Análise dos Fatores Socioambientais e Comportamentais na Identificação do Transtorno Obsessivo Compulsivo: Uma Abordagem com Dados da Pesquisa Nacional de Saúde 2019

Anna Puga Campos Rodrigues [1], **Luiz Enrique Zárate Galvez**[1]

[1]Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Av Dom José Gaspar 500, Belo Horizonte, Minas Gerais, 30535-610

annapugac@gmail.com, zarate@pucminas.br

Abstract. *Obsessive-Compulsive Disorder (OCD) is a mental distress characterized by the presence of obsessions and compulsions that significantly affect individuals' lives, as described in the DSM-5 manual. This work explores the analysis of OCD using data from the 2019 National Health Survey (PNS), addressing socio-environmental and behavioral aspects. Using the Explainable Boosting Machine (EBM) algorithm and a Decision Tree, the study identifies relevant variables for the classification of OCD, demonstrating the influence of socio-environmental factors in the identification of the disorder. Results indicate improvements in the models' metrics with the inclusion of these variables, as well as agreement with other results in the literature.*

Resumo. *O Transtorno Obsessivo Compulsivo (TOC) é um sofrimento mental caracterizado pela presença de obsessões e compulsões que afetam significativamente a vida dos indivíduos, conforme descrito no manual DSM-5. Este trabalho explora a análise do TOC utilizando dados da Pesquisa Nacional de Saúde (PNS) 2019, abordando aspectos socioambientais e comportamentais. Utilizando o algoritmo Explainable Boosting Machine (EBM) e uma Árvore de Decisão, o estudo identifica variáveis relevantes para a classificação do TOC, demonstrando a influência de fatores socioambientais na identificação do transtorno. Resultados indicam melhorias nas métricas dos modelos com a inclusão dessas variáveis, assim como concordância com outros resultados da literatura.*

1. Introdução

O Transtorno Obsessivo Compulsivo (TOC), estabelecido pelo *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* [Association 2013], publicado pelo *American Psychiatric Association*, é caracterizado pela presença de obsessões, compulsões ou ambas no dia-a-dia da pessoa avaliada. De acordo com o manual, obsessões são definidas por pensamentos e impulsos recorrentes que o indivíduo, de forma consciente, tenta afastar ou neutralizar através de outros pensamentos e ações. Estas tentativas de suavização dos impulsos são categorizadas como compulsões, ou seja, atos repetitivos ou atos mentais como resposta do indivíduo à uma obsessão ou sensação de obrigação para agir de acordo com regras inflexíveis e imutáveis. Este sofrimento mental causa prejuízo no convívio social, profissional e em outros aspectos da vida do indivíduo.

Após um estudo preliminar exploratório, percebe-se que os esforços para compreender o TOC, em geral, abrangem três vertentes: 1) entendimento do funcionamento do

cérebro sob uma perspectiva química e biológica; 2) a interligação dos sintomas a outros transtornos psiquiátricos; e 3) análise dos aspectos comportamentais da pessoa. Durante o estudo, viu-se que maioria das investigações sobre o tema concentra-se na primeira vertente, onde abordagens de classificação foram utilizadas para identificar pacientes com TOC com base em imagens do cérebro. Na segunda vertente analisam-se sintomas similares ou sobrepostos do TOC com outros distúrbios, com o objetivo de reduzir o diagnóstico equivocado. A terceira abordagem é utilizada em grande parte dos trabalhos com foco em um único aspecto comportamental, apresentando uma visão expositiva da existência de tal prática ou sintoma em uma pessoa já anteriormente diagnosticada com TOC, ou da interferência de tal aspecto na qualidade de vida.

Na literatura, existe uma menor quantidade de estudos atuais focados na análise do perfil global de pessoas com TOC, incluindo as características socioambientais da vida do indivíduo, principalmente em relação à população brasileira. Porém, há evidência da correlação de eventos traumáticos e fatores sociais na apresentação dos sintomas da condição. Foram encontrados, por exemplo, potenciais fatores de risco em áreas amplas, como complicações perinatais, ciclo reprodutivo e eventos estressantes da vida [Brander et al. 2016].

Reconhecida a limitação dos estudos nesta área, o presente trabalho busca contribuir com a expansão da área de pesquisa das características gerais do TOC, buscando responder se fatores socioambientais impactam o diagnóstico do TOC no Brasil e, caso a resposta seja positiva, quais os fatores mais relevantes. Ademais, pretende-se responder também se fatores comportamentais descritos na literatura como ligados ao TOC se mostram presentes na classificação.

Para isso, entrevistas sistêmicas com profissionais da área da saúde mental foram conduzidas. Nelas, realizamos perguntas diretas sobre a identificação das principais dimensões e aspectos ligados ao transtorno, com o objetivo entender mais sobre a abordagem clínica para diagnóstico. Tais conversas trouxeram compreensões diferentes sobre o problema e sobre a forma como cada especialista compreende os sintomas e os gatilhos do TOC. A partir disso um mapa conceitual foi formulado, buscando juntar todas as perspectivas em uma só visualização e guiar o processo de descoberta de conhecimento para aspectos específicos de um paciente com o transtorno.

