Modelagem estatística do autismo

Adrian Filipe de Castro Alves

Junho de 2024

Resumo

Este estudo emprega métodos estatísticos avançados para investigar os fatores que influenciam a presença de traços autísticos em indivíduos. Utilizando dados demográficos e respostas ao Autism Spectrum Quotient, exploramos correlações e padrões significativos que possam auxiliar em diagnósticos mais precisos e intervenções eficazes. Através da análise exploratória, regressão logística, Random Forest, entre outros métodos, buscamos identificar as características mais determinantes para o espectro autista

Sumário

1 Introdução

1.1 Contexto

O Transtorno do Espectro Autista (TEA) é uma condição de desenvolvimento que se caracteriza por desafios na interação social, comunicação e, frequentemente, comportamentos repetitivos. As manifestações do autismo variam significativamente, justificando a designação de "espectro", pois afetam indivíduos de maneiras diferentes e em graus variados. Considerando a complexidade e a variabilidade do TEA, o diagnóstico e a intervenção precoces são fundamentais, podendo melhorar significativamente a qualidade de vida das pessoas afetadas pelo autismo.

O conjunto de dados que utilizaremos é essencial para alcançar esses objetivos. Ele inclui uma gama de variáveis coletadas de pacientes através da ferramenta de triagem Autism Spectrum Quotient (AQ) de 10 itens, que contém pontuações (de A1 Score a A10 Score) refletindo respostas a questões destinadas a avaliar a presença de traços autísticos. Além disso, o conjunto de dados engloba informações demográficas como idade, gênero e etnia, e outros fatores relevantes, como ocorrência de icterícia ao nascer, autismo em familiares imediatos, histórico de triagens anteriores e o país de residência do paciente. A principal variável de saída, 'Class/ASD', indica se o paciente é classificado como estando no espectro autista.

Este conjunto de dados fornece uma base sólida para análises detalhadas, úteis para médicos, cuidadores e pesquisadores. A correlação entre variáveis demográficas e as pontuações AQ pode revelar padrões e possíveis correlações entre características específicas ou fatores ambientais e os diagnósticos de TEA. Essas análises são essenciais para a pesquisa médica e psicológica, o aprimoramento de ferramentas de triagem e a melhoria das intervenções.

Nosso estudo visa entender quais variáveis têm o maior impacto no diagnóstico do TEA e quais modelos são melhores para este estudo. Exploraremos influências de fatores genéticos, prevalência do TEA em diferentes contextos geográficos e a eficácia de métodos de triagem previamente utilizados.

Ao explorar este conjunto de dados, pretendemos aprofundar nossa compreensão do TEA e contribuir para o desenvolvimento de terapias e suportes mais direcionados, promovendo assim uma melhor qualidade de vida para aqueles no espectro.

1.2 Tratamento dos Dados

Inicialmente, realizaremos o tratamento e limpeza dos dados para otimizar nossas análises.

Começaremos pela variável 'Class/ASD', que foi renomeada para 'resultado autismo'. Esta variável é classificada como '0' para ausência de autismo e '1' para presença de autismo, proporcionando uma clara distinção no resultado do diagnóstico.

As variáveis de 'A1 score' até 'A10 score', agora chamadas 'pontuação1', 'pontuação2', ..., representam pontuações baseadas na ferramenta de triagem de 10 itens do Quociente do Espectro do Autismo (AQ). Cada uma dessas variáveis é avaliada como '0' ou '1', indicando a presença ou ausência de características específicas associadas ao autismo.

A variável 'gender', agora renomeada para 'gênero', identifica o gênero do paciente. Originalmente dividida entre 'f' (feminino) e 'm' (masculino), essa classificação será alterada para '0' se masculino e '1' se feminino, simplificando a análise de dados por gênero.

A variável 'age', renomeada para 'idade', continua a representar a idade do paciente em números contínuos, fundamental para análises que consideram a faixa etária.

Para a variável 'ethnicity', agora chamada 'etnia', que é uma variável categórica ordinal, continua a fornecer informações sobre a etnia do paciente, o que pode ser relevante em análises epidemiológicas ou demográficas.

Modificamos também as variáveis 'jaundice' e 'autism', agora chamadas 'icterícia' e 'família', respectivamente. Estas indicam se o paciente teve icterícia ao nascer e se um membro imediato da família foi diagnosticado com autismo. Ambas serão convertidas de 'no' para '0' e 'yes' para '1', padronizando as respostas para facilitar a análise.

Adicionalmente, as variáveis categóricas binárias ‘countryofres‘, ‘usedappbefore‘ e ‘relation‘, renomeadas para ’País’, ’triagem antes’ e ’relação’, informam, respectivamente, o país do paciente, se ele foi submetido a um teste de triagem anteriormente e a relação do indivíduo que completou o teste. Essas informações são essenciais para compreender o contexto de cada caso.

Por fim, removeremos a coluna ‘agedesc‘, que descreve a idade mas é redundante, pois todos os registros contêm a mesma descrição e também removemos a coluna ID que descreve qual participante é qual. Essa eliminação torna o conjunto de dados mais enxuto e focado nas variáveis que impactam diretamente as análises.

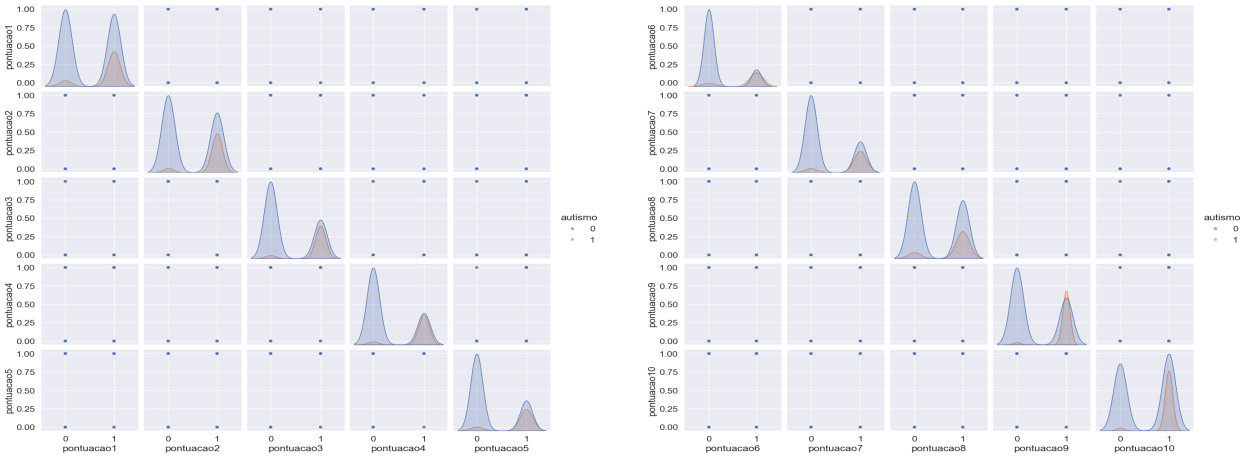
Tabela 1: Variáveis

Variáveis	**Tipo**	**Descrição**
pontuacao1 a 10	Categórica Binária	Pontuação baseada na ferramenta de triagem de 10 itens do Quociente do Espectro do Autismo (AQ).
idade	Contínua	Idade do participante.
genero	Categórica Binária	Gênero do participante (ex: masculino, feminino).
Etnia	Categórica Ordinária	Classificação étnica do participante.
Icteria	Categórica Binária	Presença de icterícia no nascimento.
Familia	Categórica Binária	Histórico familiar de condições semelhantes.
País	Categórica Ordinária	País de origem do participante.
triagem antes	Categórica Binária	Se houve triagem prévia para a condição estudada.
pontuação teste	Contínua	Pontuação obtida no teste realizado.
relação	Categórica Ordinária	Tipo de relação com outros participantes no estudo.
resultado autismo	Categórica Binária	Resultado do teste para autismo.

