Relações causais no contexto da saúde mental

Cristiane Freitas [1] **Felipe Cadar** [1] **Gabriel Diniz** [1] **Thamara Dias** [1]

Abstract

O aumento alarmante do número de pessoas afetadas por doenças mentais tornou-se um dos principais problemas de saúde pública enfrentado pelos governos em todo o mundo. A preocupação com esse cenário foi intensificada com o surgimento da nova síndrome respiratória aguda (SARS-CoV-2), parcialmente devido às medidas de isolamento social, que são necessárias para evitar a propagação do vírus (COVID-19), mas que podem agravar o problema de saúde mental da população. Neste projeto, estamos interessados em utilizar dados de classificação clínica para identificar os possíveis fatores causais que possam estabelecer relações com a ideação suicida. Para tanto, utilizaremos métodos de inferência causal e aprendizado de máquina para analisar comportamentos, identificar as possíveis causas e classificar um indivíduo quanto a ideação suicida.

1. Introdução

O suicídio está entre as 20 causas de morte mais comuns segundo avaliação da Organização Mundial da Saúde (Organization et al., 2019), superando malária, câncer de mama ou homicídio. Nos últimos 20 anos, houve um aumento de 24% dos casos (Sullivan et al., 2013), o que piorou com a pandemia da COVID-19. Apesar da situação alarmante, observou-se pouco progresso nos últimos 50 anos na compreensão do suicídio e na melhoria dos tratamentos de indivíduos em risco (Franklin et al., 2017).

Diferentes métodos de aprendizado de máquina já foram empregados para lidar com o desafio de detectar transtornos mentais, como feito por (Orabi et al., 2018). Apesar dos potenciais oferecidos por esses modelos, existem limites fundamentais para o raciocínio baseado apenas na previsão. Por exemplo, o que acontecerá se uma pessoa com ideação suicida deixar de ingerir bebida alcoólica? Tais questões não podem ser respondidas por um modelo correlativo construído sob dados previamente observados, pois envolvem uma possível mudança do usuário com relação a sua conduta. Como demonstrado em (Cloudera, 2020), sistemas preditivos puramente correlativos não estão equipados para raciocinar sob tais intervenções e, portanto, para a tomada de decisão informada por dados sob intervenção, precisamos de causalidade.

O presente trabalho visa identificar a melhor combinação de técnicas para construção de um modelo de classificação causal, objetivando classificar o risco de ideação suicida dos pacientes ambulatoriais diagnosticados com depressão.

2. Referencial Teórico

O suicídio é um importante problema de saúde pública, com impactos na sociedade como um todo. De acordo com a (Organization et al., 2019) ocorre um suicídio a cada 40 segundos no mundo e mais de 700 mil pessoas morrem por suicídio anualmente.

O notável progresso no campo preditivo de ideações suicidas alcançado nos últimos anos foi derivado do uso de arquiteturas neurais para extrair padrões de dados em grande escala (Lee & Toutanova, 2018; Liu et al., 2019). A questão é que essas arquiteturas não fazem distinção entre causas, efeitos e fatores de confusão: uma *feature* pode ser um poderoso preditor, mesmo que não tenha relação causal direta com a classe esperada (Feder et al., 2021). O conhecimento da relação causal entre os dados observados e *features* pode ser usado para formalizar correlações espúrias (Bühlmann, 2020; Veitch et al., 2021). Por essas razões, uma linha crescente de pesquisa tem tentado reorientar o aprendizado de máquina em torno de fundamentos causais (Schölkopf, 2022; Schölkopf et al., 2021).

Na saúde mental, focar nos indivíduos é fundamental, pois experiências passadas podem mudar a forma como as pessoas veem ou lidam com as situações. Pensando nesse aspecto, (Marchezini et al., 2022) propuseram um novo quadro para tratar do problema de modelagem do raciocínio contrafactual (situação ou evento que não aconteceu, mas poderia ter acontecido) em cenários onde, além das variáveis endógenas observadas, verifica se existe uma variável latente que afeta os resultados. Foram realizados experimentos com conjuntos de dados sintéticos e do mundo real,

[*]Equal contribution [1]Departamento de Ciência da Computação UFMG, Brasil. Correspondence to: <>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

tentando predizer como as mudanças nas ações das pessoas podem levar a resultados diferentes em termos de sintomas de doença mental e qualidade de vida. Os resultados mostraram que o modelo tem bons resultados com casos não lineares e multiplicativos que geralmente são características de modelos do mundo real; e responde a perguntas contrafactuais que são suportadas pela literatura médica, tendo o potencial de recomendar pequenas mudanças na vida das pessoas que podem alterar completamente sua relação com a doença mental.

Em se tratando do campo da saúde de forma mais ampla do que saúde mental, estudos que combinam técnicas causais com aprendizado de máquina tem sido conduzidos, como em (Veloso & Ziviani, 2020). Neste artigo, os autores utilizaram dados relacionados à pandemia do COVID-19 em diversos países para predizer a taxa de mortes por dia no país e como esta taxa evolui ao longo do tempo. Os métodos causais foram aplicados para garantir explicabilidade do modelo e para gerar raciocínios contrafactuais e predizer como será a resposta em termos de taxa de mortes por dia caso uma determinada política seja aplicada no país.

Apesar das diversas pesquisas e trabalhos citados, há uma carência de estudos que combinem extração de causalidade com a utilização de dados clínicos de pacientes ambulatoriais. Por esse motivo pretendemos utilizar os dados do ensaio clínico **STAR*D** de (Rush et al., 2004), visando identificar as causas e classificar um indivíduo quanto a ideação suicida.

STAR*D é um ensaio clínico, randomizado e multicamadas aplicado em pacientes ambulatoriais com transtorno depressivo. O estudo compara várias opções de tratamento para aqueles que não alcançaram uma resposta satisfatória com o medicamento Citalopram, um inibidor seletivo de recaptação de serotonina. Foram abordados 4.000 adultos (idades entre 18 e 75 anos) sem características psicóticas. O resultado primário de avaliação clínica utiliza 17 itens da escala de classificação Hamiltoniana de depressão colhidos no início e fim de cada nível de tratamento através de entrevistas telefônicas. Os resultados secundários incluem sintomas depressivos autorrelatados, sintomas físicos e mentais, efeito colateral, satisfação do paciente e custo com cuidados com a saúde. Os participantes com resposta sintomática adequada podem entrar na fase de acompanhamento de 12 meses com breves avaliações trimestrais e avaliações mensais mais completas.

3. Metodologia

Neste projeto utilizamos técnicas de Aprendizado Supervisionado Causal para identificar os padrões existentes na predição de ideação suicida, e, Explicabilidade para entender o impacto da causalidade nos resultados.

3.1. Base de Dados

Utilizamos dados de classificação clínica; medidas comportamentais, dimensionais e qualidade de vida; testes neurocognitivos do conjunto de dados **STAR*D** de (Rush et al., 2004).

3.2. Aprendizado Supervisionado Causal

Um dos maiores problemas em aprendizado de máquina ocorre quando os dados do mundo real possuem distribuições diferentes daqueles observados nos dados de treino e teste. Dessa forma, correlações espúrias são incorporadas à modelagem e os resultados são prejudicados.

Para trabalhar essa questão utilizamos grafos causais para isolar as correlações internas não-causais entre as *features*. Com o grafo causal das *features* da base de dados, mantivemos os nós com influência causal direta ao nosso objetivo de classificação e os nós que possuem influência causal a esse conjunto anterior (nós com até dois níveis de distância do nosso objetivo de classificação).

Após a construção do *dataset* com os atributos selecionados por meio de técnicas causais, a base foi utilizada para treinar modelos de classificação.

