

CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC		
(X) PRÉ-PROJETO	() PROJETO	ANO/SEMESTRE: 2022/2

ANÁLISE DO COMPORTAMENTO DEPRESSIVO E SUICÍDIA DO MUNICÍPIO DE BLUMENAU

Martha Lanser Bloemer

Prof. Aurélio Faustino Hoppe - orientador

1 INTRODUÇÃO

O suicídio é um fenômeno complexo, multifatorial, que ocorre em todas as regiões do mundo. Segundo a Organização Mundial de Saúde - OMS (2021), mais de 700 mil pessoas morrem todos os anos no mundo em decorrência do suicídio. Dessa forma, pode-se afirmar que ocorre em média, um suicídio a cada 45 segundos no mundo. No Brasil, a cada 45 minutos um brasileiro morre vítima de suicídio. Assim como, para cada caso de suicídio, há entre 10 e 20 tentativas, ou seja, quem tentou suicídio está muito mais vulnerável e suscetível a cometê-lo ao longo do tempo.

Segundo Sousa (2019), uma tentativa de suicídio é o maior preditor de nova tentativa ou do suicídio pois quase 100% das pessoas que cometeram suicídios já enfrentavam problemas mentais, principalmente relacionados à depressão. Moraes (2021) destaca que os principais fatores de risco para o comportamento suicida são: transtornos mentais que estão presentes em mais de 90% das pessoas que cometem suicídio, sexo, idade, ambiente familiar desestruturado, uso abusivo de substâncias alcoólicas e psicoativas, doenças, especialmente aquelas que causam invalidez ou dor crônica, desemprego e pobreza. Entretanto, segundo o autor, muitas pessoas justificam o suicídio apenas pelos fatores primários, como brigas em relacionamentos, desemprego, perda de um parente, etc. Já Sousa (2019) ressalta que os fatores são mais abrangentes, pois aquelas pessoas que apresentam comportamento suicida já carregam outros fatores, tendo uma predisposição que pode ser potencializada quando expostos a outros problemas, seja problemas familiares, econômicos, uso de substâncias psicoativas e álcool.

Segundo a Organização Mundial de Saúde (2021), algumas pesquisas apresentam determinados fatores como desencadeadores do comportamento suicida, entre eles estão: transtornos de humor e depressão. Além disso, a maioria das pessoas com transtornos mentais não tem acesso a tratamento eficiente em razão do serviço e suporte não estarem disponíveis, acessíveis ou fora da sua capacidade financeira assim como, o preconceito as impede de buscar ajuda. Ainda segundo a OMS, a relação entre suicídio e transtornos mentais, em especial depressão e alcoolismo, está bem definida em países desenvolvidos. Ao mesmo tempo, aponta que diferentes padrões internacionais de suicídio, mudanças nas taxas, características e métodos de suicídio, destacam a necessidade de cada país melhorar sua compreensão, qualidade e pontualidade nos seus dados relativos a suicídio.

No Brasil, o Ministério da Saúde, por meio da Agenda de Ações Estratégicas para a Vigilância e Prevenção do Suicídio e Promoção da Saúde no Brasil, aponta para a necessidade de sensibilizar e mobilizar toda a sociedade civil e os vários setores (educação, justiça, social, previdência, agricultura e outros) nas esferas federal, estadual e municipal, para atuar sobre os determinantes sociais relacionados a ocorrência do suicídio, dentre eles os socioeconômicos, étnico-raciais, de trabalho e ocupação (MINISTÉRIO DA SAÚDE, 2017).

Em Santa Catarina, considerando as três cidades mais populosas, Blumenau, Florianópolis e Joinville, no ano de 2021, o número de suicídio a cada 100 mil habitantes foi de 9,82, 8,71 e 5,95, respectivamente. Em Blumenau, a incidência de suicídios foi 12,77% maior do que em Florianópolis e 65,03% maior em comparação a Joinville (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2022). Contudo, para viabilizar uma análise representativa dos determinantes sociais e seus impactos nos fenômenos de depressão e suicídio do município, faz-se necessário o acesso a diversos bancos de dados de instituições, que coletam informações dos munícipes, que estejam vinculadas tanto aos fenômenos quanto aos determinantes sociais, sendo eles: (i) os dados referentes ao consumo de medicação psicotrópica sob poder da Vigilância Sanitária, (ii) dados referente as mortes no município e de violência, inclusive por autolesão, da Vigilância Epidemiológica (VE), (iii) dados sobre os atendimentos psicológicos e psiquiátricos dos Centros de Atendimento Psicossociais (CAPS) do município de Blumenau, (iv) dados do censo do Instituto Brasileiro de Geografia e Estatística (IBGE), e (v) dados dos atendimentos psiquiátricos de Blumenau do Corpo de Bombeiros Militar de Santa Catarina (CBMSC).