Em seguida, foram utilizados os dados da Pesquisa Nacional de Saúde (PNS) 2019 [Instituto Brasileiro de Geografia e Estatística (IBGE) 2020], levantamento populacional do Brasil, e a utilização de modelos de Classificação para interpretação das relações entre aspectos relevantes. O trabalho considera os indivíduos que indicaram durante a pesquisa que foram, ou não, diagnosticados com TOC. As informações englobam características do domicílio e do trabalho, hábitos de saúde, relação familiar, dentre outros. Ao focar na população brasileira, este estudo busca revelar padrões de comportamento e especificidades socioambientais que possam ser exclusivos dessa população, e que não necessariamente afetam pessoas com TOC em outros países. Dessa forma, espera-se gerar um conhecimento mais profundo e contextualizado sobre o TOC no Brasil.

A partir dos dados da PNS, foram realizadas transformações para gerar atributos com maior poder informativo relacionados às informações do mapa conceitual e ao conhecimento repassados pelos especialistas entrevistados, criando assim um novo

conjunto de dados. Em seguida, o algoritmo *Explainable Boosting Machine* (EBM) [The InterpretML Contributors 2023] foi utilizado com a nova base para interpretar a importância de cada atributo para a classificação. Finalmente, as regras geradas por uma Árvore de Decisão própria da plataforma *KNIME* [KNIME 2023] foram avaliadas para verificar a presença de padrões comportamentais mencionados na literatura e determinar se questões socioambientais podem, também, influenciar a identificação do transtorno.

Ao final, a utilização conjunta de atributos socioambientais e comportamentais resultou em um aumento da *F1-score* de 30,8% para o EBM e de 24,3% para a Árvore de Decisão, em comparação com a utilização apenas dos atributos comportamentais. Foram encontrados, também, resultados coerentes com a literatura, como a relação entre TOC e a baixa qualidade de sono [Segalàs et al. 2021], assim como a conexão entre a ausência de outros sofrimentos mentais e a presença do TOC.

Assim, pode-se afirmar que este estudo engloba partes tanto da segunda quanto da terceira abordagem de pesquisa citadas anteriormente, pois correlaciona a presença de outros transtornos mentais com o TOC e, também, utiliza de características comportamentais para classificação do transtorno.

Este trabalho está estruturado em quatro seções principais, além desta introdução: Revisão de Literatura, que aborda as principais linhas de pesquisa sobre o TOC, Metodologia, que abordará a montagem do mapa conceitual e do domínio do problema, as transformações de dados feitas e a experimentação com os modelos EBM e Árvore de Decisão, Resultados, Conclusões e Trabalhos Futuros.

2. Revisão de Literatura

Durante estudos preliminares, foram percebidas três principais vertentes de estudo do TOC. Na neuro anatomia, o objetivo principal é analisar as causas do TOC. [Hu et al. 2016] utilizaram a Análise de Padrão Multivariado (MVPA) sobre imagens estruturais de alta dimensão para discriminar entre pacientes com TOC e sujeitos saudáveis (HCS). Usando classificadores *Support Vector Machine* (SVM) e *Gaussian Process Classifier* (GPC), analisaram as diferenças no volume de matéria cinzenta (GM) e branca (WM). As análises demonstraram uma acurácia acima de 75% para ambos os classificadores, mostrando que características anatômicas de GM e WM são úteis na diferenciação de pacientes com TOC e HCS.

Na análise da interligação do TOC com outros distúrbios psiquiátricos, destacam-se o espectro autista, transtorno de ansiedade generalizada (TAG), e transtornos alimentares (TAs). [Højgaard et al. 2023] buscaram diferenciar crianças com TOC e traços autísticos subclínicos daquelas com TOC sem esses traços. A presença de TDAH e transtornos de tiques, assim como sintomas de TOC relacionados à organização, foram significativamente associados à presença de TOC com traços autísticos. Em [Bang et al. 2020] os autores avaliaram a presença de sintomas de TAs em pacientes com TOC. Os resultados sugerem que um subconjunto considerável com TOC pode ter um TA clínico ou estar em alto risco de desenvolver.

Nos estudos sobre características comportamentais, avaliam-se sintomas específicos e seu impacto na qualidade de vida. [Segalàs et al. 2021] mediram a qualidade do sono de pacientes com TOC e grupo controle usando regressão linear múltipla. Pacientes com TOC apresentaram baixa qualidade do sono e mais distúrbios comparados ao

grupo controle. Sintomas de depressão e traços de ansiedade gerados pelo TOC estavam correlacionados com a baixa qualidade do sono.

Finalmente, características socioambientais também são estudadas. Boger e Werner analisaram o impacto dos maus-tratos na infância na presença e severidade do TOC no futuro e concluíram que a presença de abuso na infância está relacionada a maior severidade de sintomas de TOC na vida adulta [Sabrina Boger and Werner 2020].

Ademais, estudos que utilizam técnicas de *Machine Learning* para classificação e estudo de transtornos mentais e condições de saúde são comuns na literatura. O estudo de [Souza et al. 2020] explora técnicas de *Deep Learning* para desenvolver um classificador para a identificação automática de depressão, ansiedade e suas comorbidades, a partir de um conjunto de dados extraídos do Reddit. Além disso, [Cazzolato et al. 2021] fazem a análise de similaridade e correlações entre tuplas e atributos de Registros Eletrônicos de Saúde (EHRs) de pacientes com COVID-19.

Conclui-se que existem diversos estudos sobre causas do TOC e características singulares que interferem no diagnóstico. No entanto, há poucos trabalhos com uma visão global e sistêmica dos sintomas do TOC, especialmente na população brasileira. Este trabalho pretende contribuir para essa abordagem, explorando características socioambientais correlacionadas ao TOC.

