valores ausentes

Entendendo Base de Dados

Ao visualizar a dispersão dos dados através de box-plots, percebemos a presença de alguns dados muito diferentes na coluna de idade.

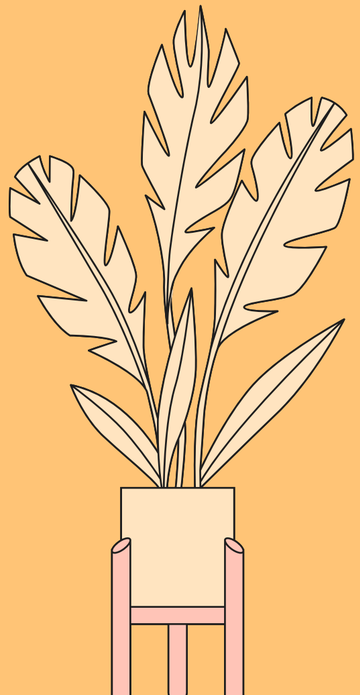


Presença de Outliers

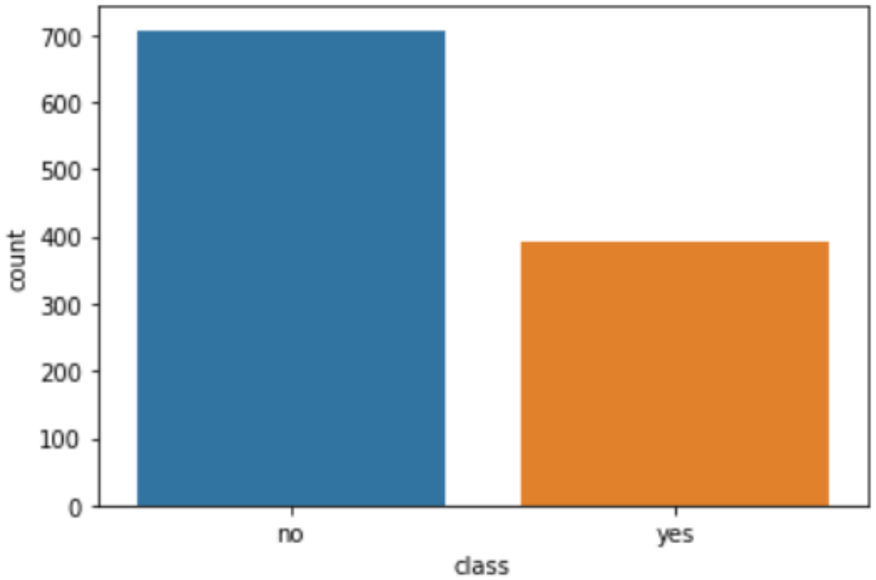


Entendendo Base de Dados

Por últimos, analisamos como está a disposição dos dados na coluna class

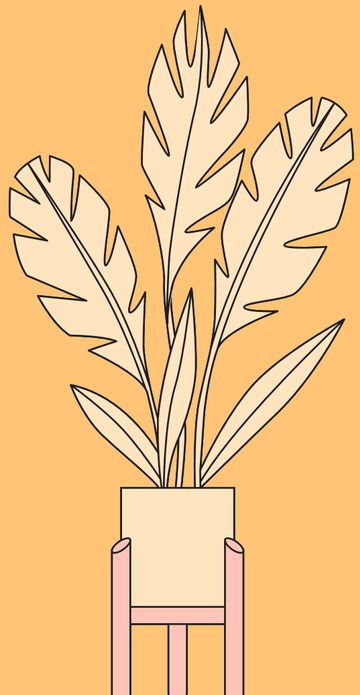


Verificando o balanceamento da classe



A classe é desbalanceada, temos mais instância com diagnóstico negativo para o TEA do que positivo.