Machado e Pereira (2020), apontam que a literatura evidencia um crescimento no uso de técnicas de Inteligência Artificial (IA) para apoiar profissionais da saúde quanto aos diagnósticos e tratamentos de doenças a partir de análise de dados ou imagens. Diante deste contexto, este trabalho propõe uma solução envolvendo análise de dados e algoritmos de aprendizado de máquina no intuito de estabelecer o panorama municipal de pessoas com depressão e suicidas da cidade Blumenau não apenas do ponto de vista estatístico, mas sim, em relação ao aspecto comportamental.

1.1 OBJETIVOS

O objetivo deste trabalho é a criação de um modelo computacional que seja capaz de estabelecer o panorama municipal de pessoas com depressão e suicidas da cidade de Blumenau.

Os objetivos específicos são:

- verificar se técnicas de Machine Learning (ML) e análise de dados podem auxiliar no que diz respeito a caracterização do comportamento depressivo ou suicidas;
- utilizar e combinar modelos preditivos no intuito de identificar as variáveis preditoras que possuem maior relevância em relação ao comportamento suicida;
- avaliar a eficiência do modelo desenvolvido;
- comparar os padrões encontrados com as características do suicídio encontradas na literatura.

2 TRABALHOS CORRELATOS

Nesta seção são descritos e comparados trabalhos que possuem características semelhantes aos principais objetivos deste projeto. A seção 2.1 aborda a investigação de que a melhora da depressão pode aumentar a ideação suicida com a aplicação de Machine Learning (ML) para descoberta de padrões em dados longitudinais (GONG; SIMON; LIU, 2019). A seção 2.2 discorre sobre a análise do ambiente de trabalho como fator de risco para o suicídio de trabalhadores assalariados utilizando Machine Learning (PARK; LEE, 2022). Por fim, a seção 2.3 apresenta um estudo que examina o uso de técnicas de ML aplicados em dados de registros nacionais suecos para previsão de tentativa ou ato de suicídio (CHEN *et al.*, 2020).

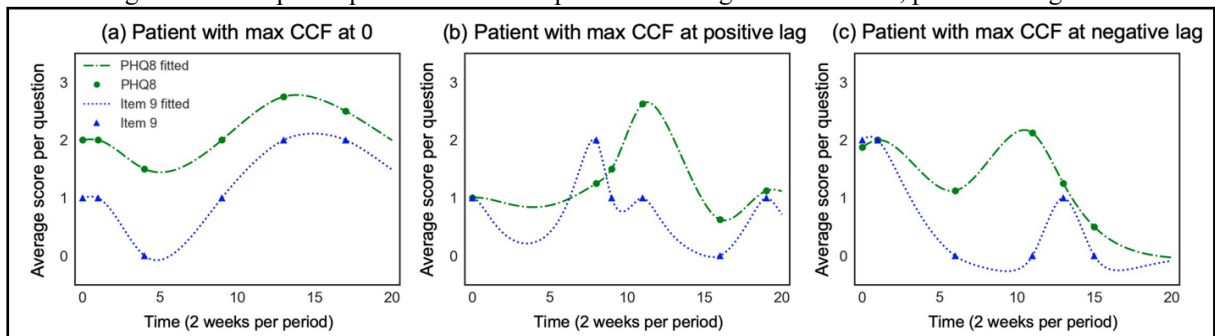
2.1 A MACHINE LEARNING DISCOVERY OS LONGITUDINAL PATTERNS OF DEPRESSION AND SUICIDAL IDEATION

Gong, Simon e Liu (2019) focaram na correlação temporal entre a depressão e a ideação suicida, assumindo que existem padrões na progressão da depressão e trajetórias de ideação suicida. Os autores, tinham como intuito estabelecer padrões heterogêneos de trajetória de depressão e investigar a afirmação feita por Mittal *et al.* (2009) de que a melhora no quadro de depressão pode ser acompanhada de ideação suicida.