1.3 Análise Exploratória dos Dados

Para fazer o nosso modelo, primeiramente vamos fazer uma análise exploratória de dados, que é quando vamos analisar cada coluna, suas distribuições, quantidade e correlações entre si e a variável resposta, pois assim podemos avaliar todas as variáveis para ajustar no nosso modelo

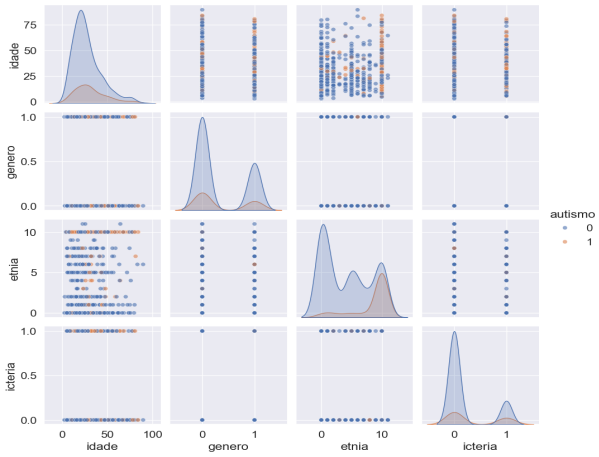
Primeiramente vamos usar os pair plots ou gráficos de pares, e são usados para explorar visualmente as relações entre múltiplas variáveis em um conjunto de dados. Cada painel (ou célula) mostra a relação entre duas variáveis diferentes com histogramas ou estimativas de densidade de probabilidade (KDE) para cada categoria de uma variável categórica, que neste caso é rotulada como ”autismo”com categorias 0 e 1.



Os gráficos ao longo da diagonal principal (da esquerda superior para a direita inferior) são gráficos univariados. Eles mostram a distribuição de cada variável individual para cada categoria do autismo.

Já os gráficos fora da diagonal principal mostram as relações bivariadas entre duas variáveis distintas, como "pontuacao1"vs "pontuacao2". No entanto, os gráficos acima não possuem esses gráficos bivariados, apenas pontos únicos indicando a falta de repetição ou variação na combinação de duas variáveis específicas.

Se a cor azul (autismo = 0) e a cor laranja (autismo = 1) têm formas ou picos muito diferentes, isso indica que as características da variável podem ser diferentes entre os grupos. Se há uma grande sobreposição entre as curvas azuis e laranjas, pode ser difícil distinguir entre os grupos com base nessa variável específica.



Acima no primeiro gráfico as variáveis analisadas são idade, gênero, etnia, icterícia. Na diagonal principal temos idade que mostra as distribuições de idade para indivíduos sem e com autismo. Pode-se ver que ambas as distribuições são semelhantes, mas o grupo com autismo (laranja) tem uma leve concentração em idades mais jovens.

Em gênero as distribuições binárias para cada categoria de gênero, comparando a frequência de autismo. Parece haver uma pequena diferença entre os gêneros.

Na etnia a distribuição de etnia mostra pouca variação entre os grupos, sugerindo que a etnia pode não ser um fator distintivo significativo para o autismo neste conjunto de dados.

Na icterícia a distribuição parece indicar uma diferença mais clara, com o grupo com autismo mostrando maior frequência de icterícia.

Já fora da diagonal estes gráficos mostram a relação entre duas variáveis diferentes para cada categoria de autismo. Por exemplo, idade vs. gênero, idade vs. etnia, etc. Os pontos indicam como as características se distribuem cruzadamente nos dois grupos.

Já na segunda as variáveis analisadas são família, país, triagem prévia, pontuação do teste, relação.

Na diagonal principal temos família e a distribuição mostra se há histórico familiar de autismo. Os dados sugerem uma discrepância, com mais históricos familiares no grupo com autismo.

Em país, similar à etnia no primeiro gráfico, mostrando a distribuição de casos por país. Esta distribuição pode ajudar a identificar se certas localidades têm prevalência maior ou menor de autismo.

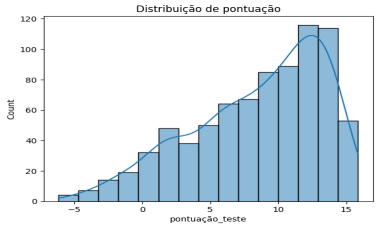
Na triagem prévia indica se houve uma triagem prévia para autismo, mostrando uma diferença significativa entre os grupos.

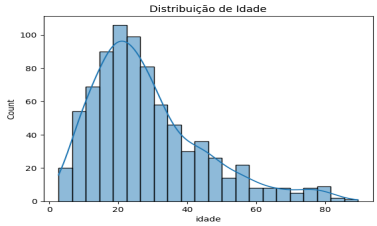
Na pontuação do teste a distribuição das pontuações de um teste específico, indicando diferenças notáveis entre os dois grupos.

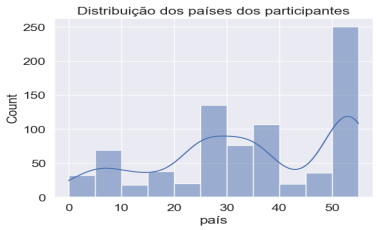
Na relação a distribuição que mostra o nível de relação (provavelmente familiar ou social), com diferenças evidentes entre os grupos.

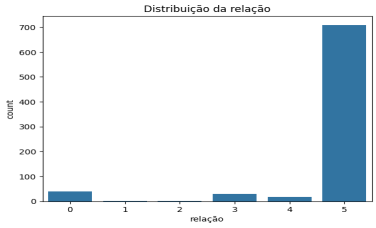
Fora da diagonal estes gráficos exploram como duas variáveis interagem entre os indivíduos com e sem autismo, ajudando a identificar padrões ou correlações significativas entre as variáveis.