3. Metodologia

A fim de identificar as variáveis mais relevantes para a identificação do TOC na base de dados PNS 2019, foram executados diversos passos do processo de *Knowledge Discovery in Databases* (KDD) descritos pelo método PICTOREA [Montevecchi and Zárate 2014], e utilizados em conjunto com o método CAPTO [Zarate et al. 2023], ambos processos de descoberta de conhecimento e entendimento de domínio em Ciência de Dados. Estas etapas auxiliaram na extração de conhecimentos significativos dos dados que possam ser essenciais para a identificação e compreensão do TOC e proporcionaram uma base sólida para a interpretação dos fatores-chave associados ao TOC.

3.1. Domínio do Problema, Espaço Solução e Mapa Conceitual

O objetivo inicial do processo de descoberta de conhecimento é a definição de um Mapa Conceitual (MC) para representação do domínio do problema. O mapa é uma representação gráfica das áreas globais do perfil de uma pessoa com TOC e suas relações, com o fim de organizar informações de forma hierárquica e ajudar a visualizar a estrutura do problema. Após a consolidação do MC, o utilizamos para a transformação dos atributos da base original, gerando uma base menor e mais concisa, com maior poder de informação.

Para iniciar o MC, é necessária a exploração do Espaço Problema ep e a definição do Espaço Solução es. A exploração do ep é a pesquisa sobre o tópico desejado e a definição do problema a ser trabalhado. A definição do es é a especificação das técnicas que serão executadas sobre ep, seguida da definição das expectativas sobre o resultado e as saídas esperadas. Neste passo, o método selecionado foi a Classificação.

Para o primeiro passo, foram feitas entrevistas com especialistas no domínio de transtornos psicológicos, tanto da área da psicologia quanto da psiquiatria. Durante os questionamentos, abordamos as principais características comportamentais de um indivíduo com TOC e como elas podem ser utilizadas para a identificação do transtorno, assim

como características socioambientais que influenciam na apresentação dos sintomas da condição. Estas discussões geraram dimensões e visões alternativas sobre o problema, portanto cada perspectiva foi comparada com a dos outros profissionais, gerando o MC unificado. As dimensões do MC podem ser vistas na Tabela descritiva (Tabela 1).

Esta Tabela contém as Dimensões, Aspectos e Atributos associados ao domínio "Transtorno Obsessivo-Compulsivo (TOC)". Este é o último passo do método PICTO-REA, ou seja, a caracterização do problema através de atributos. Durante este passo, os atributos da base de dados foram mapeados para cada uma das dimensões, e estas foram subdivididas em aspectos relevantes para entender as variáveis da base que são ligadas ao TOC na realidade.

Domínio de problema: Transtorno Obsessivo Compulsivo (TOC)		
Dimensão: Critérios Clínicos de Diagnóstico DSM-5		
Aspectos	**Atributos**	**Atributos Mapeados**
Doenças Crônicas	Tipo, Severidade	Q00201, Q03001, Q060, Q06306, Q068, Q074, Q079, Q088, Q11604, Q120, Q124
Sofrimentos Mentais Diagnosticados	Tipo, Severidade	Q092, Q11006, Q11007, Q11008, Q11010
Medicação	Tipo, Quantidade, Frequência	Módulo Q - Doenças crônicas
Álcool e Tabaco	Tipo, Quantidade, Frequência	ALCOOL: P027, P02801 TABACO: P050 até P053
Dimensão: Sentimentos e Gatilhos Psicológicos		
Aspectos	**Atributos**	**Atributos Mapeados**
Traumas e Violências	Presença de Violência Física, Psicológica, Sexual	Módulo V - Violência (Para pessoas de 18 anos ou mais de idade)
Trabalho	Presença de Indicadores de Vulnerabilidade	Módulo E - Características de trabalho das pessoas de 14 anos ou mais de idade
Moradia	Presença de Indicadores de Vulnerabilidade	Módulo A - Informações do Domicílio
Hábitos Sanitários	Frequência da Limpeza Bucal	U00204, U00101
Características do Indivíduo	Raça, Gênero, Idade, Estado Civil, Nível de Instrução	C006, C008, C009, C011, VDD004A
Dimensão: Estatísticas e Estudos de Instituições Globais		
Aspectos	**Atributos**	**Atributos Mapeados**
Utilização de Serviços de Saúde	Tipo, Frequência	J012
Alimentação	Indicativos de Transtorno Alimentar	P00601 até P02602
Exercício Físico	Frequência, Quantidade	P034 até P03702
Plano de Saúde	Existência de Plano de Saúde Contratado	I00102, I001021
Uso de Telas	Tempo	P04501 e P04502
Sono	Qualidade	Q132 até Q134, N010

Tabela 1. Descrição do Domínio "Transtorno Obsessivo Compulsivo (TOC)"

3.2. Transformações de Dados

A fim de gerar uma nova base de dados com variáveis preditivas com maior poder de informação, transformações de dados foram aplicadas na base original, utilizando os aspectos mapeados na Tabela de descrição do domínio e as instâncias da base com classificação de TOC definida, ou seja, Q11009 (diagnóstico de TOC) = $[1, 2]$. Foram removidos os módulos que não seriam utilizados, as instâncias sem classificação de TOC e todas as instâncias que apresentavam idade menor que 18 anos. A última remoção foi necessária pois vários dos atributos que seriam utilizados nas transformações continham valores nulos para este caso. Ao final, a base foi balanceada em relação a ambas as classes, resultando em 370 instâncias para a classe TOC = Não e 270 instâncias para a classe TOC = Sim.