Limpeza de dados

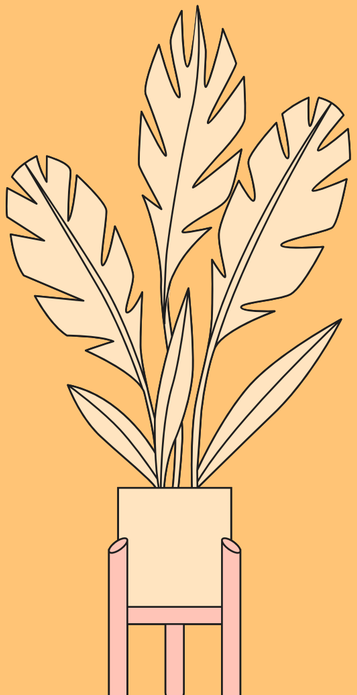


1. Valores ausentes

2. Mudança dos tipos das variáveis

3. Outliers

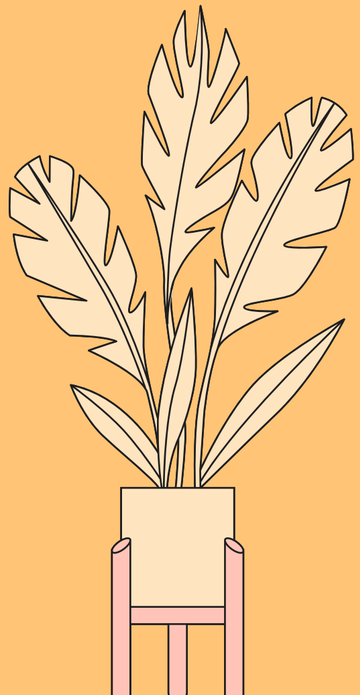
Limpeza de dados



Valores ausentes

Temos um total de 145 dados ausentes na nossa base de dados, como os atributos de ethnicity e who_answer são dados categóricos podemos substituir esses valores pela moda da coluna. No caso da idade como juntamos 3 bases de dados que eram separadas por faixa etária, iremos substituir pelo valor mais próximo.

Limpeza de dados



Mudança dos tipos das variáveis

- Nossa base de dados tinha muitos dados categóricos que não podiam ser convertidos para flot ou int;
- Realizamos um mapeamento atribuindo aos valores categóricos de cada variável

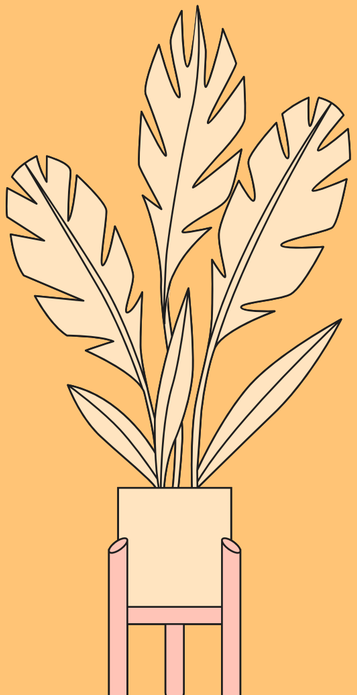
```
Coluna gender: ['m' 'f']
Coluna ethnicity: ['others' 'middle eastern
 'pasifika' 'hispanic' 'turkish' 'latino']
Coluna jundice: ['no' 'yes']
Coluna austim: ['no' 'yes']
```

Antes

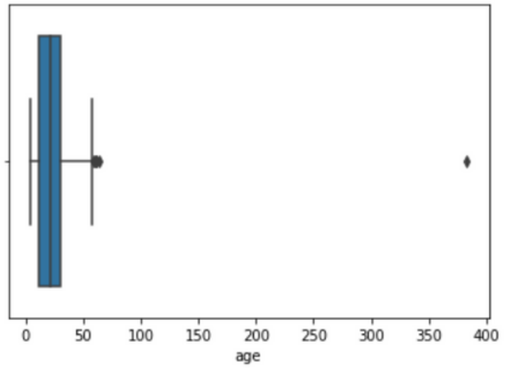
```
Coluna gender: [0 1]
Coluna ethnicity: [0 1 2 3 4 5 6 7 8 9]
Coluna jundice: [0 1]
Coluna austim: [0 1]
```