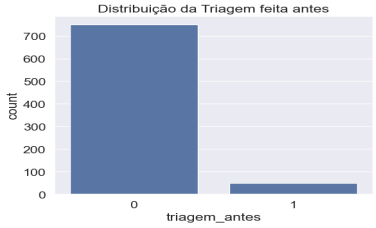
Agora vamos analisar a distribuição das variaveis:

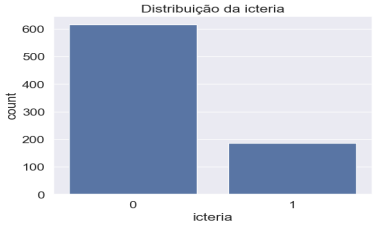


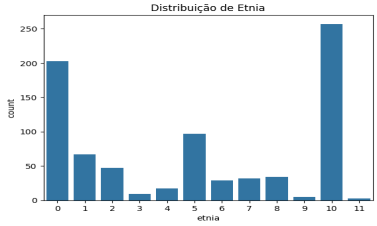


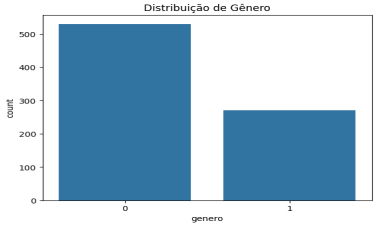


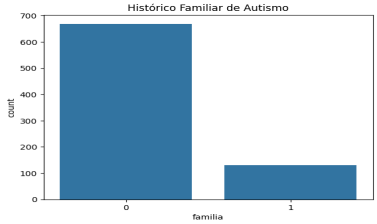


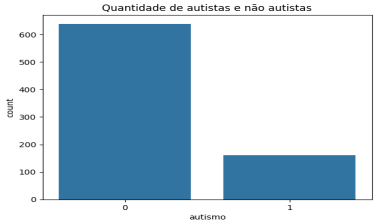










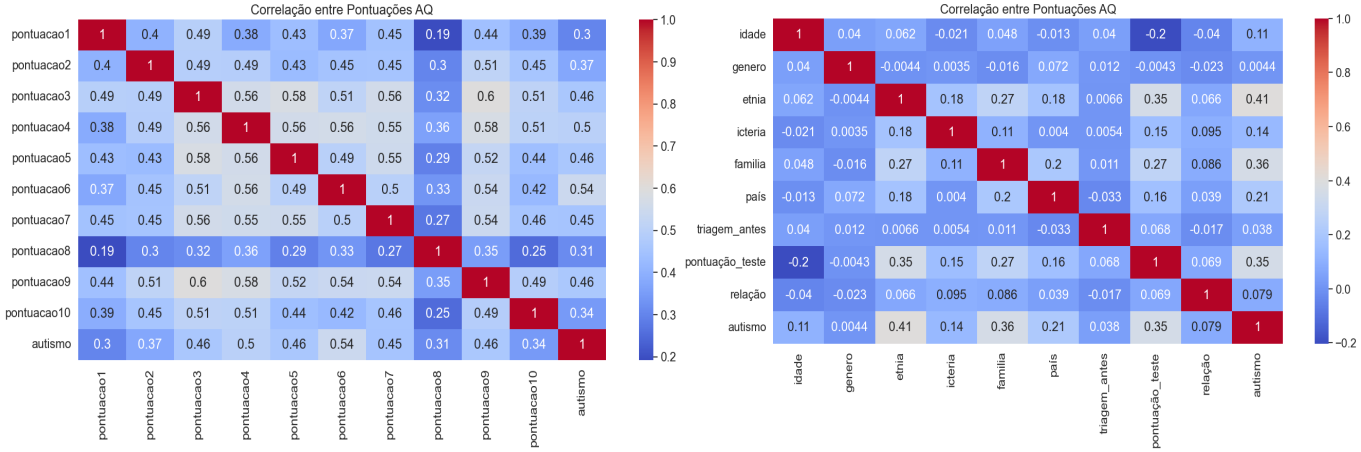


Observe que as distribuições das pontuações dos testes e das idades dos participantes são bem distribuídas. Adicionalmente, também examinamos a distribuição dos países de origem dos participantes e suas relações. É importante notar que temos representantes de 55 países, resultando em 55 categorias distintas. Esse número elevado de categorias pode impactar negativamente a eficácia do nosso modelo categórico. No que se refere à variável 'relação', observa-se uma predominância significativa de um tipo específico, o que pode limitar a capacidade do modelo de se adaptar a outras situações menos frequentes.

Também observamos diferenças nas distribuições das variáveis de triagem prévia e icterícia. Na icterícia, ainda possuímos uma quantidade suficiente de dados para treinar a categoria '1'. Em contrapartida, na triagem prévia, embora a presença do '1' seja menos frequente, ainda é possível trabalhar com essa categoria, considerando que a maioria das pessoas não realizou triagem prévia ou teve icterícia ao nascer.

Quanto à etnia, que possui 15 categorias, há algumas diferenças entre elas, o que é esperado devido à diversidade estatística inerente às categorias étnicas. Em relação ao gênero, embora haja uma maior representatividade masculina, as diferenças não são tão marcantes.

No que diz respeito ao histórico familiar e ao número real de autistas no conjunto de dados, não observamos grandes diferenças entre os grupos. No entanto, é evidente que a proporção de autistas é consideravelmente menor em ambos os casos, o que corresponde às expectativas, visto que estatisticamente uma em cada 36 pessoas é diagnosticada com autismo.



Nos gráficos de correlação apresentados, avaliamos a força da relação entre variáveis numéricas, com coeficientes variando de -1 a 1, onde valores próximos a 1 ou -1 indicam uma forte correlação positiva ou negativa, respectivamente, e valores próximos a 0 indicam pouca ou nenhuma correlação. O foco principal dessas análises é entender como diferentes variáveis estão relacionadas à variável resposta, o autismo.

Observamos que algumas variáveis, como gênero, triagem prévia e relação, têm correlações muito baixas com o autismo, sugerindo que elas possuem pouca influência direta na previsão da condição. Por outro lado, variáveis como etnia e pontuação do teste mostram correlações mais significativas, indicando que podem desempenhar um papel mais crítico na modelagem.

Essas análises são cruciais para otimizar nossos modelos. A partir desses dados, podemos considerar a exclusão de variáveis com baixa correlação, pois elas podem não contribuir significativamente para a precisão do modelo e podem até introduzir ruído. Além disso, o foco em variáveis com maiores correlações pode nos permitir aprimorar a precisão das previsões, simplificando o modelo sem comprometer seu desempenho.

Adicionalmente, as métricas derivadas dessas correlações também serão úteis para refinar ainda mais a seleção de variáveis e ajustar os parâmetros do modelo, garantindo uma análise mais eficiente e orientada por dados. Portanto, esses gráficos de correlação não apenas ilustram relações lineares entre as variáveis mas também guiam a tomada de decisão estratégica na modelagem preditiva do autismo.

2 Metodologia

2.1 Regressão Logística

Para um problema onde a variável resposta é binária um bom método e á regressão logística, destrincharemos melhor a regressão logística e abordaremos outros modelos que pode ajudar a comparar para avaliar a eficácia do modelo

A regressão logística é um método estatístico utilizado para modelar problemas de classificação binária, onde a variável de resposta é categórica e possui dois possíveis resultados: 0 ou 1. Diferente da regressão linear, que assume uma relação linear entre as variáveis independentes e a variável dependente, a regressão logística utiliza uma função logística para modelar a probabilidade de ocorrência de um evento. A função logística, também conhecida como função sigmoide, transforma a saída da combinação linear das variáveis preditoras em uma probabilidade entre 0 e 1.