Gong, Simon e Liu (2019) se basearam em dados processados a partir de questionários Patient Health Questionnaire (PHQ-9) autoadministrados que visam avaliar a frequência e gravidade dos sintomas de depressão nas últimas duas semanas dos pacientes. Os PHQ-9 são formados por 9 questões que se propõem a medir o grau de depressão e ideação suicida. A partir deles, busca-se a frequência nas últimas duas semanas de sintomas característicos de depressão como (i) falta de vontade ou prazer, (ii) sono excessivo, (iii) cansaço excessivo, (iv) distúrbios alimentares e (v) sentimento de que estar morto seria melhor. Em específico, a última questão (item 9) do questionário refere-se a ideação suicida. Para cada uma das respostas é atribuída uma pontuação que varia de 0 a 3 de acordo com a frequência relatada, sendo 3 pontos para frequência quase diária e 0 pontos para nenhuma frequência.

Gong, Simon e Liu (2019) utilizaram o *dataset* disponibilizado pelo Mental Health Research Network (MHRN) que continha a participação de 9306 indivíduos. Os autores selecionaram 610 indivíduos que possuem pelo menos 6 pontos registrados de PHQ-9 durante 20 períodos consecutivos de duas semanas da amostra original. No pré-processamento, segundo Gong, Simon e Liu (2019), os registros foram transformados em trajetórias contínuas, baseado no método regressão de processos gaussianos aplicado a dados clínicos (LASKO; DENNY; LEVY, 2013). Posteriormente, Gong, Simon e Liu (2019) calcularam a Cross Correlation Function¹(CCF) entre os pontos das primeiras 8 questões e o item 9 do PHQ-9, conforme demonstra a Figura 1.

Figura 1 - Exemplo de paciente com CCF positivo com lag com valor zero, positivo e negativo



Fonte: Gong, Simon e Liu (2019).

¹ CCF mede a similaridade entre duas séries em função de um deslocamento relativo, ou seja, uma defasagem de k unidades de tempo (GONG; SIMON; LIU, 2019).

Segundo Gong, Simon e Liu (2019), os valores de CCF variam entre -1 e 1, tendo alta similaridade quando próximo de 1, indicando uma correlação negativa quando próximo de -1 e baixa ou inexistente quando próximo de 0. Nestes cenários, com $k = 0$ demonstrado na primeira coluna na Figura 1 (item a), o comportamento da CCF aconteceu simultaneamente nas duas séries, caso k seja positivo uma série esta adianta em relação a outra, visto na coluna do item b da Figura 1 e com k negativo a série está atrasada em relação outra apresentado coluna do item c da Figura 1.

Para examinar a crença de que a melhora da depressão e o aumento da ideação suicida podem ocorrer simultaneamente, Gong, Simon e Liu (2019) utilizaram o Coeficiente de Correlação de Spearman (CCS). O CCS, segundo os autores, não requer dados contínuos, logo ele foi aplicado para medir a força e a direção de associação entre as variações de pontos das 8 questões e a variação de pontos do item 9. Além disso, também foram coletados os seguintes dados das variações de pontuação: (i) taxa de variação em 1 mês, (ii) a primeira observação (mês 0) e a (iii) média de pontos entre o mês inicial (mês 0) e os meses 3, 6 e 9.

A partir dos dados coletados, Gong, Simon e Liu (2019) aplicaram os métodos propostos por Lasko, Denny e Levy (2013), para descobrir fenótipos computacionais em dados clínicos esparsos, irregulares e com ruído. Para isto, utilizaram uma rede neural artificial (RNA) do tipo autoencoder para extrair os subtipos de depressão. De forma simplificada, um autoencoder é uma rede que procura copiar os dados de entrada para a saída, e internamente possui um código usado para representar a entrada. A camada de entrada tem m nós representando as trajetórias ajustadas, a camada latente possui h nós que representam as características ocultas pela transformação de entrada, e a camada de saída possui m nós que representam uma reconstrução dos dados de entrada transformados da camada latente.