3.3. Explicabilidade

Utilizamos esse método para obter a quantificação da influência causal de cada *feature* no modelo.

Para isso utilizaremos o método SHAP, que efetua uma quantificação da influência de cada *feature* de entrada utilizando função de custo e estratégias de amostragem.

Por meio da explicabilidade é possível contrastar o que de fato está sendo utilizado pelo modelo de classificação para efetuar as predições com o que era o esperado de acordo com o grafo e com o conhecimento causal.

3.4. Contrafactual

Usando como base a tese de mestrado do aluno Guilherme F. Marchezini, chamada de ”Counterfactual Inference with Latent Variable and its Application in Mental Health Care”, reproduzimos a metodologia de predição contrafactual. A ideia principal do trabalho é como isolar fatores endógenos, chamados de individualidade, que descrevem um individuo para que, na hora de fazer modificações ao cenário desse indivíduo, o calculo do cenário contrafactual possa levar em conta essa individualidade. A rede possui duas etapas, a primeira é estimar um valor de individualidade, em seguida tratamos esse valor como uma nova *feature* e usamos uma segunda rede para reconstruir a variável preditora. Após o treino das redes podemos estimar um cenário contrafactual de um indivíduo especifico usando a individualidade

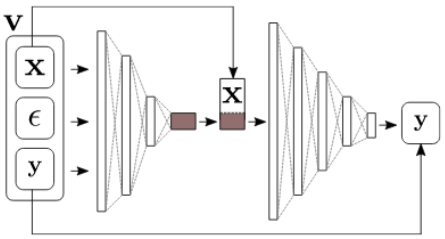


Figure 1. Arquitetura de predição contrafactual.

original e as *features* modificadas.

Para a adaptação da arquitetura aos nosso problema, escolhemos como variáveis de interesse a ideação de suicídio e a ansiedade. Então conseguimos selecionar indivíduos específicos e modificar uma característica específica para estimar como seria o relacionamento desse indivíduo com as variáveis de interesse.

Arquitetura. A arquitetura da rede de predição contrafactual pede alguns outros componentes presentes neste artigo. Para levar em consideração a estrutura causal do problema, usamos o grafo gerado na sessão anterior para selecionar apenas as *features* com relação causal direta como nossas variáveis de interesse. Também usamos os conhecimentos adquiridos na sessão de classificação para validar a eficacia da previsão usando um modelo estrutural causal. A arquitetura final é composta de duas etapas: 1) Estimar o fator de individualidade do paciente e 2) usar a individualidade e o exemplo contrafactual para prever as variáveis de interesse. Podemos ver esses componentes na figura 1. Esse modelo pode ser treinado usado os exemplos reais (não contrafactuais) para aprender a estimar corretamente a individualidade de cada paciente.

Detalhes de implementação. Para o treino, primeiramente criamos duas redes, uma para cada variável de interesse. Treinamos essas redes e salvamos os erros de regressão de cada indivíduo. Adicionamos os erros como novas *features* e usamos esse novo dado para treinar a rede contrafactual. A cada rede é uma simples MLP de 5 camadas. Todas foram treinadas com o otimizador Adam com *learing rate* 0.001 e um decaimento de 0.00001.

4. Resultados

4.1. Construção do *Dataset*

Como mencionado anteriormente, utilizamos o ensaio clínico **STAR*D** para construir o *dataset*, que compreende variáveis divididas em características pessoais (sexo, idade, renda mensal, etc), emocionais (ansiedade, sentimentos de culpa, etc), familiares (número total de pessoas que residem com o paciente, impacto da família e amigos, etc) e físicas (renal, fígado, neuro, etc).

Algumas variáveis são definidas com base em uma escala de pontuação, como por exemplo, a variável 'Suicídio' que é referente à propensão do indivíduo quanto à ideação suicida. Um valor 0 é atribuído à ausência desse pensamento, 1 quando a pessoa sente que a vida está vazia, 2 para aqueles que apresentam desejos de morte, 3 para quem tem pensamentos suicidas ativos e 4 para tentativa grave de suicídio.

A seguir, algumas análises gráficas de dados foram realizadas, considerando 'Suicício' como *target*.

Os gráficos da Figura 2 mostram como algumas variáveis se comportam em relação ao suicídio e podemos verificar que em ambos os casos há uma relação positiva; ou seja, quanto mais o indivíduo se sente ansioso, deprimido e apresenta sentimentos de culpa; mais pensamentos suicidas ele tem. Isso também pode ser observado com a variável 'Trabalho e interesses'; ou seja, quanto maior o sentimento de incapacidade, improdutividade e perda de interesse em atividades laborais e de lazer, maior é a ideação suicida.

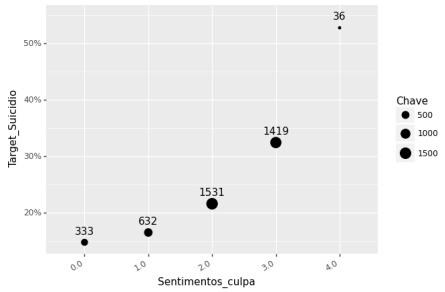
A variável 'Pontuação total' é uma escala hamiltoniana de classificação para depressão-clínico, em que um score com mais de 25 pontos caracteriza pacientes gravemente deprimidos, entre 18 a 24 pontos caracteriza pacientes moderadamente deprimidos e escores entre 7 e 17 pontos caracterizam pacientes com depressão leve. No gráfico 4(d), podemos visualizar aproximadamente uma relação linear entre essa variável e o *target*, o que é intuitivo e fácil de compreender, uma vez que quanto mais o indivíduo se sente deprimido, maior será a ideação suicida, confirmando a análise da 2(b).

Já em relação ao volume de pacientes, podemos visualizar que há um maior número de indivíduos caracterizados pelas escalas 2 (Leve - ideias de culpa ou ruminação sobre erros passados ou sentimentos pecaminosos) e 3 (Moderada - a doença atual é uma punição - ilusões de culpa); e menos indivíduos na escala 4 (Grave - ouve vozes acusatórias ou denunciatórias e/ou experiências que ameaçam alucinações visuais).

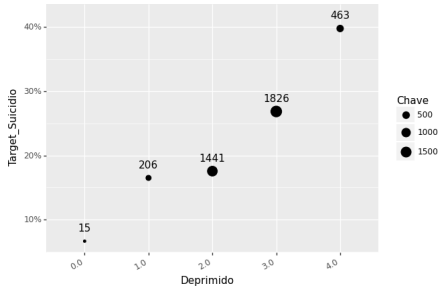
Para 'Deprimido', o maior volume também se concentra nas escalas 2 (Leve - sente-se triste, sem esperança, indefeso, inútil não mais do que 2 dias) e 3 (Moderado). Porém o menor volume está na escala 0 (humor deprimido ausente); o que de certa forma seria esperado visto que acreditamos que exista alguma relação entre as variáveis 'Deprimido' e 'Suicídio'.

De maneira análoga, os demais gráficos podem ser analisados.

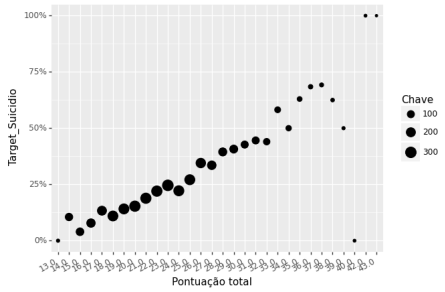
No que diz respeito à variável 'sexo', considerando 1 como 'Feminino' e 2 como 'Masculino', podemos ver na Figura 3 que mulheres apresentam menos propensão a pensamentos suicidas do que homens.