Resumindo, regressão logística modela a probabilidade p de que a variável dependente Y seja igual a 1, dado um vetor de variáveis explicativas $X = (X_1, X_2, \ldots, X_k)$. Matematicamente, isso é escrito como:

$$P(Y = 1|X) = \frac{1}{1 + e^{-z}}$$

onde z é a combinação linear das variáveis explicativas, incluindo um termo de interceptação:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

A função logística transforma a combinação linear z de forma que o resultado esteja sempre entre 0 e 1, adequado para modelar uma probabilidade. A função inversa da função logística é a função logit, que

transforma probabilidades em valores que podem variar de $-\infty$ a ∞:

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = z$$

Este lado esquerdo é conhecido como o logit da probabilidade $P(Y=1|X)$, e é igual à combinação linear das variáveis explicativas.

Os parâmetros $\beta_0, \beta_1, \ldots, \beta_k$ são geralmente estimados usando o método de máxima verossimilhança, que busca encontrar os valores de β que maximizam a probabilidade dos dados observados

No nosso conjunto de dados, vamos usar a regressão logística para prever se um paciente está dentro do espectro do autismo (Autismo = 1) ou não (Autismo = 0). As variáveis preditoras incluem informações demográficas e resultados de testes (como idade, gênero, pontuações de triagem, histórico familiar de autismo, etc.). Após treinar o modelo, interpretaremos os coeficientes para entender a influência de cada variável preditora e calcularemos as odds ratios para facilitar a interpretação. A acurácia, precisão, recall e F1-score serão usadas para avaliar o desempenho do modelo.

2.2 Arvore de Decisão - Random Forest

Árvores de decisão são um tipo de modelo preditivo que usa uma estrutura de árvore para tomar decisões. Imagine que cada "nó"na árvore faz uma pergunta sobre os dados e, com base na resposta, você segue para o próximo nó até chegar a uma "folha"da árvore, que dá a previsão final. Essas perguntas são geralmente baseadas em características ou atributos dos dados, que no nosso caso pode ser a etnia, país, o histórico familiar de autismo etc.

A árvore é construída dividindo os dados de treinamento em subconjuntos de forma iterativa. Este processo é conhecido como "particionamento recursivo"e é feito de maneira a maximizar a homogeneidade dos subconjuntos resultantes. Ou seja, cada subconjunto, ao final do particionamento, deve conter dados tão similares quanto possível em relação ao alvo (como uma classe específica). A medida de homogeneidade (ou impureza) mais comum é a entropia ou o índice Gini para classificação e o erro quadrático médio para regressão.

Random Forest é um método que utiliza múltiplas árvores de decisão para melhorar a precisão da previsão e reduzir o risco de overfitting. Cada árvore no modelo é treinada em um subconjunto aleatório dos dados e das características, e a decisão final é obtida através da média (para regressão) ou da maioria dos votos (para classificação) das previsões das árvores individuais. Random Forest é robusto a outliers e variáveis irrelevantes, e pode capturar interações não lineares entre variáveis preditoras.

Aplicaremos o modelo de Random Forest para prever a classificação de autismo. Usaremos as mesmas variáveis preditoras mencionadas anteriormente. Este modelo será treinado e avaliado usando um conjunto de validação, e as métricas de desempenho como acurácia e F1-score serão comparadas com outros modelos. Além disso, analisaremos a importância das variáveis fornecida pelo Random Forest para identificar quais características são mais influentes na previsão.

2.3 Outros Modelos Para Comparar

2.3.1 Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) é uma técnica que cria um modelo forte a partir de uma sequência de modelos fracos (geralmente árvores de decisão). Cada novo modelo é treinado para corrigir os erros cometidos pelos modelos anteriores, utilizando o gradiente do erro da função de perda. XGBoost, LightGBM e CatBoost são implementações populares do GBM que oferecem melhorias em termos de velocidade e desempenho.

Usaremos GBM para prever a classificação de autismo no conjunto de dados. As implementações específicas, como XGBoost e LightGBM, serão aplicadas para maximizar a eficiência e a acurácia das previsões. Esses modelos serão treinados utilizando as variáveis preditoras e avaliados com um conjunto de validação. Ajustaremos hiper parâmetros para otimizar o desempenho e utilizaremos métricas como acurácia, precisão, recall e F1-score para a avaliação.

2.3.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) são modelos de aprendizado supervisionado que podem ser utilizados tanto para classificação quanto para regressão. Para problemas de classificação, SVM encontra o hiperplano que maximiza a margem entre as classes no espaço de características. O modelo pode usar diferentes tipos de kernels (linear, polinomial, RBF) para lidar com dados que não são linearmente separáveis.

Aplicaremos o SVM para classificar se um paciente está dentro do espectro do autismo. Escolheremos o kernel mais apropriado (por exemplo, linear ou RBF) baseado na performance inicial. Os dados serão escalonados, pois SVM é sensível à escala das variáveis. Avaliaremos o desempenho do modelo utilizando as métricas de acurácia, precisão, recall e F1-score no conjunto de validação.

2.3.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) é um algoritmo de aprendizado supervisionado simples e intuitivo utilizado para problemas de classificação e regressão. O princípio básico do KNN é que uma amostra é classificada com base na maioria dos votos dos seus k vizinhos mais próximos no espaço de características. O valor de k é um hiperparâmetro que precisa ser ajustado para otimizar o desempenho do modelo.

Utilizaremos o KNN para prever a classificação de autismo. Antes de aplicar o modelo, os dados serão escalonados para garantir que todas as características tenham o mesmo peso na definição dos vizinhos mais próximos. Testaremos diferentes valores de k para encontrar o que oferece o melhor desempenho. Avaliaremos o modelo usando o conjunto de validação e métricas como acurácia, precisão, recall e F1-score para determinar sua eficácia.

2.4 Odds ratio e interpretação dos coeficientes

O Odds Ratio (OR) e a interpretação dos coeficientes são aspectos fundamentais na análise de modelos de regressão logística, particularmente úteis para entender a relação entre as variáveis independentes e a variável dependente binária.

Formulação Matemática: Para um modelo de regressão logística dado por:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

onde p é a probabilidade do evento de interesse (por exemplo, diagnóstico de autismo),

o Odds Ratio para um coeficiente β_i é calculado como:

$$\text{OR} = e^{\beta_i}$$

Interpretação de Coeficientes:

- Intercepto (β_0): Este coeficiente representa o log-odds do evento quando todas as variáveis explicativas são zero.
- Coeficientes (β_i): Estes coeficientes representam a mudança no log-odds do evento para um aumento de uma unidade na variável independente correspondente, mantendo todas as outras constantes.

Interpretação do Odds Ratio:

- Um OR > 1 indica que o evento é mais provável à medida que a variável independente aumenta.
- Um OR <1 indica que o evento é menos provável à medida que a variável independente aumenta.
- Um OR $= 1$ sugere que a variável independente não tem efeito sobre as chances do evento ocorrer.