Os novos atributos foram separados em duas categorias: Comportamentais e Socioambientais. As características comportamentais incluem Ingestão de Álcool, Consumo de Tabaco, Utilização de Serviços de Saúde, Frequência Limpeza Bucal, Exercício Físico, Qualidade do Sono, Uso de Telas, Qualidade da Alimentação e Uso de Medicação. Já as características socioambientais incluem Sexo, Idade, Cor/Raça, Estado Civil, Nível de Instrução, Indicadores de Violência (psicológica, física e sexual), Vulnerabilidade no Trabalho, Vulnerabilidade de Moradia, Plano de Saúde, Doenças Crônicas e Sofrimentos Mentais. As regras de transformação e fusão são demonstradas nas Tabelas 2 e 3 e a nomenclatura dos atributos pode ser encontrada no dicionário de microdados da PNS 2019.

Comportamental			
Atributos	**Regras**		**Novo Valor**
Álcool	P027	1	Nunca
	P027 P02801	2 NULL	<1x por semana
	P027 P02801	2 ou 3 0	<1x por semana
	P027 P02801	3 entre 1 e 3	1-3x por semana
	P027 P02801	3 entre 4 e 6	4-6x por semana
	P027 P02801	3 >= 7	Todo dia
Tabaco	P052	3	Nunca fumou
	P050 P052	3 1 ou 2	Já foi fumante, mas não fuma atualmente
	P050	2	Fuma ocasionalmente
	P050	1	Fuma diariamente
Frequência escovação dentária	U0024	2	<1x por dia
	U0024 U00101	1 4	<1x por dia
	U0024 U00101	1 3	1x por dia
	U0024 U00101	1 2	2x por dia
	U0024 U00101	1 1	3x ou mais por dia
Execicio Fisico	**Tempo de Exercicio por semana em horas = (P025 * P03701) + (P03702 / 60)**		
	P034 P035	2 0	Não pratica
	P034 TEMPO	1 <2	<2h por semana
	P034 TEMPO	1 entre 2 e 4	2-4h por semana
	P034 TEMPO	1 entre 4 e 6	4-6h por semana
	P034 TEMPO	1 >= 6	>= 6h por semana

Comportamental			
Atributos	**Regras**		**Novo Valor**
Qualidade do Sono	**Sistema de Pontuação: Q132 : Sim = -1 e Não = 0 Q133 : 1-3 dias: -1 , 4-7 dias: -2 , 8-14 dias: -3 Q132 : Sim = -2 e Não = -1**		
	Resultado	entre -1 e 1	Ruim
	Resultado	entre 2 e 4	Moderado
	Resultado	5	Bom
Uso de Telas	P04501 P04502	6 6	Não usa telas
	P04501 + P04502	entre 0 e 3	Baixo
	P04501 + P04502	entre 3 e 5	Moderado
	P04501 + P04502	>5	Alto
Qualidade Alimentação	**Sistema de Pontuação: P006 = 5, P00901 = 8, P01101 = 4, P013 e P015 = 4, P023 = 2, P01601 = 3, P018 = 4, P02002 = -5, P02001= -4, P02501 = -8, P02602 = -8**		
	Dias * pesos	entre -11 e 6	Muito Ruim
	Dias * pesos	entre 6 e 10	Ruim
	Dias * pesos	entre 10 e 14	Moderada
	Dias * pesos	entre 14 e 24	Boa
Medicamento Controlado	**Atributos da Tabela Descritiva**		
	Qualquer Atributo	1	Faz uso de medicamento controlado
	Todos os Atributos	!= 1	Não faz uso de medicamento controlado

Tabela 2. Transformações - Aspectos Comportamentais

Os valores dos atributos provindos de cálculos numéricos foram discretizados a partir da divisão em percentis. Cada grupo de valores dentro desses percentis foi transformado em uma categoria para o novo atributo discretizado. As características socioambientais, principalmente as de Vulnerabilidade de Trabalho e Moradia, tiveram como referência os Determinantes Sociais de Saúde [Buss and Pellegrini Filho 2007] e os estudos de [Azeredo et al. 2007] e [Pasternak 2016] para a definição de habitações irregulares.

Após a finalização das transformações, os atributos Sexo, Utilização de Serviços de Saúde, Violência Sexual e Doenças Crônicas foram removidos da base de dados, pois experimentos preliminares apontaram que a inclusão desses atributos prejudica o desempenho do modelo.