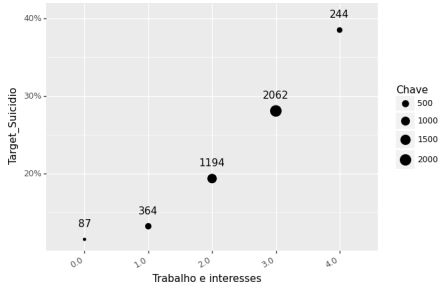
(a) Sentimentos de culpa



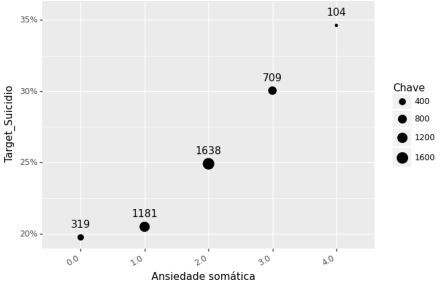
(b) Deprimido



(c) Pontuação total



(d) Trabalho e interesses



(e) Ansiedade somática

Figure 2. Target: Suicídio

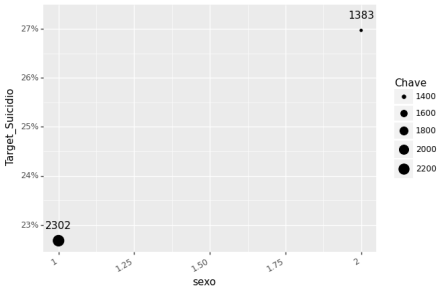
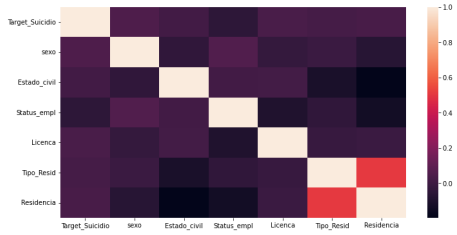
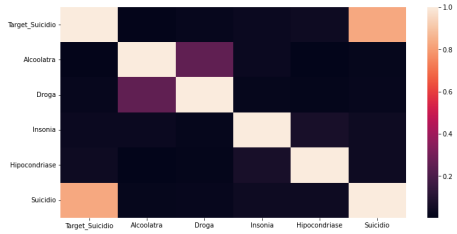


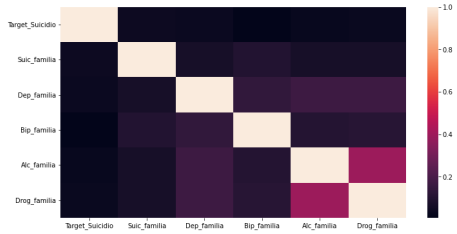
Figure 3. Sexo



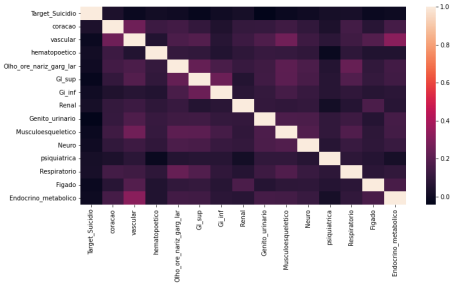
(a) Variáveis Pessoais



(b) Variáveis emocionais

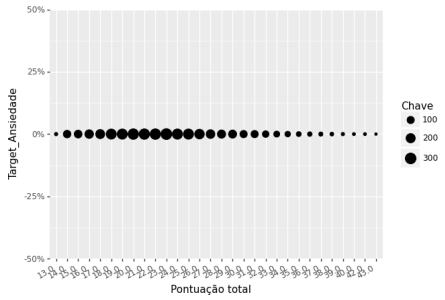


(c) Variáveis familiares

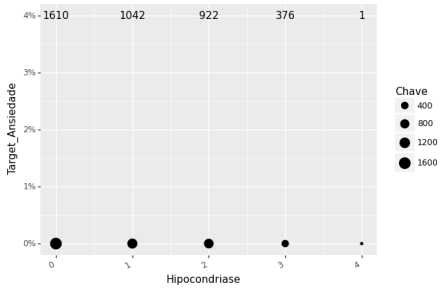


(d) Variáveis Físicas

Figure 4. Heatmaps



(a) Pontuação Total



(b) Hipocondríase

Figure 5. Heatmaps

Ainda com o *target* 'Suicidio', foram gerados alguns heatmaps para mostrar as correlações entre diferentes variáveis do mesmo grupo. Na Figura 4, é possível verificar em todos os grupos que a maior parte das variáveis não apresentam correlações fortes entre si.

As mesmas análises anteriores, podem ser feitas alterando o *target*, por exemplo para 'Ansiedade'. Contudo, para 'Pontuação total' e 'Hipocondríase' observamos que 'Ansiedade' fica no nível 0 para ambas as variáveis, como demonstrado na Figura 5.

4.2. IRT

Para a construção do grafo de causalidade, foi necessário utilizar como *target* uma variável contínua. Porém, as variáveis possíveis de se tratar como *target* para os fins do presente trabalho, sendo elas o nível de ideação suicida ou o nível de ansiedade, são variáveis categóricas.

A transformação das variáveis categóricas em variáveis contínuas foi feita utilizando um método de *Item Response Theory* (IRT), como proposto pelos autores (Kennedy et al., 2020). Neste artigo, é aplicado um modelo de IRT da família Rasch, também chamados de IRT 1-PL, sobre um conjunto de features categóricas para a obtenção de uma única variável contínua.

A implementação do modelo Rasch IRT foi feita utilizando a

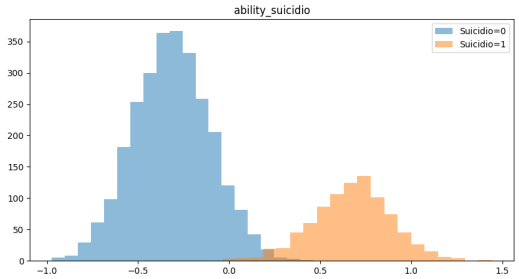


Figure 6. Distribuição dos resultados após aplicar IRT sobre a variável de suicídio

Table 1. Resultados do treinamento do Modelo IRT 1-PL para o nível de ideação suicida

Epoch	Loss	Best Loss
1	1780217.1250	1780217.1250
101	46220.5439	2680.0400
150	32870.9593	2311.0153

biblioteca *Python* py-irt[1], que utiliza a metodologia proposta em (Natesan et al., 2016).

A partir da biblioteca utilizada, foram feitas duas transformações: uma aplicando o modelo de IRT somente sobre a feature de nível de ideação suicida e outra somente sobre o nível de ansiedade.

Ambas as transformações foram feitas utilizando 150 épocas e uma taxa de aprendizagem de 0,1.

No dataset inicial, ambas as features utilizadas são categóricas e podem assumir valores inteiros de 0 a 4. Para que fosse possível a utilização do modelo de IRT, as features foram convertidas em features binárias. Assim, os valores 0 e 1 foram convertidos para 0 e os valores 2, 3 ou 4 foram convertidos para 1.

A distribuição dos resultados da transformação da feature de ideação suicida pode ser vista na figura 6.

Apesar de o nível de ideação suicida e o nível de ansiedade serem variáveis categóricas, é possível afirmar que a intensidade dessas features cresce à medida que o número que as representa cresce. Isso acontece pela própria definição dos níveis citadas no item 4.1 deste trabalho. Sendo assim, é esperado que ao aplicar a transformação por meio de IRT nessas variáveis, ainda seja possível observar uma relação crescente entre os níveis. Essa premissa foi confirmada com os dados observados no gráficos da figura 6. A mesma relação é observada para a transformação da feature referente à ansiedade.