Segundo Gong, Simon e Liu (2019), na aplicação da validação cruzada, as amostras de treinamento foram divididas aleatoriamente em cinco subamostras de igual tamanho. Uma das subamostras foi usada como dado de validação e as outras quatro para treinamento. O processo de validação cruzada foi repetido 5 vezes com cada uma das cinco subamostras, para que cada amostra fosse usada como validação. Além disso, os autores também apontam que o critério de erro entre a trajetória original e a reconstruída foi o erro quadrático médio.

Cada paciente possui um vetor de ativação que representa a semelhança com cada grupo. Segundo Gong, Simon e Liu (2019) é possível agrupá-los usando o vetor de ativação para descobrir os subtipos de depressão. Para isso, utilizou-se o algoritmo k-means, sendo que a quantidade de clusters foi definida através do método da curva do cotovelo (Elbow Curve). Os autores destacam que todas as análises foram realizadas na linguagem Python e os algoritmos GPR e k-means foram utilizados a partir do pacote de código aberto scikit-learn. Já o pacote Theano foi utilizado para a aplicação da rede neural.

De acordo com Gong, Simon e Liu (2019), através da CCF e da correlação de Spearman confirmou-se a hipótese que os pontos das 8 questões e do item 9 tendem a mudar na mesma direção, onde a CCF no seu maior valor tinha o lag de 0. Além disso, os autores também constataram a existência de uma forte correlação temporal entre os pontos das primeiras 8 questões e o item 9, tendo acontecido em sincronia em 45% dos pacientes. Neste ponto observaram também uma distribuição simétrica para lag positivo e negativo, indicando que há pacientes em que a ideação suicida precede ou sucede os sintomas de depressão. Os autores afirmam ainda que seu método traz informações úteis para as práticas de monitoramento de depressão e prevenção de suicídio, se for possível determinar o padrão de progressão de um paciente usando a análise de correlação cruzada, saber a trajetória principal pode fornecer uma forte evidência da trajetória seguinte.

Segundo Gong, Simon e Liu (2019) este é o primeiro trabalho que analisa a correlação temporal entre os valores das 8 questões de depressão e o item 9 de ideação suicida utilizando valores do PHQ-9 de prontuários médicos eletrônicos. As conclusões foram similares na análise de correlação cruzada utilizando as trajetórias ajustadas pelo método gaussiano, e na análise de mudança de padrão utilizando valores não ajustados de PHQ-9. Dentre os registros dos pacientes analisados, verificou-se que os valores das 8 questões de depressão tendem a mudar na mesma direção do item 9. Além disto, encontraram que de 8% a 13% dos pacientes possuem aumento da ideação suicida juntamente com a melhora da depressão, sendo este o primeiro estudo que mostrou evidências deste fenômeno utilizando dados de prontuário médico eletrônico.

Como limitações, Gong, Simon e Liu (2019) destacam primeiramente que a regressão do processo gaussiano não é o mais indicado quando os dados são esparsos e o item 9 está frequentemente zerado. Neste sentido, os autores apontam que a aplicação da interpolação com spline linear poderia gerar melhores resultados. Gong, Simon e Liu (2019) também ressaltam que o grupo de pacientes estudados foi pequeno, devido a seleção em relação a frequência no monitoramento, o que pode tornar o estudo enviesado para casos mais severos. Por último, a estrutura latente é um padrão local, isso significa que não é especificado o início da trajetória latente no histórico da doença. Isso limita a aplicação deste modelo para outras doenças crônicas, pois nem todas apresentam periodicidade e descobrir seu padrão local não garante a tendência futura.

2.2 A MACHINE LEARNING APPROACH FOR PREDICTING WAGE WORKERS SUICIDAL IDEATION

Quando um trabalhador sofre com problemas mentais, ele pode ter várias dificuldades como baixa autoconfiança, tensão, erros frequentes, cansaço excessivo e entrar em conflito com colegas, levando-o a um baixo desempenho. Além disso, em casos extremos, os problemas com a saúde mental podem levar ao suicídio. Diante disso, Park e Lee (2022) tinham como intuito analisar se a ideação suicida de trabalhadores assalariados pode ser prevista com características individuais, estados emocionais e características do ambiente de trabalho. Foram observados também características demográficas, histórico médico e diferenças no estilo de vida pessoal de um grupo de trabalhadores com pensamentos suicidas, assim como variações no estado emocional, níveis de estresse, situação de saúde e fatores de depressão. A partir disso, Park e Lee (2022) procuraram determinar quais características afetam a ideação suicida, analisando aspectos como jornada de trabalho, tipo de trabalho e profissão e, se técnicas de aprendizado de máquina podem identificar de forma eficaz quais fatores estão relacionados ao ambiente de trabalho ou emocionais.

