Predição de Autismo

Antonio Jonas Gonçalves de Oliveira (2016021023)
Thales Monteiro Soares (2016019642)
Thiago Alves Araujo (2016019787)

Introdução

Atualmente, a aplicação de Machine Learning entre assuntos interdisciplinares se mostrou bastante comum e bem-sucedido, principalmente em campos como biologia e neurologia. É possível gerar padrões e tendências para grande base de dados com conteúdo médico, o que se torna foco de pesquisas.

A base selecionada foi a combinação de três bases de dados sobre pessoas que sofrem com o Transtorno do Espectro Autista (TEA). As bases possuem as mesmas classes e atributos, diferenciando apenas que são sobre crianças, adolescentes e adultos. Foram usados métodos de Machine Learning como SVM, KNN, Naive Bayes, para mostrar uma predição baseando-se no estudo feito dos atributos desta base, e ver que alguns apresentam resultados melhores que outros.

Motivação

O TEA é um transtorno que não possui cura, e afeta a vida de 1 a cada 100 crianças recém nascidas, apenas no Brasil, dados apontam que são 150 mil casos novos anuais. Ainda é uma condição que apresenta um grande tabu na sociedade, pois atinge principalmente a interação social, o comportamento e a linguagem, o que faz os portadores enfrentarem dificuldades diárias. Devido ao fato de ser um transtorno ainda pouco conhecida quando olhamos pelo ponto de vista de diagnósticos e tratamentos, usar métodos de IA para criar formas de auxiliar a área da saúde na identificação e tratamento correto, assegurando uma melhora na qualidade de vida das pessoas acometidas dessa condição.

Metodologia

Foi utilizado a ferramenta Jupyter para a criação do código, permitindo a execução e documentação do que foi feito, na linguagem Python 3. Foram usadas bibliotecas como:

- Numpy
- Pandas
- Scipy
- Sklearn
- Seaborn

Dataset

O conjunto de dados é formado pela junção de três datasets:

- Autistic Spectrum Disorder Screening Data for Children Data Set
- Autistic Spectrum Disorder Screening Data for Adolescent Data Set
- Autism Screening Adult Data Set

Todos os três datasets possuem os mesmos atributos, variando apenas a origem das instâncias e a quantidade das mesmas.

Os dados foram fornecidos ao UCI Machine Learning Repository pelo Departamento de Tecnologia Digital, Manukau Institute of Technology, Auckland, Nova Zelândia.

Dataset

Os três datasets possuem os seguintes atributos:

- A1_Score a A10_Score : Variáveis binárias que representam as perguntas do questionário utilizado.
- Age : Idade em anos.
- Ethnicity : Etnia do participante.
- Jaundice: Se o paciente nasceu com Icterícia.
- Autism: Se há casos de Autismo na família.
- used_app_before : Se o usuário já utilizou o aplicativo de triagem anteriormente.
- age_desc : Classificação de idade .
- relation: Quem ajudou no teste.
- Class/ASD: Classificação indicativa de Autismo ou não.
- Country_of_res : País de origem

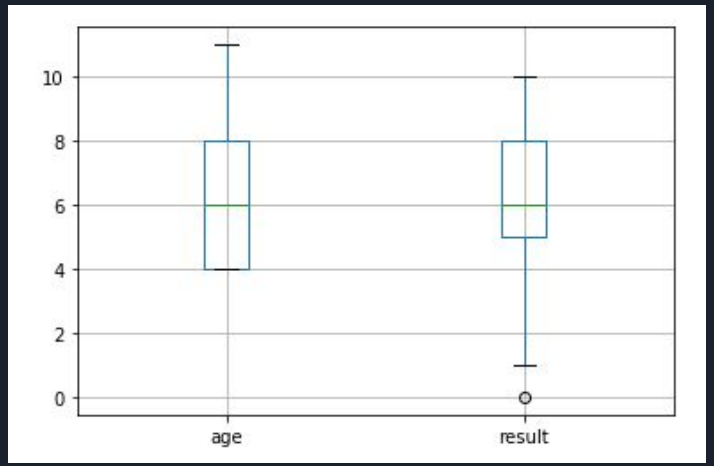
Pré-processamento Child

O pré processamento foi executado de forma independente nas três bases.