As tabelas 1 e 2 apresentam os valores de Loss obtidos du-

[1]Disponível em: https://github.com/nd-ball/py-irt

Table 2. Resultados do treinamento do Modelo IRT 1-PL para o nível de ansiedade

Epoch	Loss	Best Loss
1	544541.3125	544541.3125
101	37486.0689	7954.9687
150	4330.5803	2424.0341

rante o treinamento do modelo IRT 1-PL. Por meio dessas tabelas é possível verificar que a melhor Loss obtida no treinamento do modelo da feature do nível de ideação suicida é muito próxima a obtida para modelo do nível de ansiedade.

4.3. Modelo causal gráfico

Apesar de observarmos relações positivas entre os *targets* 'Suicídio' e 'Ansiedade' e algumas variáveis preditoras como 'Sentimentos de Culpa', 'Trabalho e interesse' e 'Ansiedade somática' usando apenas análise de correlação, não podemos afirmar que tais *targets* seriam consequências dessas variáveis.

A capacidade de identificar relações verdadeiramente causais é fundamental para efetuar intervenções que possam impactar positivamente a vida das pessoas, portanto é crucial distinguir entre eventos que causam resultados específicos e aqueles que meramente se correlacionam. Uma possível explicação para a correlação entre variáveis em que nenhuma causa a outra é a presença de variáveis de confusão que influenciam tanto o *target* quanto um direcionador desse *target*. Variáveis de confusão não observadas são ameaças graves ao fazer inferência causal em dados observacionais.

Determinar a causalidade entre variáveis pode ser um passo desafiador. Para tanto iniciamos o processo gerando o grafo acíclico dirigido (*DAG*) objetivando efetuar uma seleção de *features* identificadas como possíveis causas do nosso *target*. Focamos nas Redes Bayesianas que é um tipo de model causal estrutural. Modelos causais estruturais representam dependências causais usando grafos que fornecem uma visualização intuitiva ao representar variáveis como nós e relacionamentos entre variáveis como arestas em um grafo.

Um grafo é uma coleção de nós e arestas, onde os nós são alguns objetos, e as arestas entre eles representam alguma conexão entre esses objetos. Um grafo direcionado é um grafo no qual cada aresta é orientada de um nó para outro nó. Em um grafo direcionado, uma aresta vai de um nó pai para um nó filho. Um caminho em um grafo direcionado é uma sequência de arestas tal que o nó final de cada aresta é o nó inicial da próxima aresta na sequência. Um ciclo é um caminho no qual o nó inicial de sua primeira aresta é igual ao nó final de sua última aresta. Um grafo acíclico direcionado é um grafo direcionado que não possui ciclos.

A Rede Bayesiana consiste em um *DAG*, um grafo causal onde os nós representam as variáveis aleatórias e as arestas representam a relação entre elas, e uma distribuição de probabilidade condicional associada a cada uma das variáveis aleatórias.

Para obter as *features* identificadas como possíveis causas dos nossos *targets* seguimos as seguintes etapas:

1. Geração do grafo causal 7 e 9 para os *targets* 'Suicídio' e 'Ansiedade' respectivamente, objetivando capturar a individualidade. Geração do gráfico de barras das features com estimativa de probabilidades relacionadas ao *targets* 'Suicídio' 11.
2. Remoção de arestas com baixa estimativa de probabilidade (peso menor que 0,5).
3. Remoção das arestas que não fazem parte do *Markov Blanket* (incluindo os pais do nosso *target*, filhos e os pais de todos os seus filhos) 8 e 10.
4. Revisão da estrutura. Cada relação deve ser validada, para que possa ser afirmada como causal. Isso pode envolver inverter, remover ou adicionar arestas aprendidas ou confirmar o conhecimento especializado com especialistas da área.

Para nos suportar utilizamos a biblioteca *Python* (Beaumont et al., 2021) que permite desenvolver modelos que vão além da correlação, considerando relações causais.

Inicialmente estamos interessados em obter o grafo acíclico dirigido ótimo (*DAGs*) que descreve as dependências condicionais entre as variáveis. No entanto, o espaço de busca para isso é combinatório e cresce exponencialmente com o número de nós. Utilizamos o algoritmo *notears* dessa biblioteca que introduz uma nova heurística de otimização e abordagem para resolver esse problema, onde o tempo de execução para isso é cúbico pelo número de nós e não mais exponencial.

Notears funciona detectando se um pequeno aumento no valor do nó resultará em um aumento em outro nó. Se houver, ela será capaz de capturar isso e afirmar que esta é uma relação causal. Portanto, é altamente recomendável que o conjunto de dados a ser usado seja contínuo. Por esse motivo, normalizamos nossas *features* para dados contínuos.

É recomendado que pelo menos 1000 amostras sejam usadas para obter um desempenho satisfatório e como nosso *Dataset* tem mais de 3600 amostras, foi possível utilizar o algoritmo.

Para imprimir o grafo utilizamos o algoritmo *plot structure* da mesma biblioteca, removendo as arestas cujos pesos absolutos eram menores que 0.5.

Figure 7. Grafo Causal - Suicídio

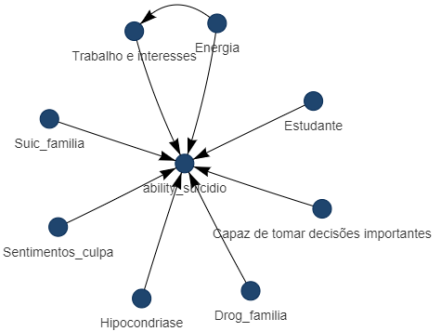


Figure 8. Grafo Causal - Suicídio (Markov Blanket)

Para a geração do grafo foram utilizadas as seguintes *features*: Sexo, Estado Civil, Tipo de Residência, Alcoólatra, Droga, Suicídio na Família, Depressão na Família, Alcoólatra Família, Droga na Família, Neuro, Psiquiátrica, Capaz de desfrutar das coisas, Impacto de sua Família e amigos, Capaz de tomar decisões importantes, Estudante, Insônia, Ansiedade, Perda de *insights*, Apetite, Perda de Peso, Ansiedade somática, Hipocondriase, Sentimento de Culpa, Trabalho e interesses, Energia, Lentidão pensamento e fala, Agitação, Libido, TOC, idade e Suicídio após aplicação descrita no item 4.2.

O grafo 8 foi validado pelo psicólogo André Santos Araújo (CRP08/23885) que trata principalmente pacientes com transtornos psicológicos bem como pacientes com ideação suicida.