Ao final do processo, a base gerada é significativamente menor em comparação

Socioambiental			
Atributos	**Regras**		**Novo Valor**
Violências	**Psicológica = V00201 até V00205 Física = V01401 até V01405 Sexual = V02701 até V02802**		
	Todos os Atributos	2	Não
	Qualquer Atributo	1	Sim
	Resposta = Sim : Quem Causou a Violência (psi, fis, sex) = [Atributos]		
	Atributos	1 ou 2 ou 3	Relação Romântica
	Atributos	4 ou 5 ou 6 ou 7	Família
	Atributos	8	Amigo, Vizinho
	Atributos	9 ou 10	Trabalho
	Atributos	11 ou 12 ou 13	Outro
Vulnerab. Trabalho	**Variáveis Auxiliares**		
	Tem Trabalho E001, E002, E003, E004, E02601	1	Sim
	Remunerado E001, E01601, E01801	1	Sim
	Aposentado F001011	1	Sim
	Múlt. Trabalhos E011	2 ou 3	Sim
	Rendimento Total E01602, E01802, F001021	Soma	$
	Regras		
	Tem trabalho Aposentado	Não Não	Sim
	Tem trabalho Remunerado	Sim Não	Sim
	Tem trabalho Múlt. Trabalhos	Sim Sim	Sim
	Tem trabalho Múlt. Trabalhos Rendimento	Sim Não <salário min.	Sim
	Outro Caso		Não

Socioambiental			
Vulnerab. Moradia	**Variáveis Auxiliares 1. (V0022) / (A01401) 2. (V0022) / (A011)**		
	Residência Inadequada A001 OU A00210 OU A004010	3 2 ou 3 4	Sim
	Sem Saneamento Básico A01501	3 ou 4 ou 5 ou 6	Sim
	Origem da Água Inadequada A005010	3 ou 5	Sim
	Depósito da Água Inadequado A00601 OU A009010	3 2 ou 6	Sim
	Instalação Sanitária Inadequada A01401 OU Conta 1	0 >3	Sim
	Destino do Lixo Inadequado A016010	3 ou 4 ou 5	Sim
	+ 3 Pessoas / Dorm. Conta 2	>3	Sim
	Falta de Eletrodomésticos Básicos A018013 OU (A018019 & A018017) OU A01901	2 2 2	Sim
	Regras		
	Qualquer Auxiliar	Sim	Habitação Não Saudável
	Todas Auxiliares	Não	Habitação Saudável
Plano Saúde	I00102 I001021	1 X	X
	I00102	2	0
Doenças Cr.	**Atributos da Tabela Descritiva**		
	Qualquer Atributo	1	Sim
	Todos os Atributos	2	Não
Sof. Mentais	**Atributos: [Q092, Q11007 até Q11010]**		
	Qualquer Atributo	1	Sim
	Todos os Atributos	2	Não

Tabela 3. Transformações - Aspectos Socioambientais

com a original, como pode ser visto na Tabela 4.

	Original	**Final**
Atributos	1090	19
Instâncias	293726	640

Tabela 4. Comparação Base Original x Final

3.3. Modelos de Classificação

Para avaliar a base de dados gerada, dois modelos de classificação foram utilizados: o *Explainable Boosting Machine* (EBM) criado pela Microsoft e o nó *Decision Tree Learner* da plataforma KNIME, ferramenta visual de análise de dados.

Na plataforma, foi criado um fluxo visual e realizada a cross-validação do modelo. Ao final, foram geradas as regras finais da árvore, selecionando somente aquelas que incluem pelo menos 10% da população da classe predita. Esta medida foi tomada para que sejam avaliadas somente as regras mais abrangentes e as características com maior poder de decisão dentre todas.

Os hiperparâmetros selecionados para a Árvore de Decisão são: Medida de qualidade Gini; poda simples; número mínimo de instâncias para dividir um nó interno da árvore = 10 e divisões binárias para valores nominais. O conjunto de dados completo foi

Base Comportamental			
	Precisão	**Recall**	**F1-score**
TOC = Sim	0.397	0.387	0.392
TOC = Não	0.613	0.624	0.619

Base Completa			
	Precisão	**Recall**	**F1-score**
TOC = Sim	0.614	0.567	0.59
TOC = Não	0.642	0.686	0.664

Tabela 5. Resultados de Teste: Árvore de Decisão

Base Comportamental			
	Precisão	**Recall**	**F1-score**
TOC = Sim	0.483	0.383	0.42
TOC = Não	0.604	0.686	0.639

Base Completa			
	Precisão	**Recall**	**F1-score**
TOC = Sim	0.534	0.549	0.533
TOC = Não	0.666	0.655	0.655

Tabela 6. Cross-Validação: Árvore de Decisão

dividido em 70% para treinamento e 30% para teste, sem amostragem estratificada em classe. Para treinamento foi aplicado o procedimento de validação cruzada com k-fold = 10. O processo de fine-tuning do algoritmo não foi realizado.

A seguir, os resultados da EBM foram utilizados para explicar a força de cada característica para a classificação final, assim como comparar o ranking de cada atributo com sua presença e importância nas regras da árvore de decisão. Os hiperparâmetros padrão do modelo foram utilizados, sem aplicação de fine-tuning, e os conjuntos de treino e teste, assim como na Árvore de Decisão, foram divididos pela proporção 70x30 não estratificada para as classes. O tempo de execução de ambos os modelos não é relevante, devido à pouca quantidade de instâncias na base de dados final.

Finalmente, os dois modelos foram executados utilizando uma variação da base de dados transformada que contém somente atributos comportamentais e a base completa, que contém atributos comportamentais e socioambientais. O objetivo da divisão é comparar o desempenho de ambos os modelos para as duas bases e entender se a adição das características socioambientais reforça ou enfraquece as métricas de comparação, a fim de determinar se questões socioambientais podem, também, influenciar na identificação do transtorno. O segundo objetivo é entender se as características comportamentais, quando utilizadas sem interferência das socioambientais, concordam com resultados encontrados na literatura e com as hipóteses discutidas com os especialistas da área de saúde mental entrevistados.