Na base Child Autism foi identificado que 43 das 288 instâncias possuíam valores nulos, estes campos nulos foram substituídos pela mediana da coluna correspondente.

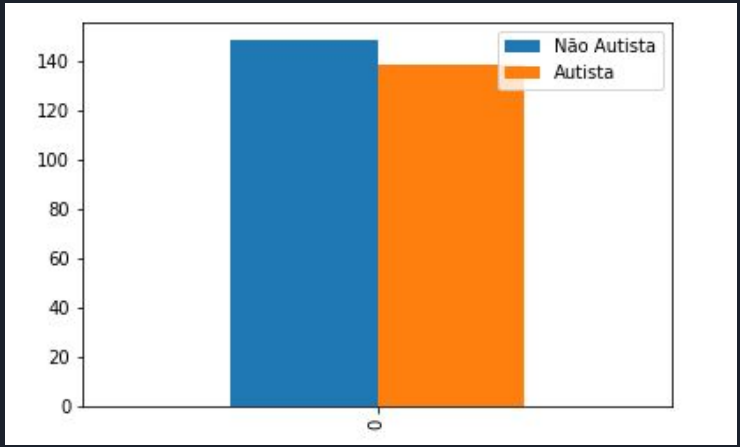
Foi encontrado duas instâncias repetidas, sendo necessário removê-las.

Não foi encontrado nenhum outlier.

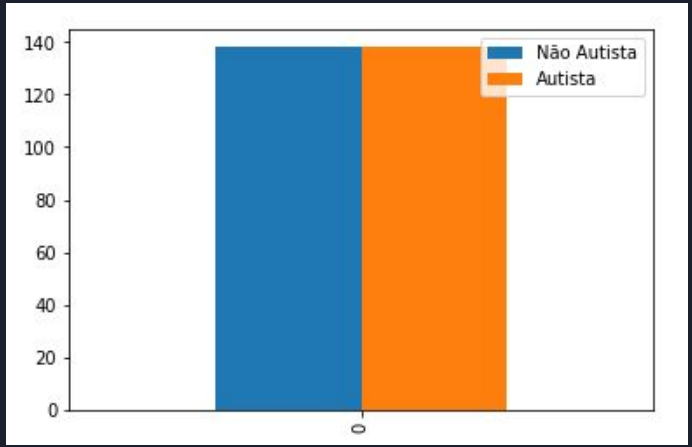


Pré-processamento Child

Observamos que a base estava balanceada, haviam 148 instâncias classificadas como Não Autistas e 138 Autistas. Para o balanceamento foi realizada um downsample da classe majoritária.



Classe desbalanceada



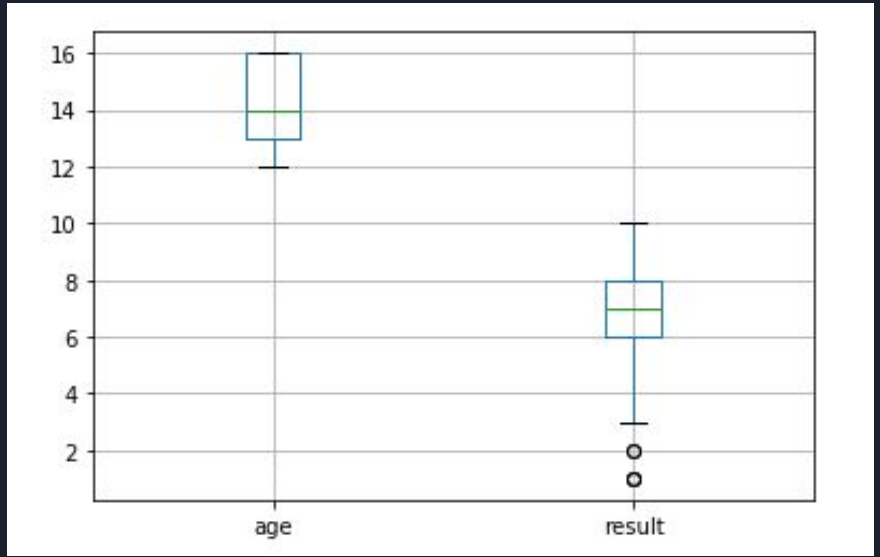
Classe balanceada

Pré-processamento Adolescent

Na base Adolescent Autism foi identificado que 6 das 104 instâncias possuíam valores nulos, estes campos nulos foram substituídos pela mediana da coluna correspondente.