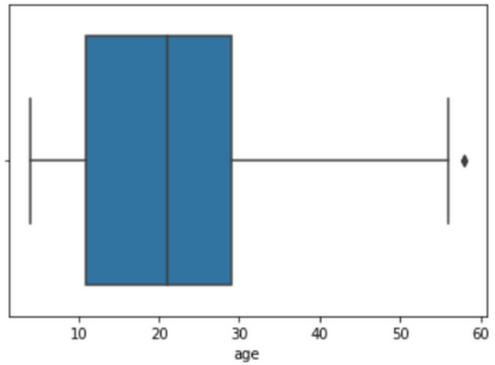
Depois

Limpeza de dados



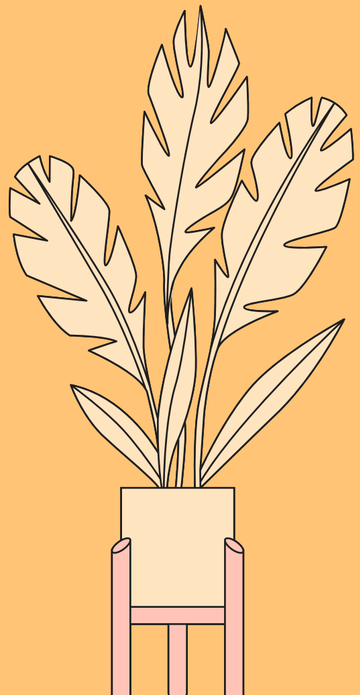
Outliers



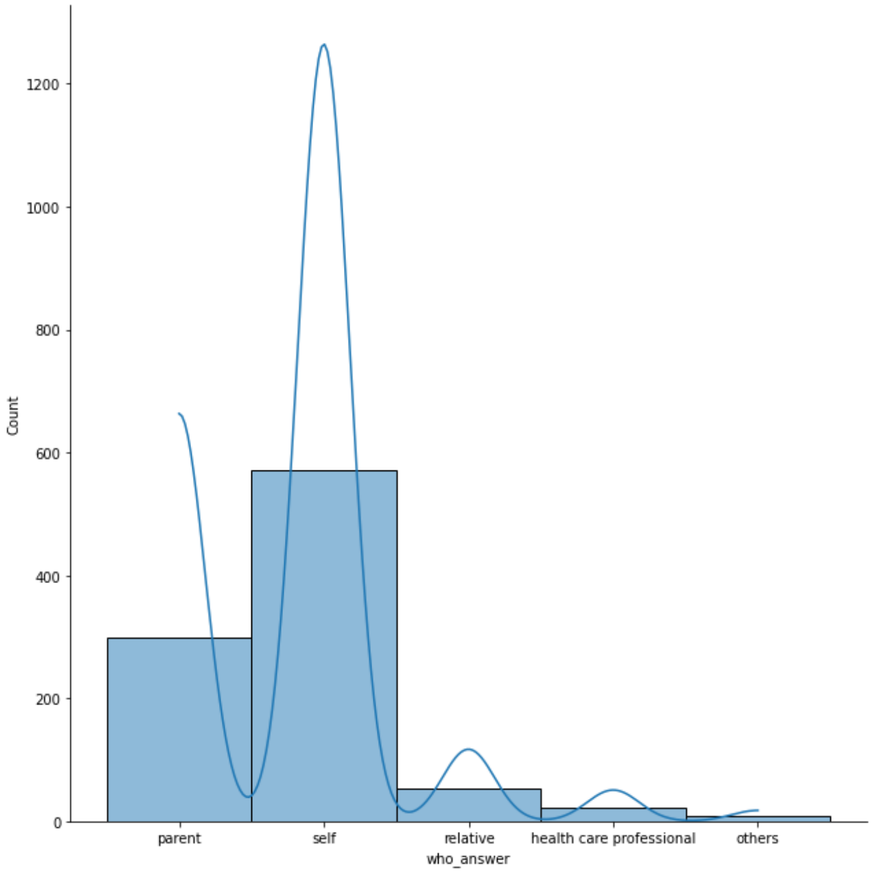


Encontramos Outliers apenas na coluna de age e para fazer o tratamento utilizamos o intervalo interquartil e removemos as instâncias que tinham esses dados discrepantes