Park e Lee (2022) utilizaram dados levantados pelo Korean National Health and Nutrition Examination Survey (KNHANES²). Foram selecionados dados das pesquisas dos anos de 2007 a 2013, de 2015 e de 2017, pois continham questões relacionadas a análise de ideação suicida, tais como: durante o ano passado, você sentiu em algum momento que estava disposto a morrer? Ao total, foram filtrados 12865 trabalhadores com idade igual ou superior a 19 anos. Destes, Park e Lee (2022) excluíram 49 pessoas que não responderam a pergunta sobre ideação suicida.

Segundo Park e Lee (2022), foram selecionadas 23 variáveis que apresentam alguma relação com a ideação suicida. Elas possuem informações sobre a saúde (histórico de variadas doenças, índice de massa corporal etc.), características individuais (educação, idade, estado civil etc.) e de trabalho (ocupação, tipo de turno, horas trabalhadas por semana etc.) dos indivíduos. A partir delas, os autores desenvolveram dois modelos computacionais, um levando em consideração todas as variáveis para previsão de ideação suicida e o outro, utilizando apenas as relacionadas ao trabalho. O algoritmo de aprendizado de máquina utilizado por Park e Lee (2022) foi a Floresta Aleatória (FA). Dentre os benefícios citados pelos autores, tem-se o fato dele ser amplamente utilizado devido a sua boa performance preditiva, especialmente em pesquisas relacionadas a saúde mental.

Inicialmente, Park e Lee (2022) subdividiram a base de dados em 70% dos dados para treinamento e 30% para teste. Para evitar *overfitting*, os autores aplicaram validação cruzada com 10 dobras e repetiram o processo cinco vezes. Para encontrar os melhores parâmetros para a Floresta Aleatória, Park e Lee (2022) utilizaram o método *grid search*. Posteriormente, a performance do modelo foi avaliada utilizando o cálculo de acurácia, sensibilidade, especificidade, valor preditivo positivo (precisão), valor preditivo negativo, *area under the curve* (AUC) e F1 Score. Na implementação, utilizou-se a linguagem R versão 4.1.2, com o pacote “mice” para processar dados faltantes, o pacote “caret” para a validação cruzada e o pacote “ROCR” para o AUC.

Park e Lee (2022) utilizaram dois modelos de floresta aleatória similares, porém diferenciando as variáveis utilizadas do *dataset*. No primeiro modelo foram utilizadas 23 variáveis, entre elas dados socioeconômicos, histórico médico, características do trabalho e estilo de vida, que obteve um F1 Score de 0,915. Já o segundo modelo, utilizou apenas as 5 variáveis relacionadas ao trabalho, sendo elas ocupação, regime de trabalho, jornada de trabalho, tipo de turno e horas trabalhadas por semana, alcançando 0,778 de F1 Score. Park e Lee (2022) ressaltam que o primeiro modelo atingiu AUC de 0,922 e o segundo 0,818. Além disso, os autores também destacam que um valor de AUC superior ou igual a 0,75 é considerado clinicamente útil para previsão de ideação suicida.

Segundo Park e Lee (2022), as principais limitações identificadas foram: (i) apenas fatores multissetoriais foram utilizados e a avaliação dos ambientes de trabalho não foi realizada. No futuro, os autores sugerem a necessidade da realização de uma pesquisa com os trabalhadores para confirmar a consistência dos dados; (ii) apenas a ideação suicida foi avaliada. Por último, (iii) a aplicação de outras técnicas de aprendizado de máquina ou inteligência artificial para previsão de suicídio.