4. Resultados

Para a Árvore de Decisão, os resultados do treinamento da base comportamental indicaram que para um intervalo de confiança de 95%, a medida *F1-score* encontra-se no intervalo [0.48,0.52]. Já para a base completa, para o mesmo intervalo de confiança, a *F1-score* encontrou-se no intervalo [0.62, 0.63]. A Tabela 5 destaca os resultados para o conjunto de teste e os resultados com a cross-validação podem ser vistos na Tabela 6. As regras geradas pela Árvore para a base comportamental e completa, respectivamente, podem ser vistas nas Tabelas 7 e 8. É importante ressaltar que as regras consideradas para análise são somente aquelas que englobam pelo menos 10% da população da classificação predita e, dentre essas, explorou-se os resultados somente daquelas com maior taxa de acerto na predição para cada classe.

Para a EBM, o resultado do treinamento retornou Precisão média de 0.733, *Recall* médio de 0.162 e *F1-score* média de 0.265 para a base comportamental. Já para a base completa, o modelo obteve precisão média de 0.741, *Recall* médio de 0.421 e *F1-*

score média de 0.573. Neste caso não foi aplicada cross-validação, pois o objetivo não é analisar suas métricas, e sim utilizar sua interpretabilidade para entender a força de cada atributo para a classificação. O ranking de importância das variáveis preditoras para a base completa pode ser visto na Figura 1.

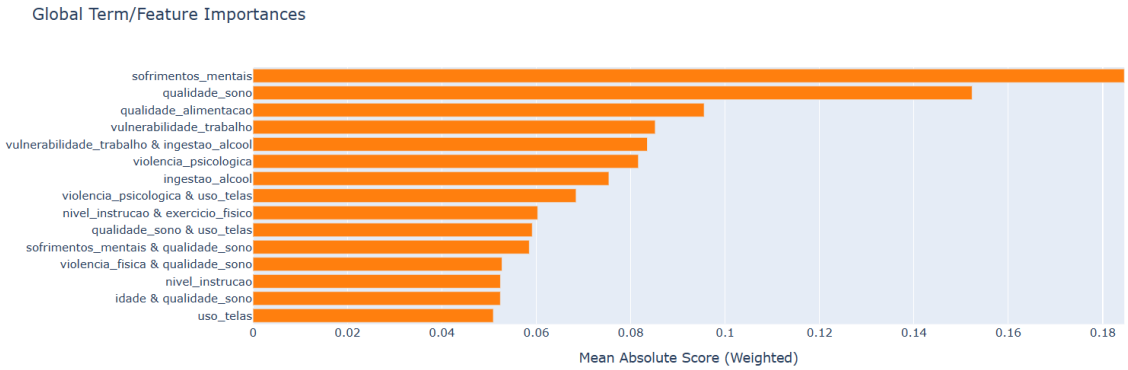

Figura 1. Importância Atributos - Base Completa

Base Comportamental			
Regras (TOC= Sim)			
Regra 1	**Valores**	**Total**	**Corretas**
Qualidade Sono	1. Ruim	59	41
Exercício Físico	1. Não Pratica Regularmente 2. <2 horas por semana 3. >= 6h por semana		
Qualidade Alimentação	1. Ruim, Muito Ruim		
Regra 2	**Valores**	**Total**	**Corretas**
Qualidade Sono	1. Ruim	30	17
Exercício Físico	1. Não Pratica Regularmente 2. <2 horas por semana 3. >= 6h por semana		
Qualidade Alimentação	1. Boa, Moderada		
Tabaco	1. Já foi fumante, mas não fuma atualmente 2. Fuma diariamente 3. Fuma ocasionalmente		

Base Comportamental			
Regras (TOC= Não)			
Regra 3	**Valores**	**Total**	**Corretas**
Qualidade Sono	1. Ruim	48	36
Qualidade Alimentação	1. Ruim, Boa, Moderada		
Uso Telas	1. Moderado 2. Alto		
Tabaco	1. Já foi fumante, mas não fuma atualmente 2. Fuma ocasionalmente		

Tabela 7. Regras Árvore de Decisão: Base Comportamental

Modelo	Precisão	Recall	F1-score
Base Comportamental	0.733	0.162	0.265
Base Completa	0.741	0.421	0.573

Tabela 9. Resultados do treinamento do modelo EBM

A partir dos resultados apresentados é possível perceber uma melhora significativa na *F1-score* de ambos os modelos com a adição da características socioambientais. Esta melhora pode ser vista principalmente para a classificação de TOC = Sim, onde a métrica subiu de 0.392 para 0.59 na Árvore de Decisão. Apesar de ainda não ser um resultado forte devido a vários desafios enfrentados, como o baixo número de instâncias para a Classe TOC = Sim, o que fez com que a representação da população tenha sido dificultada e um balanceamento cuidadoso da Classe TOC = Não tenha sido necessário, resultando na remoção de muitas linhas, já podemos afirmar que características socioambientais tem grande peso na identificação do TOC.