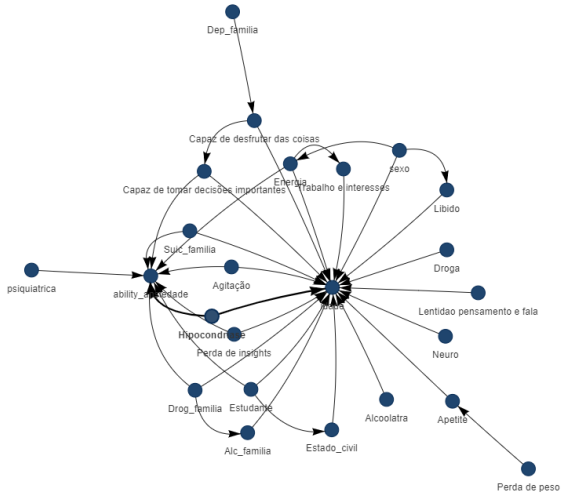


Figure 9. Grafo Causal - Ansiedade

Figure 10. Grafo Causal - Ansiedade (Markov Blanket)

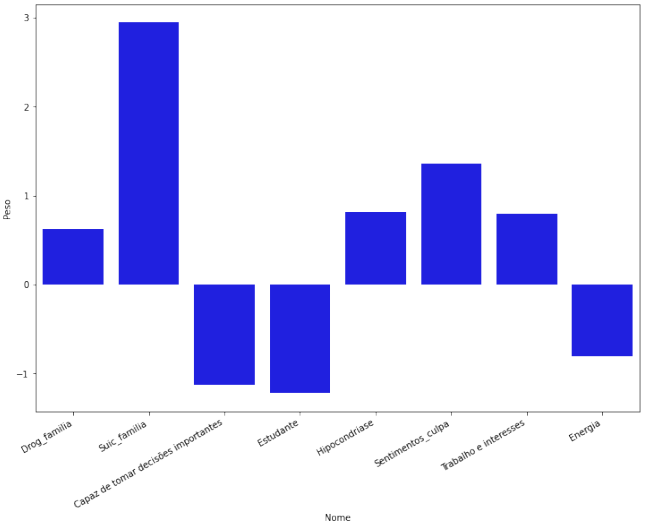


Figure 11. Features com estimativa de probabilidades

4.4. Baseline de Classificação

Para avaliar o impacto da causalidade em modelos de aprendizagem de máquina, utilizamos como *benchmark* um modelo de classificação para predição quanto à ideação suicida, considerando a variável 'Suicídio' como resposta e um total de 32 *features*. A aplicação do modelo compreende as seguintes especificações:

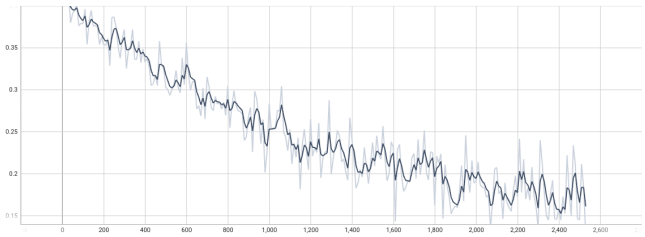
- Biblioteca: PyTorch Lightning
- Algoritmo: Rede neural
- *Layers*: 2
- Neurônios: 128
- Função de ativação: LeakyReLU
- Função perda: binary_cross_entropy
- Otimizador: ADAM
- Parada antecipada: *loss* nos dados de validação
- Taxa de aprendizagem: 0.001

O *dataset* foi dividido em dados de treino (80%) e dados de validação (20%) e posteriormente com base nas especificações acima o modelo de classificação foi empregado. As Figuras 14 e 13 apresentam o comportamento da *loss* e da acurácia do modelo durante sua execução para os dados de treinamento e validação. Como podemos verificar, a perda (*loss*) diminui nos dados de treinamento e aumenta no conjunto de validação à medida que o modelo evolui. Já em relação à acurácia, a precisão aumenta no conjunto de treinamento e diminui no conjunto de validação. Portanto, o modelo está superajustando os dados, ou seja, temos um caso de *overfitting* que pode ser explicado pela existência de correlações espúrias entre as *features*.

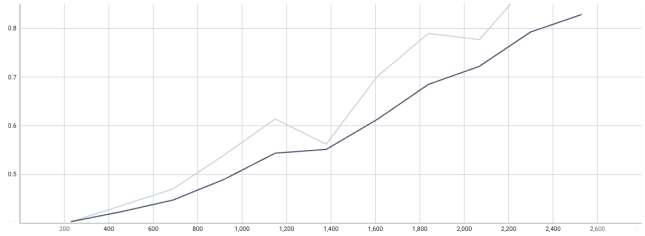
O mesmo modelo de classificação com o mesmo *target* 'Suicidio' foi aplicado ao conjunto de 8 *features* que foram propostas pelo modelo gráfico causal da Figura 8. Os resultados são mostrados na Figura 15 na cor azul em conjunto com os resultados anteriores obtidos com as 32 *features*. Uma vez que a causalidade é incorporada à seleção de *features*, resultados melhores podem ser identificados, principalmente no que concerne ao conjunto de validação. Logo, através de uma simples comparação é possível averiguar os benefícios da aplicação da causalidade em estudos de aprendizagem de máquina.

4.5. Contrafactual

Ao final do treino contrafactual podemos avaliar a acurácia da rede nos dados de teste sem alterações. Para ter um comparativo inicial, também executamos todo o *pipeline* sem

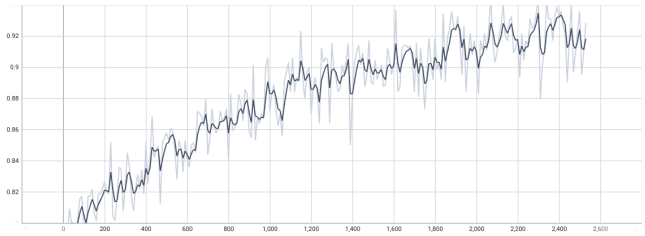


(a) *Loss* - Treino

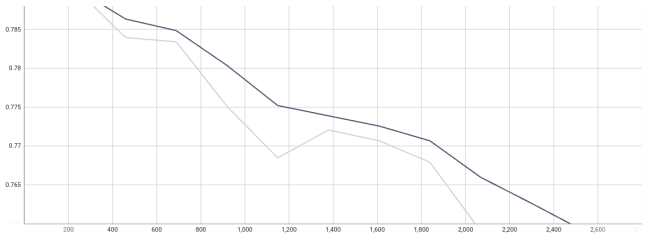


(b) *Loss* - Validação

Figure 12. Evolução *loss* - Rede neural com 32 *features*

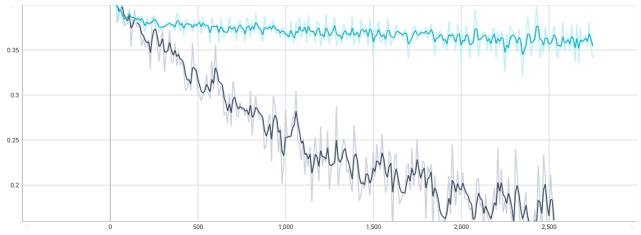


(a) Acurácia - Treino

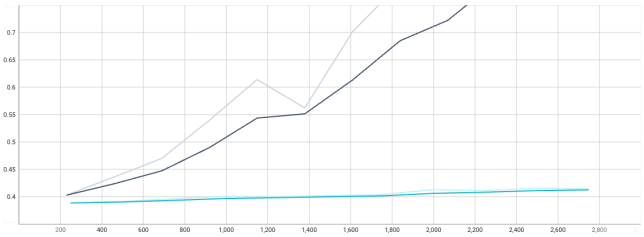


(b) Acurácia - Validação

Figure 13. Evolução da Acurácia - Rede neural com 32 *features*

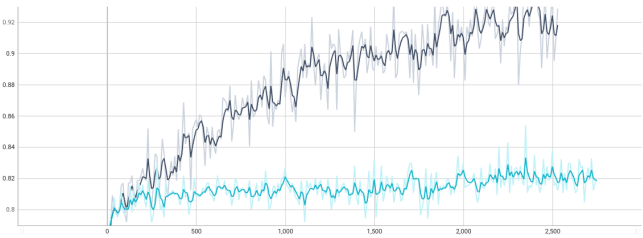


(a) *Loss* - Treino

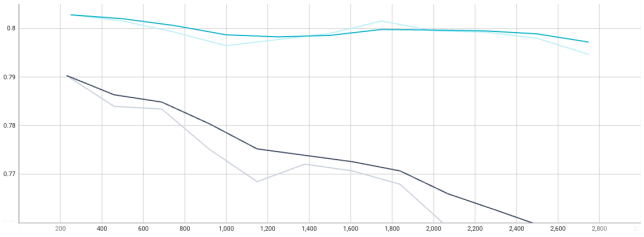


(b) *Loss* - Validação

Figure 14. Evolução *loss* - Redes neurais com 8 e 32 *features*



(a) Acurácia - Treino



(b) Acurácia - Validação

Figure 15. Evolução da Acurácia - Redes neurais com 8 e 32 *features*

	Acurácia	
Modelos	Suicídio	Ansiedade
Não Causal	0.51	0.58
Causal	0.71	0.67

Table 3. Acurácia os modelos de predição contrafactual no conjunto de teste.

nenhuma premissa causal. Na tabela 3 podemos observar que a premissa de causalidade melhorou consideravelmente a acurácia no conjunto de teste.