Suponha que um dos coeficientes, β_1, do modelo de regressão logística seja 0.5. O Odds Ratio associado a esse coeficiente seria:

$$\text{OR} = e^{0.5} \approx 1.65$$

Isso significa que, para cada aumento de uma unidade em x_1, as odds do evento ocorrer aumentam em 65

Esta análise é particularmente valiosa para identificar quais características são mais significativas na determinação do resultado de interesse e para guiar intervenções focadas ou estratégias de prevenção baseadas em evidências específicas identificadas através do modelo.

2.5 P-Valor

O p-valor é uma medida estatística que ajuda a determinar se os resultados observados em seus dados são estatisticamente significativos. No contexto dos modelos de regressão logística, é usado para testar a hipótese nula de que um coeficiente é igual a zero.

Matematicamente, para uma estatística de teste T e um valor observado t,

$$\text{p-valor} = P(T \geq t \mid H_0)$$

para testes unilaterais, ou

$$\text{p-valor} = P(|T| \geq |t| \mid H_0)$$

para testes bilaterais

O p-valor ele é derivado da distribuição da estatística de teste, como a t-distribuição no caso de coeficientes de regressão. O p-valor é calculado comparando o valor da estatística de teste com uma distribuição de referência sob a hipótese nula.

Um exemplo de uso é que se temos p-valor menor que 0,05 (usualmente usado como limite de significância) sugere que há evidências estatísticas suficientes para rejeitar a hipótese nula em favor da hipótese alternativa, indicando que o coeficiente em questão tem um efeito significativo sobre a variável dependente.

2.6 Intervalo de Confiança

Os intervalos de confiança (IC) para os coeficientes de um modelo de regressão fornecem uma faixa de valores estimados que, com uma determinada probabilidade, contêm o verdadeiro valor do coeficiente.

Para um coeficiente β, o intervalo de confiança a 95

$$\beta \pm 1.96 \times SE(\beta)$$

onde $SE(\beta)$ é o erro padrão do coeficiente.

Um intervalo de confiança que não inclui zero sugere que o coeficiente é estatisticamente significativo ao nível de confiança escolhido (frequentemente 95%).

2.7 Métricas de Desempenho

Após ser criado nosso modelo de classificação, precisaremos de métricas para avaliar o quão bom é o nosso modelo, para este vamos usar, acurácia, precisão, recall, f1 - score, roc e auc
Para entender os cálculos de acurácia, precisão e recall precisamos entender o que é a matriz de confusão e o que compõe ela

Uma matriz de confusão é uma tabela que indica os erros e acertos do modelo, comparando com o resultado esperado

Tabela 2: Matriz de Confusão

	1 - Real	**0 - Real**
1 - Previsto	Verdadeiros Positivos (VP)	Falso Positivo (FP)
0 - Previsto	Falsos Negativo (FN)	Verdadeiros Negativo (VN)

- **Acurácia:** Mede a proporção total de previsões corretas, ou seja, é a quantidade de acertos do nosso modelo divido pelo total da amostra.

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

- **Precisão:** Importante quando o custo de um falso positivo é alto, ou seja, de todos os dados classificados como positivos, quantos são realmente positivos

$$\text{Precisão} = \frac{TP}{TP + FP}$$

- **Recall:** Crítico quando é essencial detectar todos os casos positivos, ou seja, qual a porcentagem de dados classificados como positivos comparado com a quantidade real de positivos que existem em nossa amostra.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** Combina precisão e recall, assim essa métrica une precisão e recall afim de trazer um número único que determine a qualidade geral do nosso modelo

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

em uma única métrica que é útil quando você precisa de um equilíbrio entre precisão e recall, especialmente se há um desequilíbrio de classe.

A Curva de Característica de Operação do Receptor (ROC) e a Área Sob a Curva (AUC) são métricas utilizadas para avaliar o desempenho de modelos de classificação binária.

- Curva ROC: Plota a taxa de verdadeiro positivo (Sensibilidade) contra a taxa de falso positivo (1 - Especificidade) para diferentes limiares de decisão.

- AUC: É a área sob a curva ROC.

$$\text{Sensibilidade} = \frac{TP}{TP + FN}$$

$$\text{1 - Especificidade} = \frac{FP}{TN + FP}$$

AUC varia entre 0 e 1, onde 1 indica um modelo perfeito e 0.5 um modelo que não tem capacidade discriminativa, equivalente a uma escolha aleatória.

Essas métricas fornecem um quadro robusto para avaliar e comparar modelos de classificação, permitindo aos pesquisadores escolher o modelo que melhor se adequa ao contexto específico de suas investigações.

3 Resultados

Vamos apresentar os resultados de todos os modelos, destacando a regressão logística e o random forest, pois foram os mais estudados e utilizados neste trabalho. Os demais modelos serão utilizados para fins de comparação.

3.1 Interpretação dos Coeficientes P-Valor

Dando um summary() no R vamos obter algumas informações sobre os coeficientes do modelo, a estimativa do coeficiente para o modelo, o erro padrão, o z valor e o p-valor, que tem alguns significados:

- Estimate: São os coeficientes de regressão associados a cada variável. Um coeficiente positivo sugere que um aumento na variável está associado a um aumento na probabilidade do evento de interesse ocorrer (dependendo do contexto do modelo). Um coeficiente negativo sugere o contrário.

- Std. Error: O erro padrão dos coeficientes. Um erro padrão menor indica maior precisão da estimativa do coeficiente.

- z value:É a razão entre a estimativa do coeficiente e o seu erro padrão. É usado para testar a hipótese nula de que o coeficiente é igual a zero (ou seja, a variável não tem efeito).

- $Pr(>|z|)$**:** É o p-valor associado ao teste de que o coeficiente é zero.

Tabela 3: Coeficientes da Regressão

	Estimate	**Std. Error**	**z value**	$Pr(>\|z\|)$
(Intercept)	-6.195682	0.714790	-8.668	$< 2e-16$ ***
pontuacao1	-0.056320	0.341385	-0.165	0.868964
pontuacao2	0.383149	0.360182	1.064	0.287435
pontuacao3	0.583812	0.403189	1.447	0.147222
pontuacao4	0.988882	0.360287	2.739	0.006120 **
pontuacao5	0.615119	0.330119	1.863	0.06289 .
pontuacao6	0.661537	0.282600	3.764	0.000168 ***
pontuacao7	0.339307	0.327982	1.034	0.301391
pontuacao8	0.592069	0.285204	2.076	0.037899 *
pontuacao9	0.514318	0.476643	1.079	0.280911
pontuacao10	0.113402	0.477898	0.237	0.812429
idade	0.00427540	0.002680	1.594	0.110977
genero	-0.040342	0.253395	-0.159	0.873505
etnia	-0.030123	0.211594	-0.142	0.887224
icteria	-0.116500	0.257441	-0.453	0.650887
familia	0.279937	0.235922	1.186	0.235636
triagem_antes	0.199872	0.442435	0.451	0.652093
país	0.014046	0.031042	0.453	0.65033
pontuacao_teste	0.064817	0.036654	1.768	0.077002 .
relação	0.218357	0.187861	1.162	0.245102

Assim, de acordo com esses resultados, o 'Intercept' tem um coeficiente significativamente negativo, indicando a log-odds de baseline para o evento de interesse ser baixa. Já as variáveis como pontuacao4, pontuacao6 e pontuacao8 são significativos pelo o p-valor pois eles estão menores que 0,05.