Base Completa			
Regras (TOC= Sim)			
Regra 1	**Valores**	**Total**	**Corretas**
Sofrimentos Mentais	1. Sim	30	20
Qualidade do Sono	1. Ruim		
Violência Psicológica	1. Amigo ou Vizinho 2. Relação Romântica 3. Trabalho 4. Outro		
Regra 2	**Valores**	**Total**	**Corretas**
Sofrimentos Mentais	1. Não	30	26
Estado Civil	1. Solteiro 2. Divorciado(a) ou desquitado(a) ou separado(a) judicialmente		
Regra 3	**Valores**	**Total**	**Corretas**
Sofrimentos Mentais	1. Sim	29	19
Qualidade do Sono	1. Ruim		
Violência Psicológica	1. Não 2. Família		
Vulnerabilidade Trabalho	1. Não		
Idade	1. Meia Idade 2. Adulto 3. Idoso		

Base Completa			
Regras (TOC= Não)			
Regra 4	**Valores**	**Total**	**Corretas**
Sofrimentos Mentais	1. Sim	75	71
Qualidade do Sono	1. Bom, Moderado		
Uso Telas	1. Moderado 2. Não Usa Telas 3. Baixo		
Idade	1. Meia Idade 2. Jovem Adulto 3. Adulto 4. Idoso		
Violência Psicológica	1. Não		
Regra 5	**Valores**	**Total**	**Corretas**
Sofrimentos Mentais	1. Sim	37	28
Qualidade do Sono	1. Ruim		
Violência Psicológica	1. Não 2. Família		
Vulnerabilidade Trabalho	1. Sim		
Cor / Raça	1. Branca 2. Preta 3. Amarela 4. Indígena		

Tabela 8. Regras Árvore de Decisão: Base Completa

Além disso, é possível perceber através da Figura 1 que apesar de atributos socioambientais como Sofrimentos Mentais e Vulnerabilidade no Trabalho terem alta força de decisão na classificação, características comportamentais como Qualidade do Sono e Qualidade da Alimentação ainda se apresentam nos rankings mais altos após a adição das características socioambientais.

Este padrão concorda com os resultados encontrados na literatura, como no estudo de [Segalàs et al. 2021], onde os autores encontraram correlação entre os sintomas de depressão e ansiedade em pessoas com TOC e baixa qualidade de sono. As regras da árvore de decisão da base completa também apontam para esta conclusão, o que pode ser visto na Tabela 8 onde a característica Qualidade do Sono = Ruim teve alto poder de divisão dos nós da árvore para a Classe TOC = Sim, assim como Qualidade do Sono = [Bom, Moderado] para a classe TOC = Não. Para qualidade de alimentação, o estudo de [Bang et al. 2020] afirma que um subconjunto considerável dos pacientes com TOC pode ter um transtorno alimentar clínico ou estar em alto risco de desenvolver um. Esta correlação pode ser vista na Tabela 7, que associa Qualidade Alimentação = [Ruim, Muito Ruim] para a classificação TOC = Sim.

Finalmente, ao analisar o resultado gerado pelo EBM para os atributos socioambientais na base completa, podemos indicar Sofrimentos Mentais e Vulnerabilidade no Trabalho como as mais fortes. É necessário ressaltar que Sofrimentos Mentais apresenta a presença de outros sofrimentos mentais além do TOC, ou seja, condições pré-existentes ou diagnosticadas após o diagnóstico de TOC. No caso de pessoas que não tem TOC, significa condições mentais como depressão, ansiedade, esquizofrenia, entre outros.

De acordo com as regras geradas pela árvore, Sofrimentos Mentais = Não tem alta correlação com a presença de TOC, enquanto Sofrimentos Mentais = Sim tem correlação

com TOC = Não. Esta regra concorda com o diagnóstico clínico do DSM-5, que indica que os sintomas apresentados pelo paciente para diagnóstico de TOC não podem ser melhor explicados por critérios diagnósticos de outro transtorno mental. Isso significa que, caso a pessoa seja diagnosticada com TOC, os sintomas de ansiedade e depressão, por exemplo, apresentados pelo paciente são indicados como sintomas do TOC, e não como diagnósticos de Ansiedade ou Depressão. Portanto, pessoas diagnosticadas com TOC tem pouca probabilidade de também terem outros diagnósticos.

Em relação ao atributo Vulnerabilidade no Trabalho, encontrou-se resultados contrários aos discutidos com os especialistas da área da saúde. Durante as entrevistas, foi relatado que vulnerabilidades como estresse e ambiente desagradáveis no trabalho teriam alta relevância na apresentação dos sintomas do TOC, porém as regras encontradas pela Árvore de Decisão demonstraram o contrário, ou seja, este aspecto quando apresentado está correlacionado à ausência de TOC no paciente.

5. Conclusão e Trabalhos Futuros

A análise dos modelos preditivos baseados em variáveis comportamentais e socioambientais revelou insights valiosos sobre a classificação do TOC. A inclusão de variáveis socioambientais aumentou significativamente o *F1-score* da classe TOC = Sim, embora o impacto na classe TOC = Não tenha sido menos pronunciado.

As variáveis comportamentais, como qualidade do sono e qualidade da alimentação, são as mais informativas, conforme destacado tanto pelo modelo EBM quanto pelas regras da árvore em relação à literatura existente. Em particular, a qualidade do sono ruim foi uma constante significativa nas regras das árvores de decisão, o que reforça sua importância como fator determinante na análise comportamental. Ao adicionar variáveis socioambientais, observamos um aumento na proporção de regras relacionadas a essas características, especialmente a ausência de outros sofrimentos mentais, que é altamente correlacionada ao TOC. Isso está alinhado com o DSM-5, que indica que pessoas com TOC dificilmente têm diagnósticos prévios de outros transtornos devido à sobreposição de sintomas.