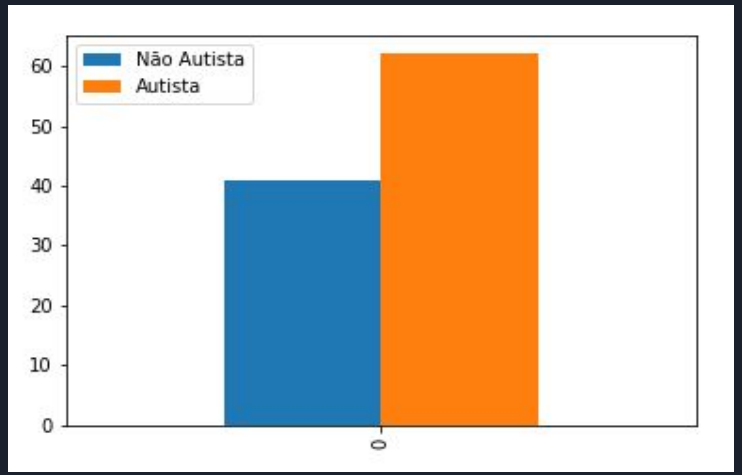
Foi encontrado uma instâncias repetidas, sendo necessário removê -las.

Não foi encontrado nenhum outlier. Pois Result igual a zero é um resultado possivel no teste.

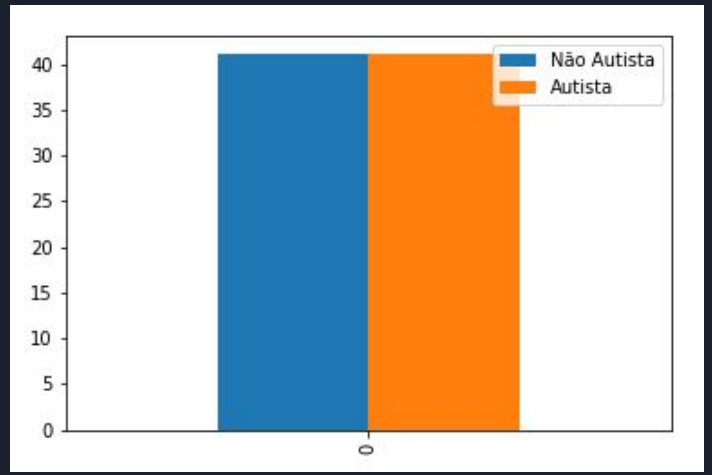


Pré-processamento Adolescent

A base com dados de adolescentes estava desbalanceada, haviam 41 instâncias classificadas como Não Autistas e 62 Autistas. Para o balanceamento foi realizada um downsample da classe Autista.



Classe desbalanceada



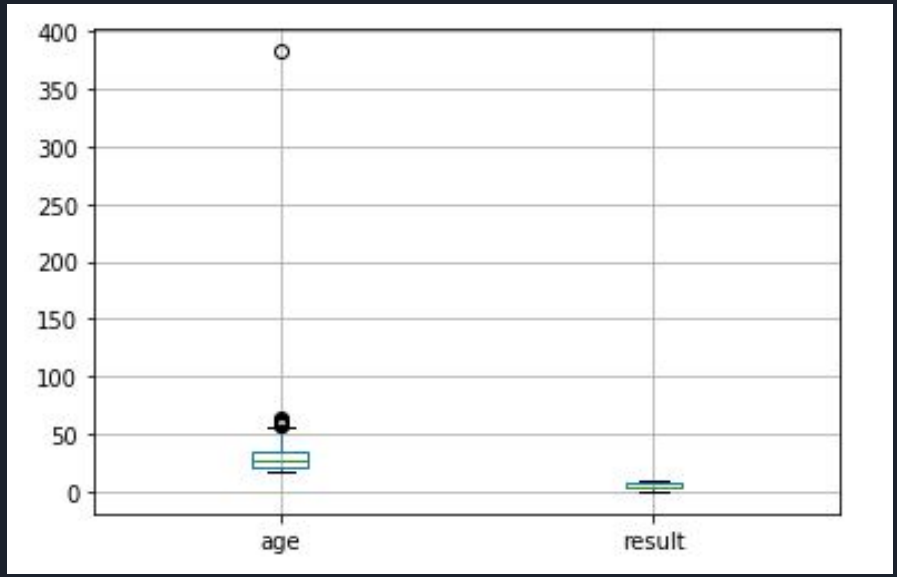
Classe balanceada

Pré-processamento Adult

Na base Adult Autism foi encontrado 95 das 702 instâncias com valores nulos, estes campos nulos foram substituídos pela mediana da coluna correspondente.