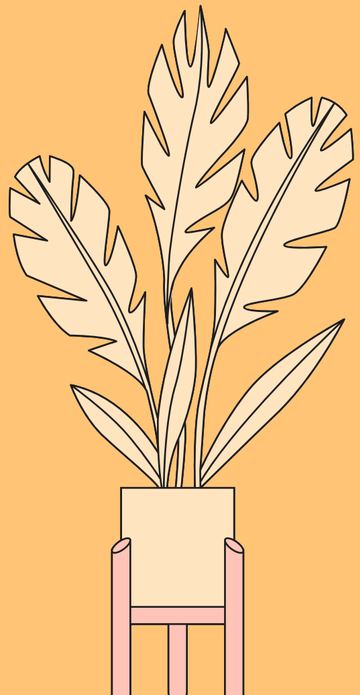
Representação Gráfica



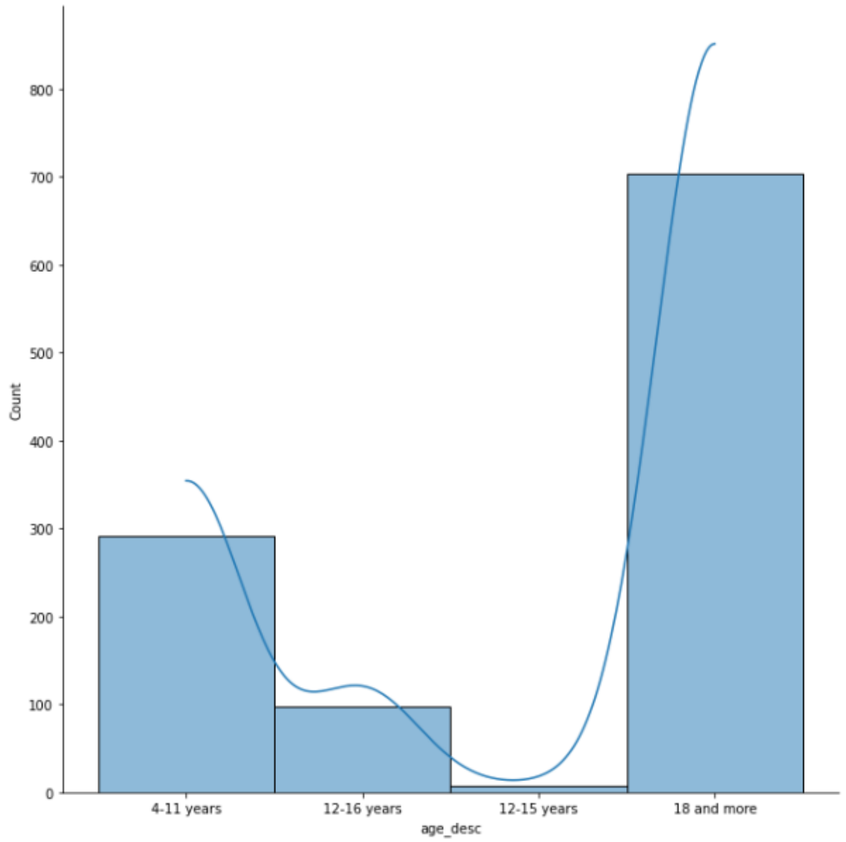
Quem respondeu o questionário



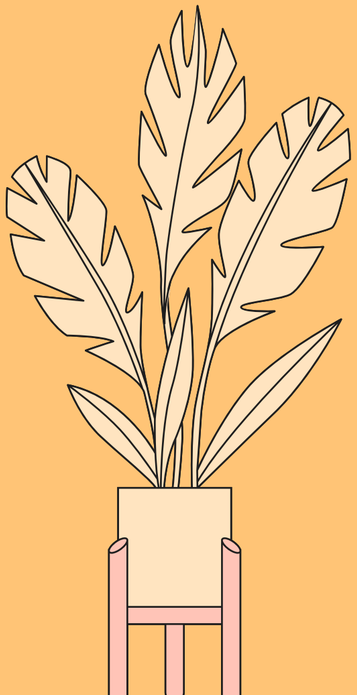
Representação Gráfica



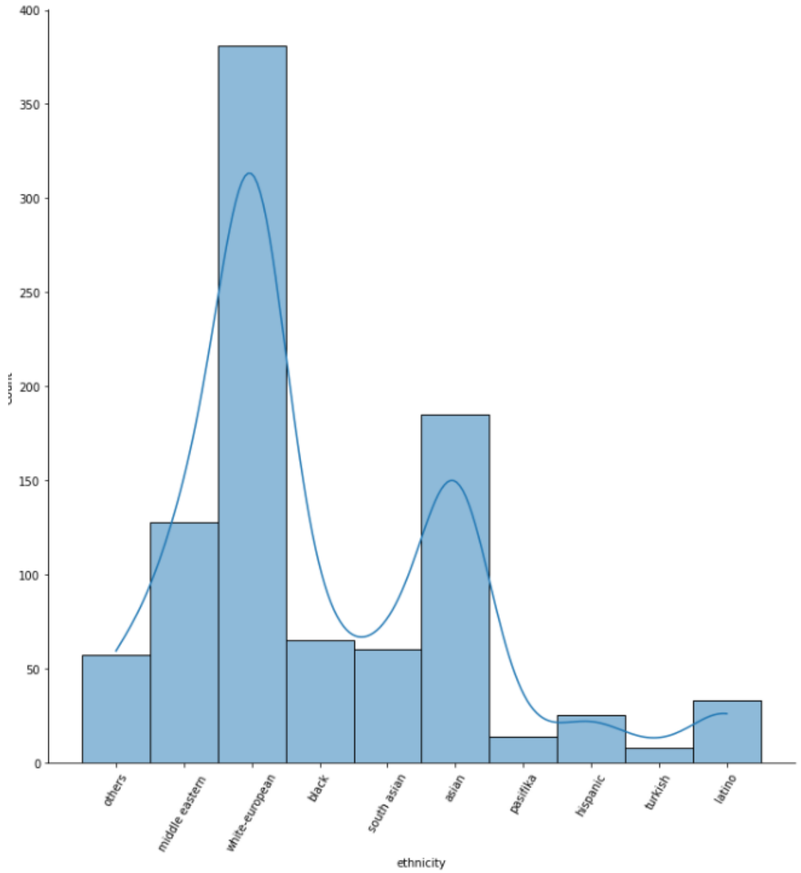
Faixa etária dos pacientes



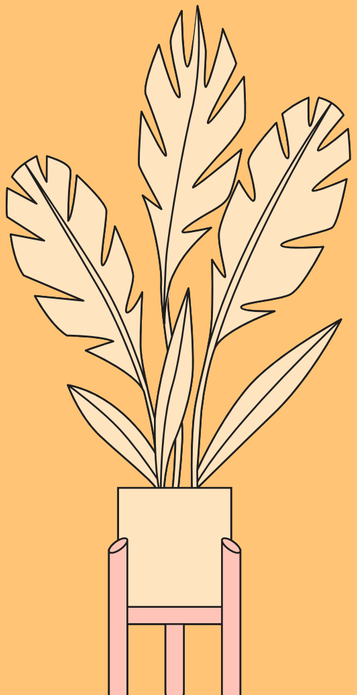
Representação Gráfica



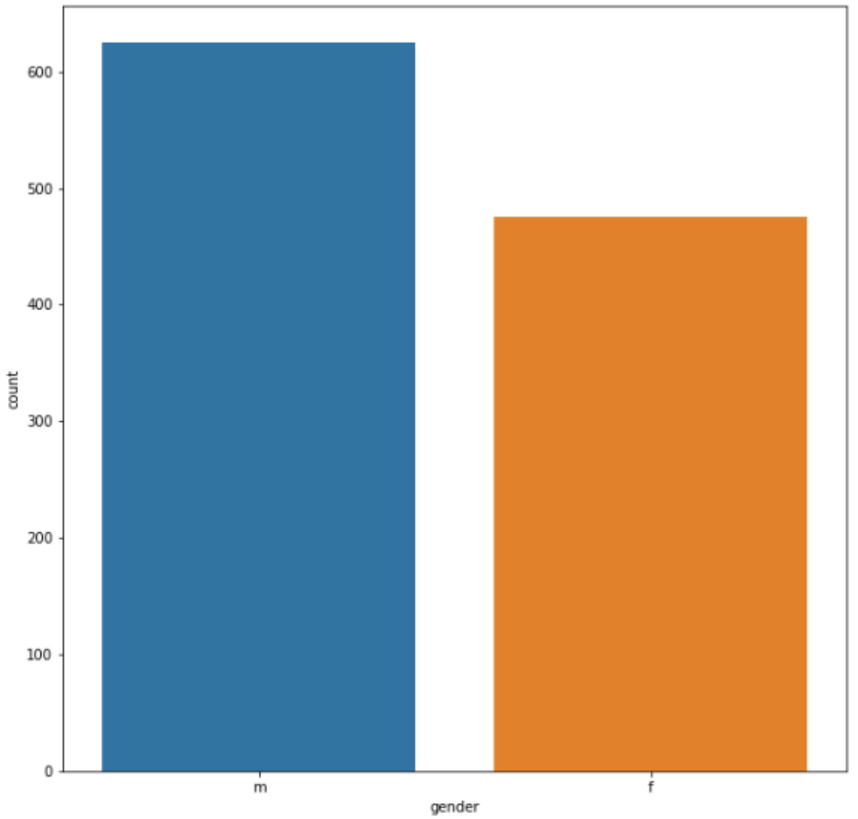
Etnia dos pacientes



Representação Gráfica



Gênero dos pacientes



Análise Estatística

Perfil dos Pacientes Positivos

01

Idade

A média de idade desses pacientes é de 19,78 anos.
Sendo a classe de 4-11 anos a de maior incidência.
Já o menor índice é em torno dos 50 anos.

02

Sexo

Há uma preponderância masculina com 54,25% de ocorrência contra os 45,75% feminina.