Alguns experimentos foram feitos também no sentido de validar que a individualidade faz diferença na classificação final. Como exemplo, escolhemos dois indivíduos aleatórios, então usamos a rede para simular como o indivíduo numero 2 se sairia se estivesse nas condições do indivíduo numero 1. Frequentemente vimos resultados em que, apenas por possuir uma individualidade diferente, o indivíduo numero dois possui valores de suicídio e ansiedade diferentes.

5. Explicabilidade

A partir do modelo de classificação apresentado na seção 4.5, aplicamos o método de explicabilidade SHAP.

A implementação deste método foi feita utilizando a classe *ShapleyValueSampling*[2] da biblioteca *Python* Captum, que utiliza a metodologia proposta em (Castro et al., 2009).

Para a aplicação do SHAP sobre o modelo treinado, foi necessário dividir os dados disponíveis entre os diferentes níveis da classe target em que cada amostra se encontra. Assim, o dataset inicial foi dividido em 5, uma divisão para as amostras com cada nível de target, sendo que os níveis possíveis são de 0 a 4.

Após a divisão do dataset e com o modelo de classificação já treinado em mãos, foi possível aplicar o SHAP a cada subconjunto do dataset.

O método utilizado pela classe *ShapleyValueSampling* tem como premissa a amostragem aleatória do dataset. Este é um passo importante, pois o método empregado é muito caro computacionalmente. Portanto, amostrar a base dados possibilita a aplicação do método em tempos de execução razoáveis. Neste trabalho, utilizamos a quantidade padrão de número de amostras desta classe, 25 amostras para cada subconjunto do dataset. A única exceção é o subconjunto representado pelas pessoas que apresentam a feature de ideação suicida no nível 4, que é o máximo para esta feature. Nesse subconjunto somente existem 9 amostras, portanto foram utilizadas todas elas.

[2]Disponível em: https://captum.ai/api/shapley$_value_sampling.html$

Este procedimento foi realizado tanto para o modelo de classificação que utiliza todas as features disponíveis no dataset quanto para o modelo que somente utiliza as features com causa direta para o target.

Ao final desta etapa, obtivemos valores de importância de cada feature na classificação das amostras em cada nível. Esses resultados podem ser vistos nas figuras de 16 a 18. O resultado detalhado do modelo de classificação que utilza somente as features presentes no Markov Blanket do grafo causal podem ser vistos na tabela 4.

Table 4. SHAP Values - Com causalidade - *Target* Suicídio

Feature	**Classe**				
	0	**1**	**2**	**3**	**4**
Drog_familia	-0.011	0.005	0.003	-0.001	-0.002
Suic_familia	0.004	-0.003	0.001	0.001	0.004
Capaz de tomar dec.	0.007	0.018	-0.017	-0.009	-0.007
Estudante	-0.002	0.003	-0.001	-0.001	0.001
Hipocondriase	0.055	-0.021	-0.006	-0.003	0.000
Sentimentos_culpa	-0.133	0.064	0.069	0.024	0.012
Trabalho e interesses	-0.205	0.125	0.084	0.021	0.009
Energia	-0.009	0.012	0.003	-0.004	-0.002

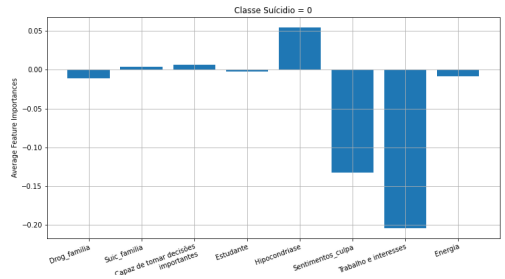
Ao comparar estes resultados entre si, é possível perceber que algumas features com importância relativamente alta no modelo de classificação não causal não estão presentes nas features que de fato têm relação causal com o target. É o caso da feature *sexo* para os níveis 0, 1, 2 e 4. Isso é uma evidência de que o modelo que não considera causalidade está aprendendo correlações espúrias e elas estão sendo relevantes nas predições feitas, o que pode acarretar em problemas ao se utilizar o modelo para prever dados inéditos.

Por outro lado, é possível também perceber que as duas features com maior importância tanto para o modelo causal quanto para o modelo com todas as features são as features de *Sentimento de Culpa* e *Trabalho e Interesse*. Este é um resultado esperado, dado que existe uma relação causal entre essas features e o target.

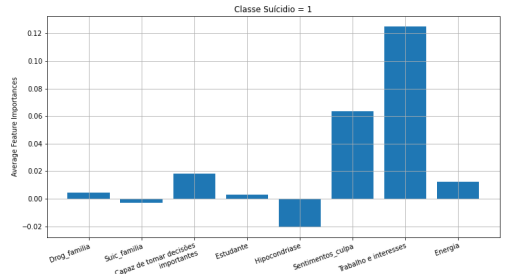
Outra comparação interessante de se fazer é entre os resultados do SHAP aplicado ao modelo de predição com as 8 features e o resultado de pesos das arestas do grafo de causalidade, apresentado na figura 11. O maior peso das arestas do grafo de causalidade se encontra na relação entre a feature *Suicídio na Família* e o target. Já ao se observar as essa mesma feature nos gráficos de importância da figura 16 é possível verificar que ela sempre possui baixos coeficientes. Já a feature *Sentimento de Culpa* é compartilhada como uma das features com aresta de maior peso e também com importância alta.

6. Experimentos Adicionais

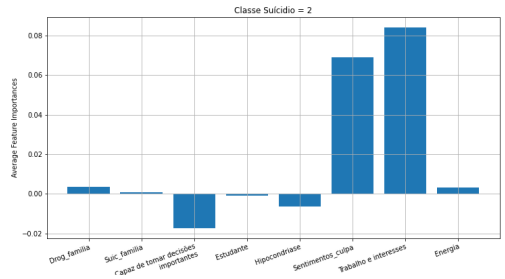
DAGs com dados categóricos. Durante o desenvolvimento do trabalho encontramos alguns desafios, como a maneira correta de lidar com dados tabulados categóricos.