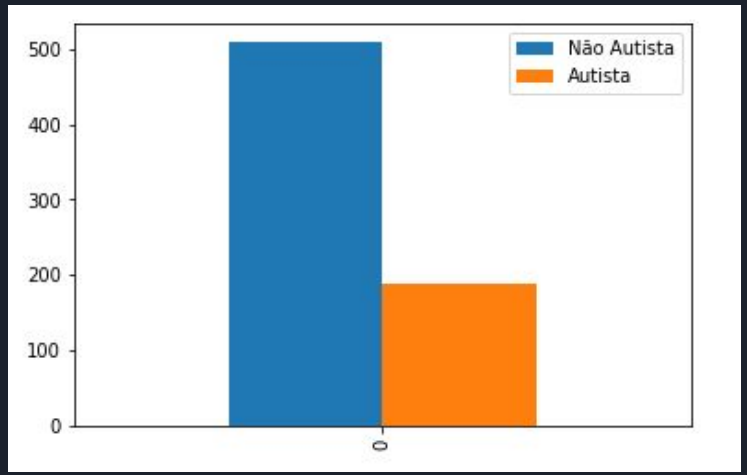
Foi encontrado 5 instâncias repetidas, sendo necessário removê -las.

Foi encontrado um outlier, Idade de 383 anos. A instância foi removida.

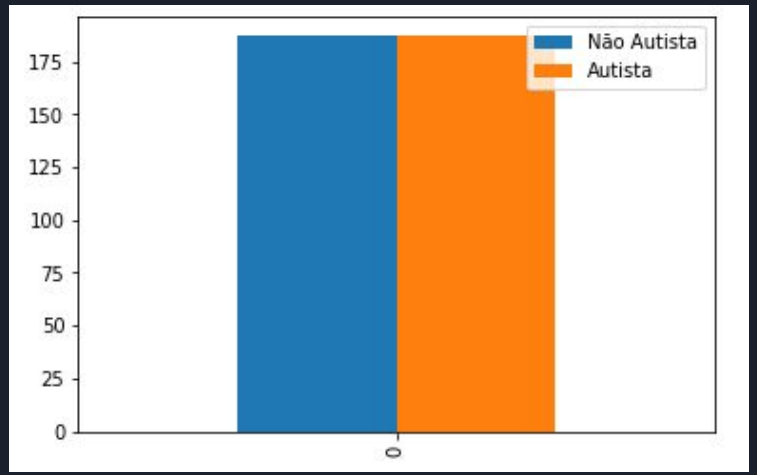


Pré-processamento Adult

A base com dados de adulto tinha 509 instâncias classificadas como Não Autistas e 187 Autistas. Para o balanceamento foi realizada um downsample da classe Não Autista.