03

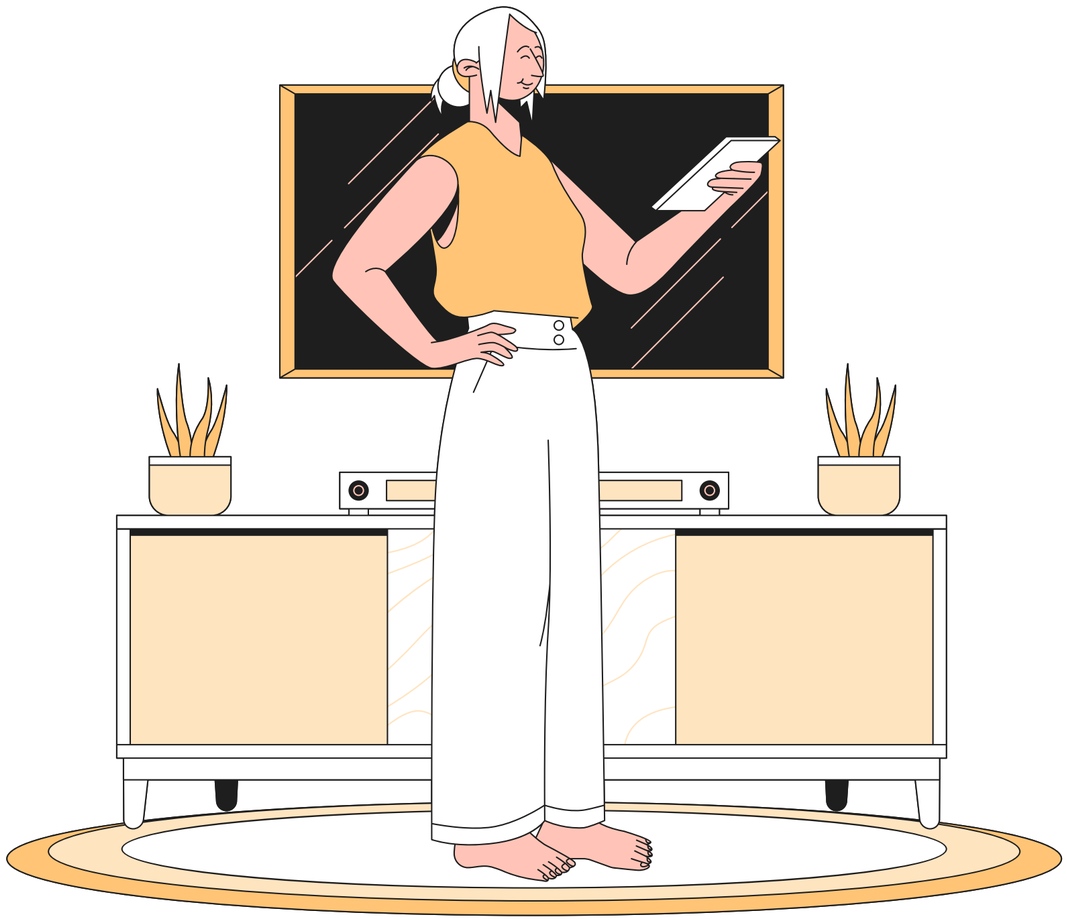
Nascidos com icterícia

18,36% dos pacientes diagnosticados dentro do espectro do TEA nasceram com icterícia.

04

Familiar incidente no TEA

18,63% desses indivíduos também tem algum familiar pertencente ao espectro.



Implementação do Classificador Ingênuo de Bayes

Equações

Equação 1

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

Probabilidade Condicional

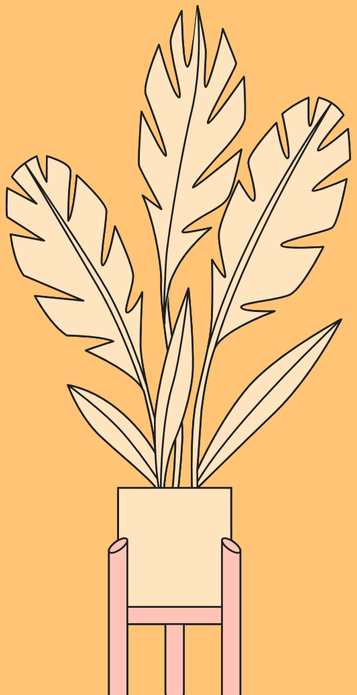
Equação 2

$$p(a_i|c_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(a_i-\mu_j)^2}{2\sigma_j^2}}$$

Probabilidade condicional em uma distribuição Gaussiana

Modelo e Experimentos

Programação do Modelo



1. Remoção das variáveis categóricas

2. Treinamento e Teste pelo Classificador Ingênuo de Bayes

3. Obtendo os resultados

Resultados

Experimento 1

Base de dados sem selecionar os atributos mais relevantes.

Acurácia = 96,80%

Precisão = 97%

Sensibilidade = 98%