3.2 Odds Ratio e seu intervalo de confiança

Para calcular o OR tiramos o exponencial dos coeficientes do modelo e usamos eles para fazer o nosso intervalo de confiança do OR:

Tabela 4: Odds Ratios e seus Intervalos de Confiança

	OR	**2.5 %**	**97.5 %**
(Intercept)	0.002038213	0.0004608599	0.007693711
pontuacao1	0.945236792	0.482213218	1.849138276
pontuacao2	1.466966660	0.789954452	2.724421932
pontuacao3	1.791975783	0.830643874	3.959603973
pontuacao4	2.266590156	1.354304853	3.791627089
pontuacao5	1.84387082	0.965491378	3.537707571
pontuacao6	1.608380245	0.734971504	3.30845675
pontuacao7	1.432335997	0.715219369	2.868953691
pontuacao8	1.807724876	0.793576792	4.066817099
pontuacao9	3.259701613	1.343842409	8.854128318
pontuacao10	1.115034504	0.47210848	2.548182322
idade	1.00427540	0.996045187	1.012544888
genero	1.00837410	0.581337304	1.749152024

Continua na próxima página

Tabela 4 continuação da página anterior

	OR	2.5 %	97.5 %
etnia	0.984010597	0.823349141	1.176509793
icteria	0.875895362	0.504522472	1.518832841
familia	1.432097504	0.781552955	2.63760668
triagem_antes	3.221477592	2.357140713	5.285731785
país	0.98553594	0.965467191	1.00534866
pontuacao_teste	1.066695331	0.994359364	1.148739587
relação	1.244030621	0.8434693964	1.774551412

O 2.5 % e 97.5 % são os limites do intervalo de confiança de 95% para o OR. Se este intervalo incluir 1, então o efeito da variável não é estatisticamente significativo ao nível de 0.05.

3.3 Intervalo de Confiança para os coeficientes

Dando um confit() no R ele vai nos dar o intervalo de confiança dos coeficientes

Tabela 5: Intervalos de Confiança para os coeficientes

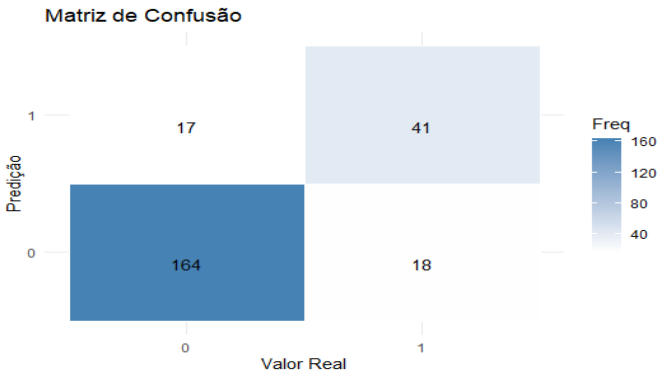
	2.5 %	97.5 %
(Intercept)	-7.682416475	-4.867352008
pontuacao1	-0.729310623	0.614719734
pontuacao2	0.316636687	1.103439254
pontuacao3	-0.185350517	1.375144013
pontuacao4	0.294277167	1.816854897
pontuacao5	-0.035114684	1.263350442
pontuacao6	0.511331453	1.623568549
pontuacao7	-0.286041539	1.093225207
pontuacao8	-0.124448172	1.015944802
pontuacao9	0.288739755	1.880889698
pontuacao10	-0.801729412	0.540278672
idade	-0.003964329	0.042521714
genero	-0.539322912	0.345298489
etnia	-0.124110046	0.429864726
icteria	-0.020546732	0.505530765
familia	-0.246468950	0.850468590
triagem_antes	-0.628598039	0.306832383
país	-0.035141328	0.30353035
pontuação_teste	-0.005659388	0.138665303
relação	-0.170231659	0.573457665

3.4 Métricas de Desempenho

3.4.1 Regressão Logistica

Para alcançar os resultados desejados com nosso modelo, o dividiremos em dois conjuntos distintos: um para treinamento e outro para teste. Essa divisão é crucial tanto para calcular eficazmente a performance do modelo quanto para prevenir o overfitting. O overfitting ocorre quando o modelo é excessivamente ajustado a um conjunto específico de dados, funcionando bem apenas para esses dados e falhando ao ser aplicado a outros conjuntos. Por isso, a separação em dois grupos é fundamental para garantir a generalização do modelo.

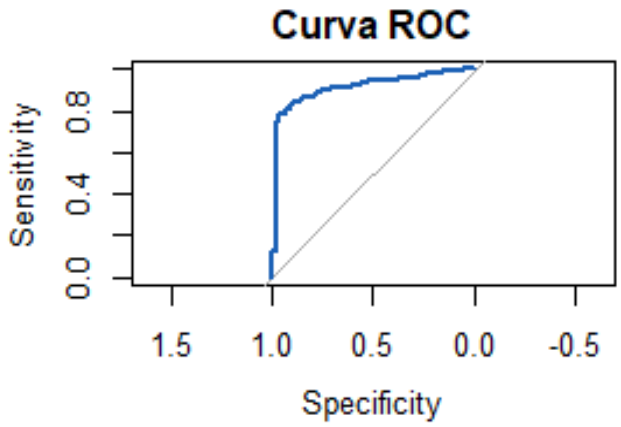
Primeiramente vamos visualizar nossa matriz de confusão que nos ajudará a calcular algumas métricas:



Usando as formulas para calcular, acuracia, precisão, recall e F1 Score temos os seguintes resultados para esse modelo (obviavente tudo calculado pelo R para melhor precisão dos calculos):

$$\textbf{Acurácia} = 0,8583 \tag{3.1}$$
$$\textbf{Recall} = 0,9060 \tag{3.2}$$
$$\textbf{Precisão} = 0,9060 \tag{3.3}$$
$$\textbf{F1-Score} = 0,90 \tag{3.4}$$
$$\textbf{AUC} = 0,913 \tag{3.5}$$

Curva ROC:



Com base nas métricas avaliadas, podemos considerar que temos um modelo robusto, especialmente considerando a complexidade inerente à previsão de autismo. A análise inicial sugere um bom ajuste, mas realizamos experimentos adicionais removendo algumas variáveis que, de acordo com a análise dos coeficientes de correlação, pareciam ter pouco impacto na variável resposta.

Interessantemente, ao remover variáveis como gênero, etnia, histórico familiar e triagem prévia, observamos uma deterioração nas métricas e nos resultados do modelo. Isso indica que, apesar das baixas correlações isoladas, essas variáveis contribuem significativamente para o modelo, possivelmente devido a interações complexas entre elas que influenciam a previsão do autismo.