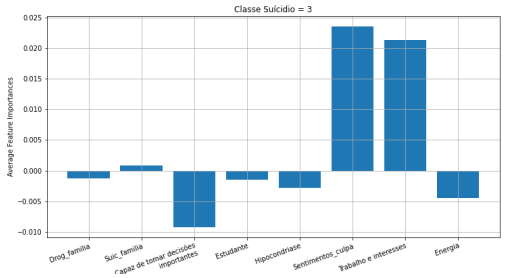
(a) SHAP Values - Categoria 0



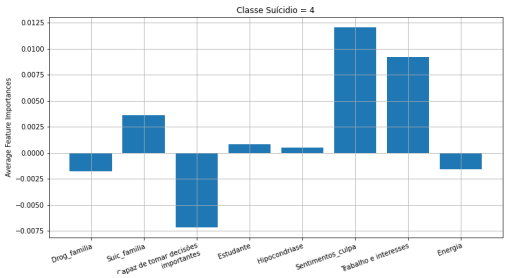
(b) SHAP Values - Categoria 1



(c) SHAP Values - Categoria 2

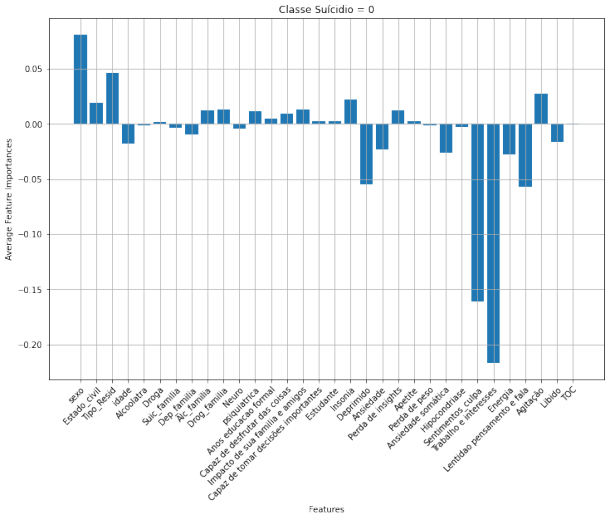


(d) SHAP Values - Categoria 3

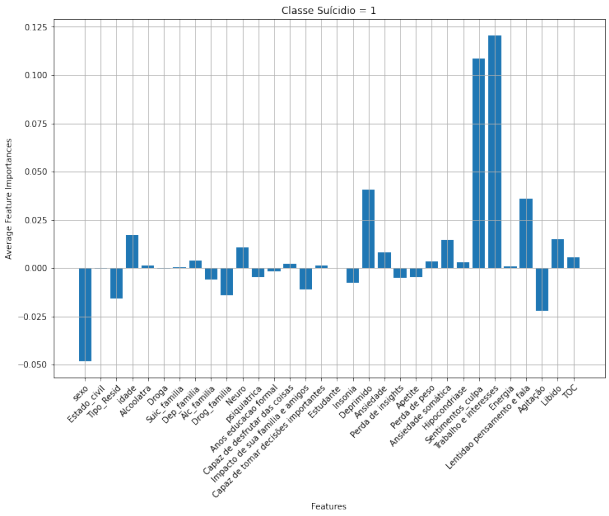


(e) SHAP Values - Categoria 4

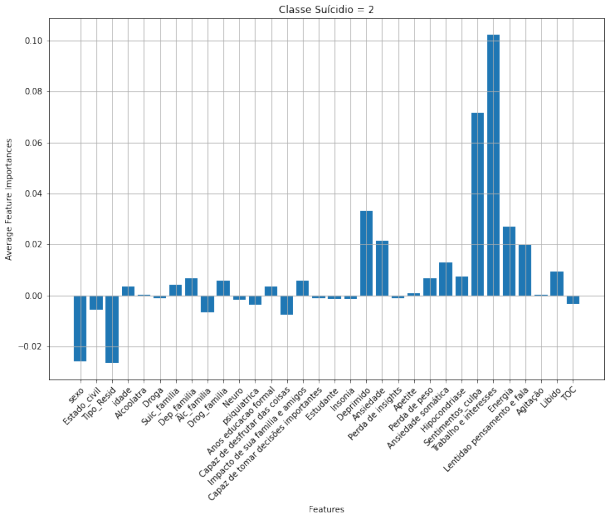
Figure 16. Com causalidade - *Target* Suicídio



(a) SHAP Values - Categoria 0

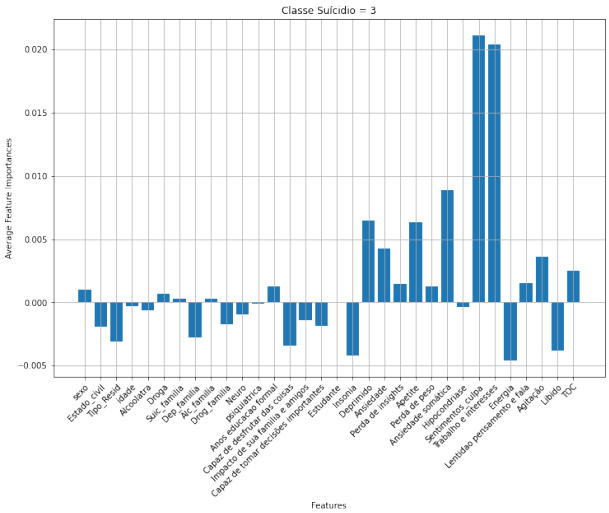


(b) SHAP Values - Categoria 1

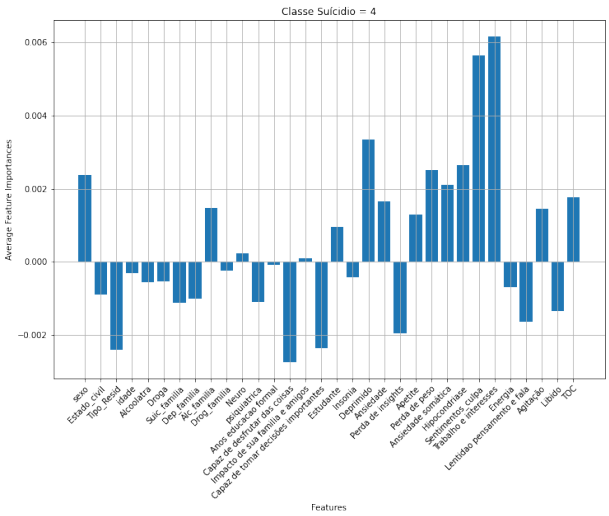


(c) SHAP Values - Categoria 2

Figure 17. Sem causalidade - *Target* Suicídio



(a) SHAP Values - Categoria 3



(b) SHAP Values - Categoria 4

Figure 18. Sem causalidade - *Target* Suicídio

Como sabemos, algoritmos de predição de estrutura causal (DAGs) precisam de atributos reais. Quando possuímos atributos categóricos, não existe um relação numérica correta entre os valores. Para resolver isso propomos uma maneira de projetar os dados categóricos em um espaço numérico onde eles se compartam da forma como esperamos. O método proposto é criar um autoencoder com *embedding*. A inspiração para esse modelo veio da recente popularidade dos grandes modelos de linguagem, e como eles projetam palavras em um espaço numérico. A ideia principal é encontrar uma tarefa simples que utilize essa representação intermediária para ser resolvida. Após alguns experimentos convergimos para um modelo que aprende um *embedding* para cada *feature* e em seguida o usa para reconstruir o dado categórico original. Essa reconstrução garante que não perderemos informação no aprendizado dos *embeddings*. Após o treino, podemos isolar a parte de projeção de features para gerar um novo conjunto de dados contínuo. Com esse conjunto conseguimos utilizar normalmente algoritmos que dependem de propagação de gradientes, como o próprio NOTEARS. Infelizmente não tivemos tempo suficiente para desenvolver mais esse modelo, já que os outros DAGs já tinham sido validados por um especialista. Na figura 19 podemos ver todos os componentes conexos gerados pelo notears. Com esse resultado podemos notar que os *embeddings* foram capaz de incorporar muitas relações causais que realmente fazem sentido.

Redes neurais causais com grafos. Outro desafio que encontramos foi como encorporar um modelo causal estrutural de maneira automática dentro de uma rede neural. A ideia é usar convoluções em grafos para agregar informações na direções causais. Para testar isso resolvemos comparar a performance de duas redes, uma MLP e uma GCN (Graph Convolutional Network) em dados simulados, onde o grafo causal correto é conhecido. Apesar de ao final ambas ser conseguirem uma performance bem similar, podemos observar que, por conseguir utilizar todas as features do grafo, a performance de teste e treino caminharam juntas, sem overfitting mesmo depois de muito treino. Como só fizemos testes com dados simulados e redes simples, ainda existe muito a ser pesquisado na área.