Experimento 2

Base de dados selecionando os atributos mais relevantes.

Acurácia = 98,63%

Precisão = 99%

Sensibilidade = 99%

Experimento 3

Base de dados adicionando dados sintéticos às classe minoritárias.

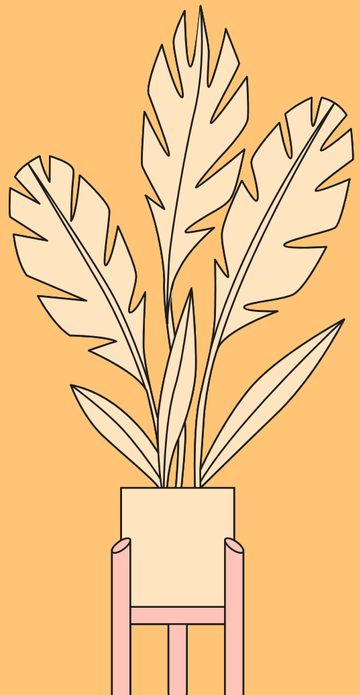
Acurácia = 94,86%

Precisão = 93%

Sensibilidade = 96%

Conclusões e Discussões

O melhor modelo a ser escolhido



Sensibilidade

Dado o contexto, apesar da importância da acurácia e precisão, deve-se priorizar a menor incidência possível de falsos negativos para que um paciente não deixe de receber o tratamento adequado.

Experimento 2

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 (recall superior a 98%).

Sistema de Predição confiável

Pelos experimentos, é possível notar que foi alcançado o objetivo de criar um sistema de predição que auxilie na identificação de casos de pacientes dentro do espectro do TEA de forma simples, acessível e eficaz.