Em síntese, embora a base de dados não apresente informações tão precisas devido a vários desafios, como o baixo número de casos de TOC, a necessidade de excluir entrevistados menores de idade e a remoção de atributos balanceados demais entre as classes, demonstrou-se que as características socioambientais são valiosas para a classificação do TOC, especialmente quando associadas a características comportamentais. Portanto, os resultados reforçam a utilidade de incluir uma variedade de fatores para melhorar a precisão e a profundidade das análises preditivas sobre TOC.

É importante ressaltar uma limitação significativa deste trabalho: a ausência de um estudo longitudinal das pessoas participantes da PNS 2019. É possível que o diagnóstico de TOC tenha ocorrido muitos anos antes das entrevistas, o que impede a afirmação de que as características socioambientais apresentadas pela pesquisa são as mesmas que existiam no momento do diagnóstico. Portanto, futuros trabalhos devem incluir um estudo longitudinal para acompanhar as mudanças nas características socioambientais ao longo do tempo e sua relação com o diagnóstico de TOC.

Além disso, deseja-se replicar o processo com uma base de dados maior, ou seja, com maior quantidade de instâncias com classificação de TOC, assim como modificar

as transformações de dados das características socioambientais, em busca de um modelo mais forte. Será executado, também, o processo de fine-tuning dos algoritmos, passo que não foi realizado neste trabalho. Deseja-se também buscar entender o motivo da relação inversa da presença de Vulnerabilidade no Trabalho com o diagnóstico, a partir da fusão de novos atributos ou exploração de características correlacionadas a este aspecto.

Referências

Association, A. P. (2013). *Manual Diagnóstico e Estatístico de Transtornos Mentais*. Artmed, 5 edition.

Azeredo, C. M., Cotta, R. M. M., Schott, M., Maia, T. d. M., and Marques, E. S. (2007). Avaliação das condições de habitação e saneamento: a importância da visita domiciliar no contexto do programa de saúde da família. *Ciência Saúde Coletiva*, 12(3):743–753.

Bang, L., Kristensen, U. B., Wisting, L., Stedal, K., Garte, M., Minde, , and Rø, (2020). Presence of eating disorder symptoms in patients with obsessive-compulsive disorder. *BMC Psychiatry*, 20(1):36.

Brander, G., Pérez-Vigil, A., Larsson, H., and Mataix-Cols, D. (2016). Systematic review of environmental risk factors for obsessive-compulsive disorder: A proposed roadmap from association to causation. *Neuroscience Biobehavioral Reviews*, 65:36–62. Epub 2016 Mar 21.

Buss, P. M. and Pellegrini Filho, A. (2007). A saúde e seus determinantes sociais. *Physis: Revista de Saúde Coletiva*, 17(1):77–93.

Cazzolato, M., Rodrigues, L., Ribeiro, M., Gutierrez, M., Jr., C. T., and Traina, A. (2021). Similarity search and correlation-based exploratory analysis in ehrs: A case study with covid-19 databases. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 25–36, Porto Alegre, RS, Brasil. SBC.

Hu, X., Liu, Q., Li, B., Tang, W., Sun, H., Li, F., Yang, Y., Gong, Q., and Huang, X. (2016). Multivariate pattern analysis of obsessive–compulsive disorder using structural neuroanatomy. *European Neuropsychopharmacology*, 26(2):246–254.

Højgaard, D. R., Arildskov, T. W., Skarphedinsson, G., and et al. (2023). Do autistic traits predict outcome of cognitive behavioral therapy in pediatric obsessive-compulsive disorder? *Research on Child and Adolescent Psychopathology*, 51:1083–1095.

Instituto Brasileiro de Geografia e Estatística (IBGE) (2020). Pesquisa nacional de saúde 2019. `https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=25921&t=resultados`. Acesso em: 2024-07-15.

KNIME (2023). Decision tree learner (3.6.0). `https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.mine.decisiontree2.learner2.DecisionTreeLearnerNodeFactory3/`. Accessed: 2024-07-12.

Montevecchi, A. and Zárate, L. E. (2014). *Pictorea: Um método para descoberta de conhecimento em bancos de dados convencionais*. Novas Edições Acadêmicas. Published by Lucky's Textbooks, Dallas, TX, U.S.A.

Pasternak, S. (2016). Habitação e saúde. *Estudos Avançados*, 30(86):51–66.

Sabrina Boger, Thomas Ehring, G. B. and Werner, G. G. (2020). Impact of childhood maltreatment on obsessive-compulsive disorder symptom severity and treatment outcome. *European Journal of Psychotraumatology*, 11(1):1753942.

Segalàs, C., Labad, J., Salvat-Pujol, N., Real, E., Alonso, P., Bertolín, S., Jiménez-Murcia, S., Soriano-Mas, C., Monasterio, C., Menchón, J. M., and Soria, V. (2021). Sleep disturbances in obsessive-compulsive disorder: influence of depression symptoms and trait anxiety. *BMC Psychiatry*, 21(1):42.

Souza, V., Nobre, J., and Becker, K. (2020). Characterization of anxiety, depression, and their comorbidity from texts of social networks. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 121–132, Porto Alegre, RS, Brasil. SBC.

The InterpretML Contributors (2023). Explainable boosting machine. Accessed: 2024-05-23.

Zarate, L., Petrocchi, B., Maia, C. D., Felix, C., and Gomes, M. P. (2023). Capto - a method for understanding problem domains for data science projects. *Concilium*.