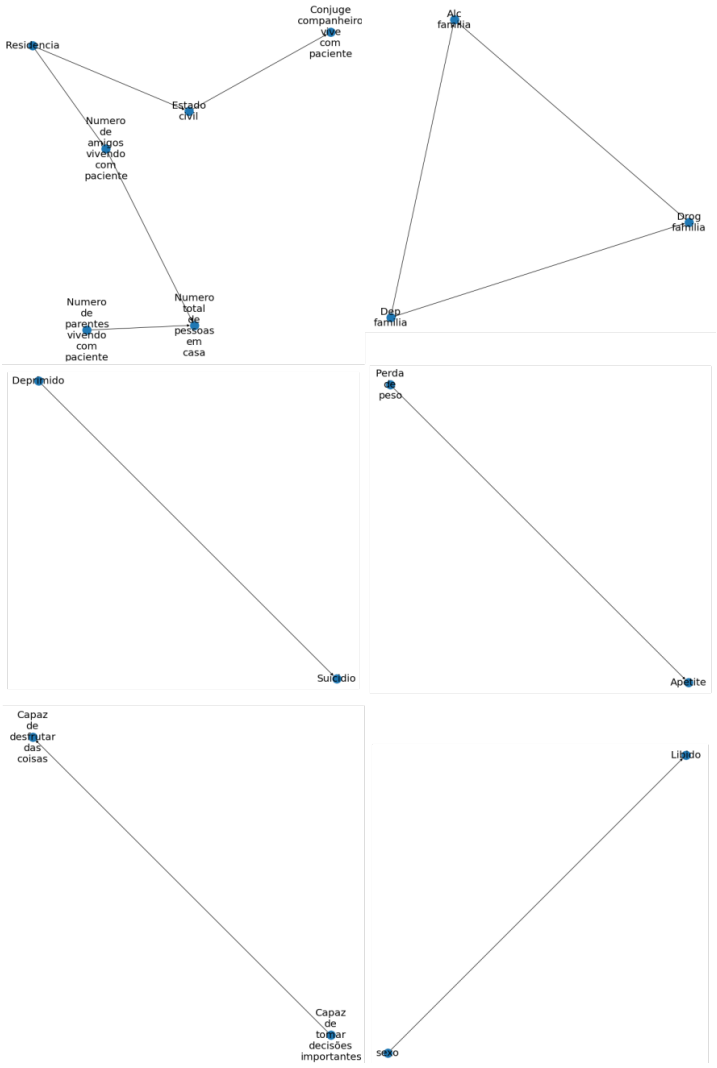


Figure 19. Grafos causais usando a estratégia de *embbedings* e *notears*.

7. Conclusão

As mortes por suicídio representam um grande problema de saúde pública. Em todo o mundo e em números absolutos, os suicídios matam mais que os homicídios e as guerras juntos. No Brasil, os suicídios representam 0,8% do total de óbitos.

A prevenção do suicídio não é uma tarefa fácil e requer ações que considerem aspectos médicos, sociais, psicológicos, familiares, culturais, religiosos e econômicos.

Pelo menos 90% dos que cometem suicídio apresentam um transtorno psiquiátrico, predominantemente depressão. Por isso, o reconhecimento da depressão é o primeiro passo para preveni-lo, especialmente quando existe comorbidade com o abuso de substâncias (Minayo & Souza, 2005).

De acordo com esse cenário, acreditamos que os resultados apresentados nesse trabalho poderão ser úteis para prevenir ideação suicida de pacientes ambulatoriais diagnosticados com depressão, tendo em vista vantagens e avanços obtidos ao incorporar causalidade em aprendizagem de máquina.

No presente trabalho, utilizamos técnicas de Aprendizado Supervisionado Causal para identificar os padrões existentes na predição de ideação suicida, incluindo um modelo contrafactual. Para investigar os resultados alcançados com a Causalidade, empregamos um modelo de classificação, que foi usado como *benchmark*, e *E*xplicadade.

Dentre os resultados alcançados, verificamos melhor desempenho do modelo de classificação com as 8 *features* propostas pelo grafo causal em comparação com o modelo que compreende todas as *features*, o que pode ser justificado pela análise da explicabilidade, que mostra relações não causais sendo consideradas como importantes no modelo com as 32 *features*. Desse modo, o modelo pode envolver correlações espúrias, o que se torna um problema, pois correlação não implica em causalidade.

Portanto, é importante perceber que nem sempre uma forte correlação entre duas variáveis estabelece uma relação de causa e efeito entre elas, principalmente em questões da área de saúde.

References

Beaumont, P., Horsburgh, B., Pilgerstorfer, P., Droth, A., Oentaryo, R., Ler, S., Nguyen, H., Ferreira, G. A., Patel, Z., and Leong, W. Causalnex, 10 2021. URL https://github.com/quantumblacklabs/causalnex.

Bühlmann, P. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Computers Operations Research*, 36(5):1726–1730, 2009. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2008.04.004. URL https://www.sciencedirect.com/science/article/pii/S0305054808000804. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

Cloudera. Causality for machine learning. https://ff13.fastforwardlabs.com/, 2020. (Accessed on 09/19/2022).

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., and Nock, M. K. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological bulletin*, 143(2):187, 2017.

Kennedy, C. J., Bacon, G., Sahn, A., and von Vacano, C. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application, 2020. URL https://arxiv.org/abs/2009.10277.

Lee, J. D. M. C. K. and Toutanova, K. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Liu, X., Liu, X., Sun, J., Yu, N. X., Sun, B., Li, Q., Zhu, T., et al. Proactive suicide prevention online (pspo): machine identification and crisis management for chinese social media users with suicidal thoughts and behaviors. *Journal of medical internet research*, 21(5):e11705, 2019.

Marchezini, G. F., Lacerda, A. M., Pappa, G. L., Meira, W., Miranda, D., Romano-Silva, M. A., Costa, D. S., and Diniz, L. M. Counterfactual inference with latent variable and its application in mental health care. *Data Mining and Knowledge Discovery*, 36(2):811–840, 2022.

Minayo, M. C. d. S. and Souza, E. R. d. Suicídio: violência auto-infligida. *Impacto da violência na saúde dos brasileiros*, pp. 205–240, 2005.

Natesan, P., Nandakumar, R., Minka, T., and Rubright, J. D. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422, 2016.

Orabi, A. H., Buddhitha, P., Orabi, M. H., and Inkpen, D. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 88–97, 2018.

Organization, W. H. et al. Suicide: one person dies every 40 seconds. *URL: https://www. who. int/news/item/09-09-2019-suicide-one-person-dies-every-40-seconds*, 2019.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M., et al. Sequenced treatment alternatives to relieve depression (star* d): rationale and design. *Controlled clinical trials*, 25(1): 119–142, 2004.

Schölkopf, B. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804. 2022.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Sullivan, E. M., Annest, J. L., Luo, F., Simon, T. R., and Dahlberg, L. L. Suicide among adults aged 35–64 years—united states, 1999–2010. *Morbidity and Mortality Weekly Report*, 62(17):321, 2013.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34:16196–16208, 2021.

Veloso, A. and Ziviani, N. Explainable death toll motion modeling: Covid-19 narratives and counterfactuals. *medRxiv*, 2020. doi: 10.1101/2020.07.04.20146423. URL https://www.medrxiv.org/content/early/2020/07/06/2020.07.04.20146423.