Classe desbalanceada

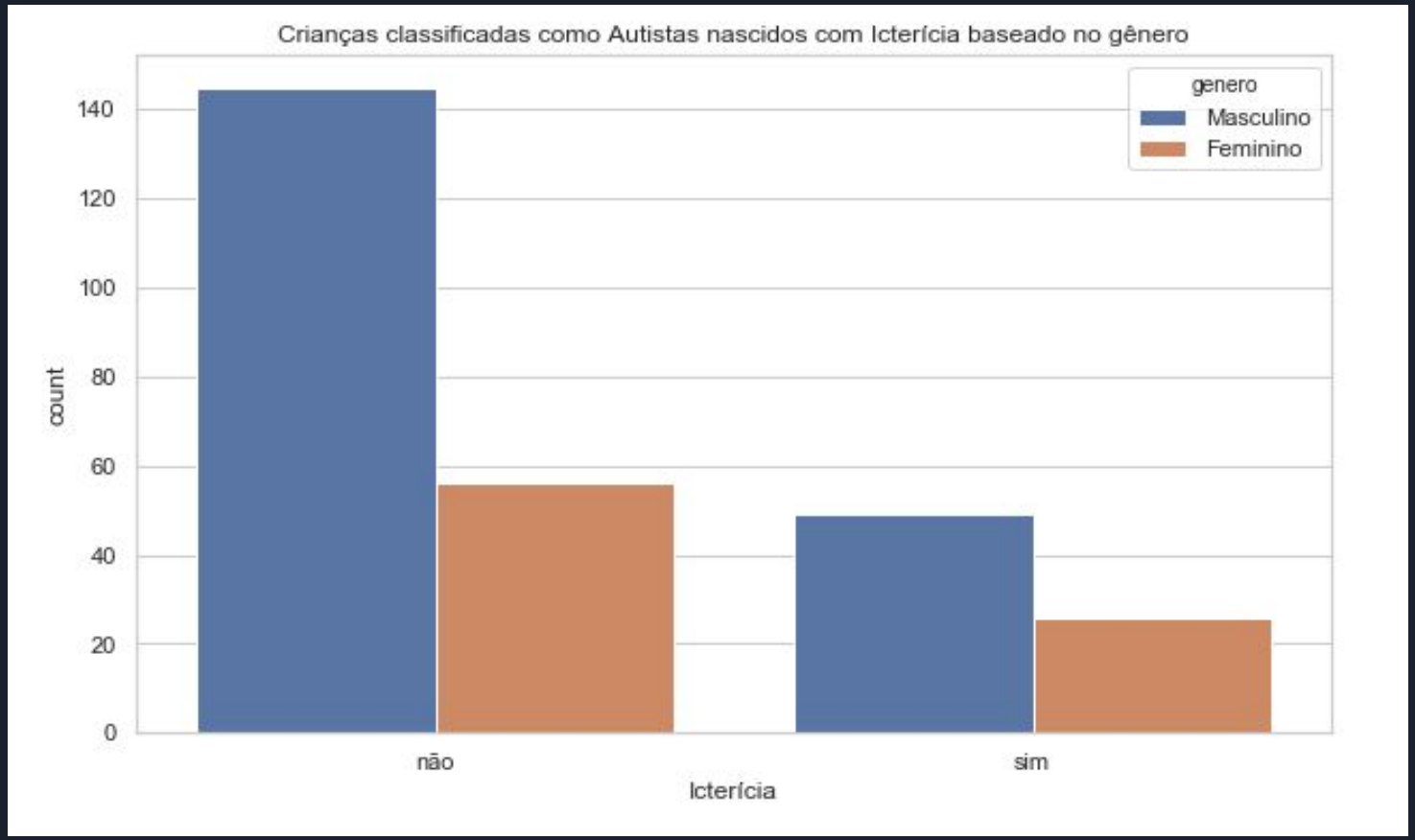


Classe balanceada

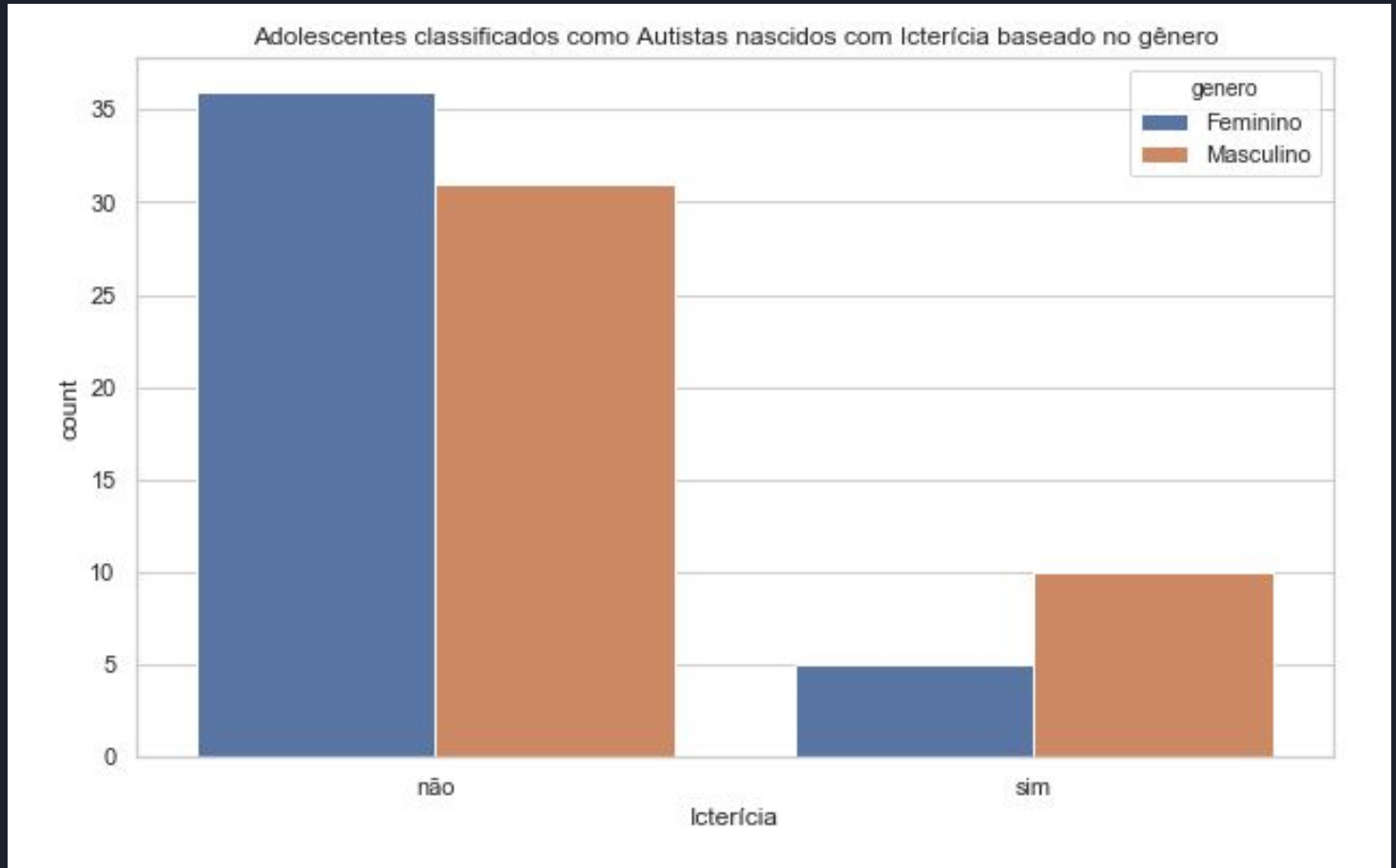
Pré-processamento

Após as três bases terem passados pela primeira fase de pré-processamento realizamos a junção dos datasets. Agora com o dataset completo foi necessário converter os valores binários para numéricos e realizar label encoder em todo o dataset.