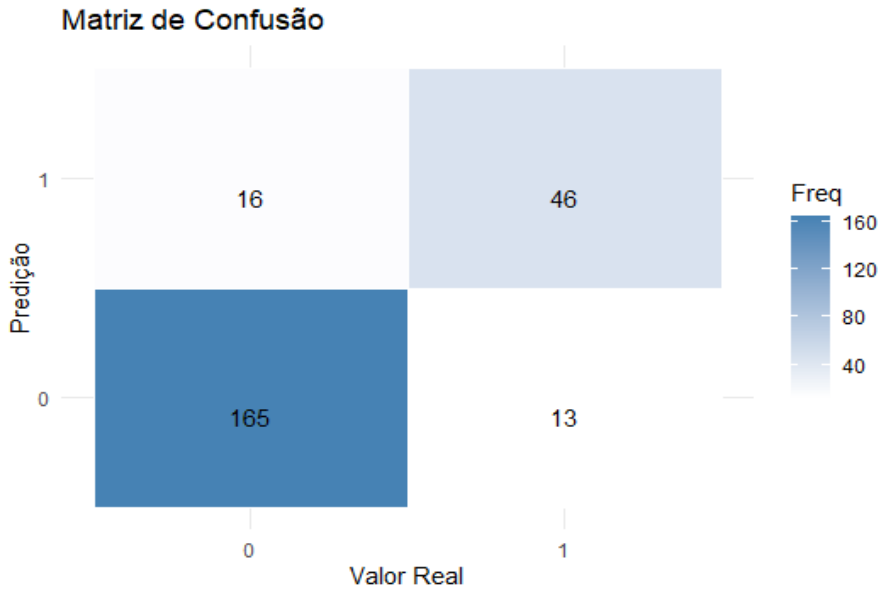
Por outro lado, a exclusão de variáveis como país, pontuação geral e tipo de relação resultou em melhorias nas métricas do modelo. Isso era esperado para a pontuação geral e a variável relação, que aparentemente resumem ou não adicionam informações significativas além das já capturadas pelas pontuações individuais de 1 a 10. A remoção da variável país, que inclui 55 categorias distintas, também melhorou o desempenho, provavelmente porque a diversidade e a distribuição desigual dos dados por país não foram adequadamente aprendidas pelo modelo, levando a um treinamento menos eficaz.

Esses resultados realçam a importância de uma análise detalhada de cada variável no contexto do modelo global, destacando como a inclusão ou exclusão de certas variáveis pode influenciar de maneira significativa a

precisão das previsões. Continuaremos ajustando o modelo para melhorar ainda mais sua eficácia, utilizando essas descobertas para guiar nossas decisões na seleção de variáveis.

Assim, o melhor ajuste do modelo nos deu as seguintes métricas:

$$\textbf{Acurácia} = 0,87916 \tag{3.6}$$
$$\textbf{Recall} = 0,9116 \tag{3.7}$$
$$\textbf{Precisão} = 0,9269 \tag{3.8}$$
$$\textbf{F1-Score} = 0,919 \tag{3.9}$$
$$\textbf{AUC} = 0,915 \tag{3.10}$$

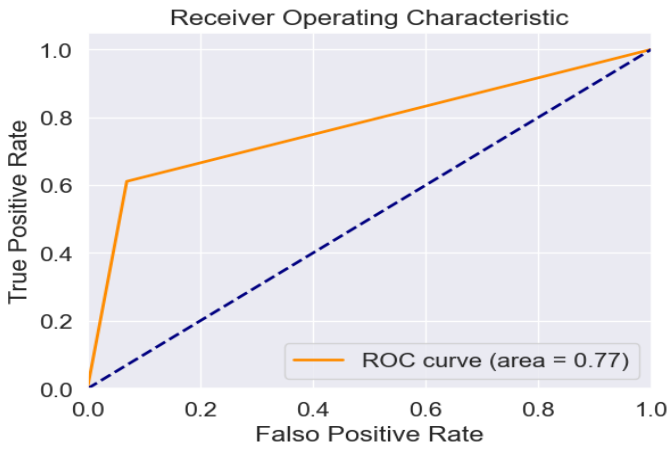


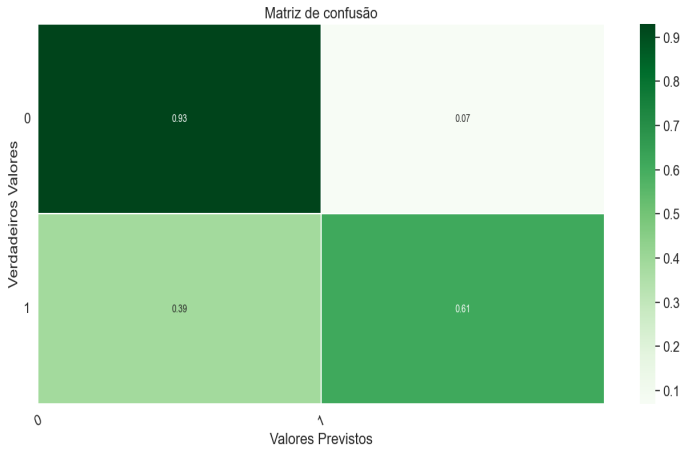
3.4.2 Random Forest

Este modelo foi desenvolvido em Python e, inicialmente, vamos apresentar os resultados sem o tratamento de dados que apliquei para aprimorar a regressão logística.

Tabela 6: Métricas Random Forest

Modelo	Classe	Precisão	Recall	F1-score
Random Forest	0	0.89	0.90	0.89
	1	0.62	0.56	0.59
Acurácia			0.86	
AUC			0.77	

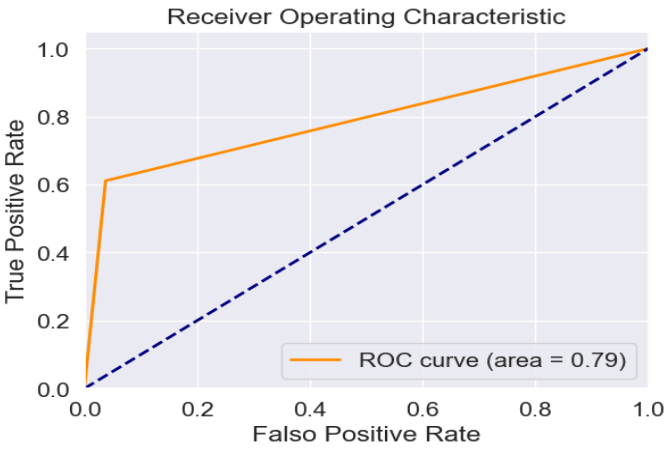


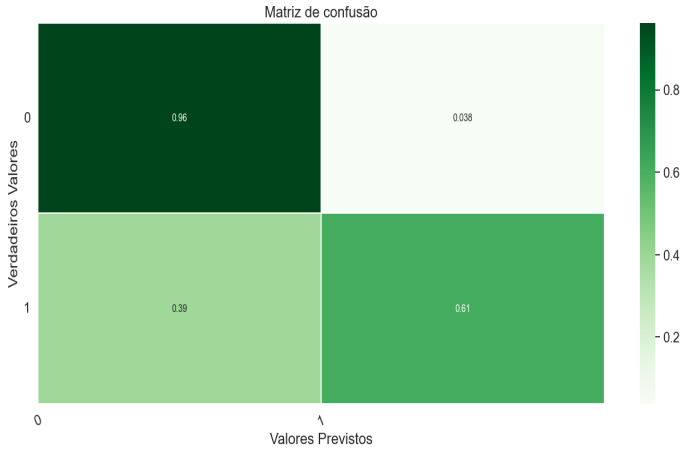


Agora aplicando o tratamento que foi feito em regressão as métricas mudaram para:

Tabela 7: Métricas Random Forest

Modelo	Classe	Precisão	Recall	F1-score
Random Forest	0	0.90	0.96	0.93
	1	0.82	0.61	0.70
Acurácia			0.88	
AUC			0.79	





3.4.3 Resultados dos outros modelos

Tabela 8: Métricas Outros Modelos

Modelo	Classe	Precisão	Recall	F1-score
GMB	0	0.88	0.90	0.89
	1	0.62	0.56	0.59
Acurácia			0.825	
AUC			0.79	
GMB Light	0	0.88	0.90	0.89
	1	0.62	0.56	0.59
Acurácia			0.86	
AUC			0.79	
SVM	0	0.79	0.97	0.87
	1	0.50	0.09	0.16
Acurácia			0.775	
AUC			0.79	
KNN	0	0.84	0.90	0.87
	1	0.55	0.43	0.48
Acurácia			0.79	
AUC			0.79	

Com as mudanças temos:

Tabela 9: Métricas Outros Modelos

Modelo	Classe	Precisão	Recall	F1-score
GMB	0	0.88	0.93	0.91
	1	0.70	0.57	0.63
Acurácia			0.85	
AUC			0.752	
GMB Light	0	0.88	0.90	0.89
	1	0.62	0.56	0.59
Acurácia			0.87	
AUC			0.769	
SVM	0	0.78	1.00	0.87
	1	0.00	0.00	0.00
Acurácia			0.78	
AUC			0.5	
KNN	0	0.87	0.91	0.89
	1	0.64	0.54	0.59
Acurácia			0.829	
AUC			0.725	

4 Discussão

Este estudo investigou a influência de variáveis demográficas e de triagem no diagnóstico do Transtorno do Espectro Autista (TEA) utilizando técnicas avançadas de modelagem estatística, como regressão logística e Random Forest, além de outras abordagens de machine learning. Nosso objetivo principal foi identificar as variáveis que mais impactam o diagnóstico do TEA e determinar o método de triagem mais eficaz.

Inicialmente, uma análise exploratória detalhada revelou correlações significativas entre variáveis demográficas e as pontuações do Autism Spectrum Quotient (AQ). Essas descobertas orientaram a seleção e o ajuste dos modelos de predição. Posteriormente, aplicamos técnicas de modelagem para avaliar a influência de cada variável no diagnóstico, removendo sequencialmente as variáveis menos impactantes.

Curiosamente, a remoção das pontuações detalhadas do teste de triagem mostrou-se benéfica, indicando que a pontuação agregada era redundante e até prejudicial ao modelo, ressaltando a importância da seleção criteriosa de variáveis. A eliminação das variáveis de pontuação detalhada (de 1 a 10) proporcionou uma análise mais granular, tornando a pontuação agregada desnecessária e até prejudicial ao modelo. A variável relacionada ao histórico familiar de autismo mostrou-se mais informativa do que a variável de relação, justificando sua exclusão. Além disso, a remoção da variável de relação foi vantajosa, pois era um derivativo de histórico complicado e pouco relevante. A remoção da variável país, devido à grande diversidade e distribuição desigual dos dados, melhorou a eficácia do treinamento, com a variável etnia capturando as influências culturais e regionais necessárias. Essas modificações ressaltam a importância de uma seleção criteriosa de variáveis em modelos preditivos, visando incluir apenas aquelas mais informativas e relevantes para maximizar a precisão e a eficiência.

Os modelos de regressão logística e Random Forest foram os mais eficazes em nosso estudo. O modelo de regressão logística destacou-se em termos de AUC, indicando uma performance geral superior. Esse modelo alcançou uma acurácia de 0.87916, recall de 0.9116, precisão de 0.9269, F1-score de 0.919 e AUC de 0.915. Por outro lado, o modelo de Random Forest também apresentou um desempenho notável, especialmente após a aplicação de tratamentos específicos. Inicialmente, o modelo apresentou uma acurácia de 0.86 e AUC de 0.77. Com as melhorias aplicadas, essas métricas aumentaram para uma acurácia de 0.88 e AUC de 0.79.

Esses modelos capturaram de forma eficaz as complexidades associadas ao diagnóstico de TEA, evidenciando a utilidade dessas técnicas avançadas em configurações médicas e psicológicas. A análise dos coeficientes no modelo de regressão logística, por exemplo, revelou a importância de variáveis como histórico familiar de autismo e pontuações detalhadas de triagem, que proporcionaram uma análise mais granular e informativa. Comparando com outros modelos, como Gradient Boosting Machines (GBM), Support Vector Machines (SVM) e K-Nearest Neighbors (KNN), a regressão logística e o Random Forest mantiveram uma performance

superior. O GBM, por exemplo, apresentou uma acurácia de 0.85 e AUC de 0.752 após ajustes, enquanto o SVM teve uma acurácia de 0.78 e AUC de 0.5, e o KNN alcançou uma acurácia de 0.829 e AUC de 0.725.

Esses resultados evidenciam que, embora todos os modelos tenham suas próprias vantagens, a regressão logística e o Random Forest se destacaram pela capacidade de lidar com a complexidade dos dados e fornecer previsões mais precisas e eficientes para o diagnóstico de TEA.

Embora nosso estudo tenha fornecido resultados valiosos, ele possui limitações, como o tamanho da amostra, que pode restringir a generalização dos resultados. Para pesquisas futuras, recomendamos a expansão do conjunto de dados com uma amostra mais diversificada e a inclusão de mais variáveis que possam influenciar o diagnóstico de TEA. Além disso, explorar modelos que integrem técnicas de aprendizado de máquina para identificar interações complexas entre as variáveis poderia proporcionar uma compreensão ainda mais profunda do TEA.

Este trabalho não só demonstra a eficácia das técnicas avançadas de modelagem estatística no diagnóstico do autismo, mas também oferece uma base sólida para futuras investigações que busquem aprimorar as práticas de diagnóstico e intervenção dentro do espectro autista.

Referências

[1] Kaggle Dataset. Autism Prediction. Available at: `https://www.kaggle.com/competitions/autismdiagnosis/data`

[2] Gelman, A., Hill, J., Vehtari, A. (2020). Regression and Other Stories. Cambridge University Press.

[3] McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and Stan. Chapman and Hall/CRC.

[4] Gelman, A., Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

[5] The Elements of Statistical Learning"por Trevor Hastie, Robert Tibshirani e Jerome Friedman

[6] Pattern Recognition and Machine Learning"por Christopher M. Bishop

[7] Machine Learning: A Probabilistic Perspective"por Kevin P. Murphy

[8] Applied Predictive Modeling"por Max Kuhn e Kjell Johnson

[9] Data Mining: Practical Machine Learning Tools and Techniques"por Ian H. Witten, Eibe Frank, e Mark A. Hall

5 Apêndice

O Github deste trabalho com códigos e tudo mais se encontra neste link: `https://github.com/adrianfilipe/Modelagem-Estatistica-Do-Autismo/tree/main`