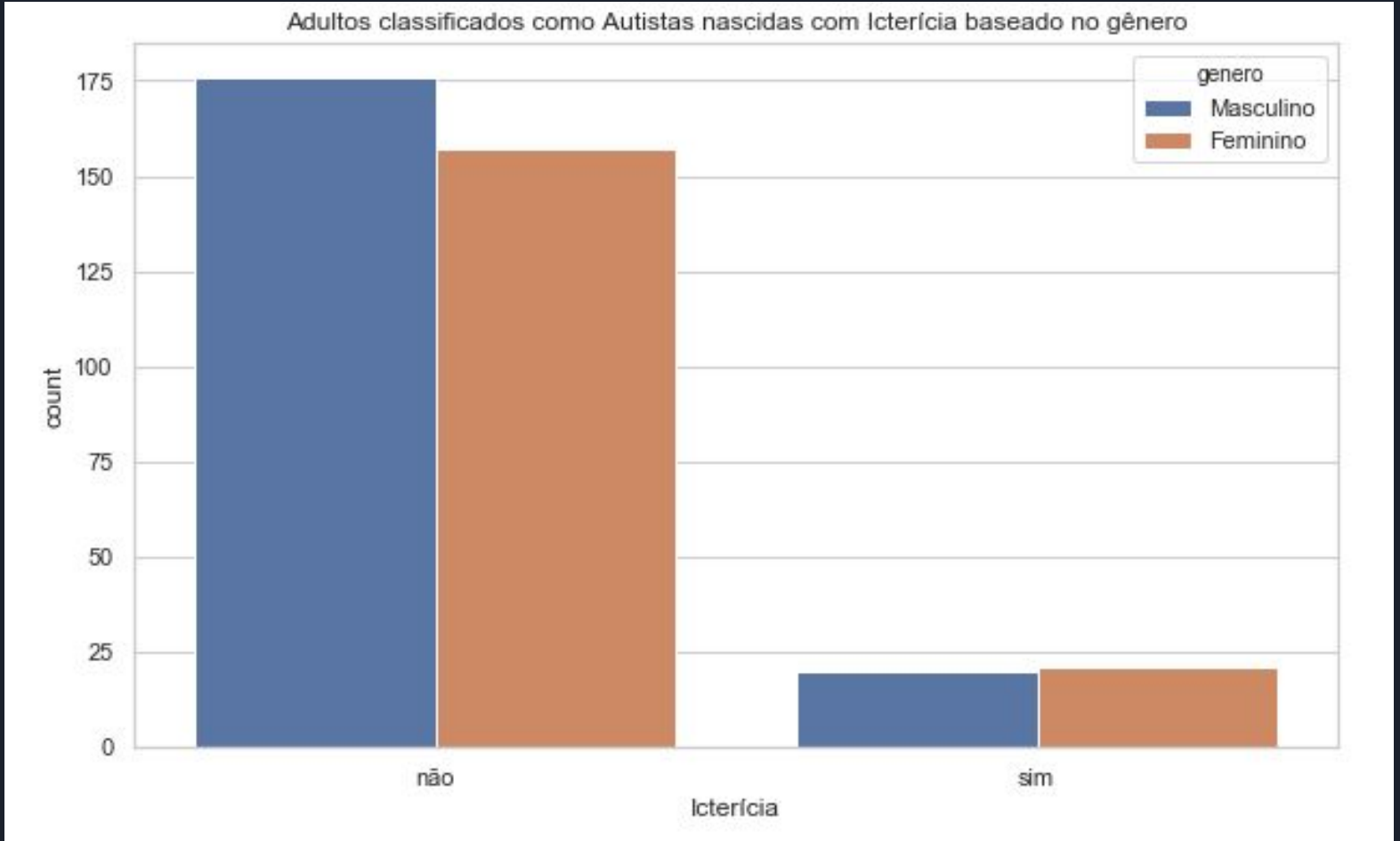
Quantidade de crianças autistas que nasceram com Icterícia.



Quantidade de adolescentes autistas que nasceram com Icterícia.



Quantidade de adultos autistas que nasceram com Icterícia.



Métodos

- **KNN**: Um dos mais famosos algoritmos de Machine Learning, é bastante versátil e possui uma facilidade no seu entendimento, pois possui uma ideia simples de aprendizado, baseando-se em quão similar um vetor é do outro.
- **SVM**: Outro algoritmo muito utilizado na área, principalmente em problemas de classificação, pois apresenta ótimos resultados. Ele utiliza a idéia de classificar determinado conjunto de pontos que são mapeados para um espaço multidimensional usando uma função Kernel.
- **Naive Bayes**: Classificador probabilístico que se baseia no teorema de bayes.

Resultados

Com os métodos definidos, foi feito o treino e teste dos modelos, e os resultados foram os seguintes:

	KNN K = 5 Metric='minkowski'	SVM Kernel = rbf	Naive Bayes Event Models = Gaussian
Acurácia	93,19%	97,95%	96,59%
F1 Score	94,11%	98,15%	96,89%
Acurácia - Validação Cruzada	87,16%	92,60%	85,77%

Conclusão

Nosso objetivo foi aplicar algoritmos de aprendizado de máquina supervisionados a um conjunto de dados derivado da pesquisa em Transtorno do Espectro do Autismo, com o intuito de classificar novas instâncias.

Usando uma validação cruzada na seleção dos dados, bem como uma partição dos dados em treinamento e teste, conseguimos construir três modelos com nível moderado de precisão (de acordo com as métricas apresentadas anteriormente) para previsão da variável target quando novos dados são fornecidos.

Referências

1. https://scikit-learn.org/stable/
2. https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult#
3. https://saude.abril.com.br/mente-saudavel/o-que-e-autismo-das-causas-aos-sinais-e-o-tratamento/