2.3 PREDICTING SUICIDE ATTEMPT OR SUICIDE DEATH FOLLOWING A VISIT TO PSYCHIATRIC SPECIALTY CARE: A MACHINE LEARNING STUDY USING SWEDISH NATIONAL REGISTRY DATA

Chen *et al.* (2020) mediram a performance de diversos modelos de aprendizado de máquina aplicados a dados de registros nacionais de saúde dos suecos. Segundo os autores, tais modelos tinham o intuito de estabelecer

² KNHANES é uma pesquisa nacional da Coreia do Sul sobre a situação de nutrição e saúde de civis não institucionalizados (PARK; LEE, 2022).

a previsão prognóstica de tentativa de suicídio ou morte por suicídio dentro de 30 ou 90 dias após uma internação ou consulta ambulatorial de atendimento psiquiátrico.

Segundo Chen *et al.* (2020), os dados utilizados são provenientes dos registros médicos de partos, registro de população total (contém dados como sexo, nascimento, morte migração e relações familiares), o registro multigeracional (vincula indivíduos com pais biológicos), integração longitudinal para seguro de saúde e estudos de mercado de trabalho (dados socioeconômicos), registro nacional de pacientes, registro de medicação prescrita e o registro nacional de crime. Ainda de acordo com os autores, o *dataset* consiste em consultas ambulatoriais ou de internação com psiquiatras por pacientes de 18 a 39 anos na Suécia entre as datas 01/01/2011 e 31/12/2012, com diagnóstico principal de qualquer transtorno mental ou comportamental de acordo com a Classificação Internacional Diagnósticos (CID-10: F00-F99). Para garantir a qualidade de dados e minimizar a quantidade de variáveis preditoras vazias, apenas pacientes nascidos na Suécia foram incluídos no estudo. Adicionalmente pacientes que emigraram antes da consulta, faleceram no mesmo dia da consulta ou não possuíam informação de um dos pais foram excluídos. No final, o *dataset* se estabeleceu com 541300 consultas de 126205 pacientes dentro do período analisado.

Inicialmente, Chen *et al.* (2020) treinaram 4 modelos: (i) Elastic Net Penalized Logistic Regression (ENPLR), (ii) Floresta Aleatória (FA), (iii) Gradient Boosting (GB) e (iv) uma Rede Neural (RN). Segundo os autores, eles optaram por estes algoritmos pois possuem uma abordagem analítica, tornando possível agregá-los com um método de ensemble³. Para cada modelo, utilizou-se o *grid search* para definir os hypermâmetros com a validação cruzada de 10 dobras e como métrica o AUC. Em seguida, comparou-se os 4 algoritmos e modelos treinados com os formados pela agregação dos algoritmos (método ensemble). Segundo Chen *et al.* (2020), selecionou-se o modelo com menor diferença entre o AUC de validação e treinamento. Ainda segundo os autores, utilizou-se o valor de AUC no conjunto de teste para avaliar a discriminação do modelo, ou seja, sua capacidade de separar os que podem ou não ter uma experiência com suicídio. Neste caso, o método de Delong's foi aplicado para determinar o intervalo de confiança do AUC do conjunto de teste. Além disso, também foram utilizadas as métricas que medem a sensibilidade, especificidade, valor preditivo positivo (VPP), valor preditivo negativo (VPN), Brier Score, *predictor importance score* e análise da curva de aprendizado do modelo. Ressalta-se que os autores utilizaram os softwares SAS 9.4 e R 3.6.1 para construção do conjunto de dados e análises descritivas. Já os pacotes Scikit-learn e XGBoost da linguagem Python 3.6.7 foram utilizados para as análises de aprendizado de máquina durante a definição e avaliação dos modelos.

Por fim, Chen *et al.* (2020) treinaram um modelo considerando um número menor de variáveis (sexo, idade na visita, diagnóstico e medicamentos controlados) assim como, utilizaram o método de diminuição de AUC proposto por Hanley e McNeil (1982). Tal técnica permite avaliar o poder preditivo de um modelo que não depende de uma integração complexa entre bases, ou seja, possibilita apenas a utilização dos dados disponíveis em um prontuário eletrônico que, segundo os autores, é mais viável.

Segundo Chen *et al.* (2020), os resultados da validação cruzada demonstraram que os AUCs dos melhores modelos com algoritmos individuais eram similares. Já o modelo com agregação dos 4 melhores algoritmos distintos apresentou um valor maior de AUC de validação. Por esse motivo, Chen *et al.* (2020) optaram pelo modelo com agregação dos 4 algoritmos distintos (ENPLR, FA, GB e RN) aplicado a uma versão com previsão de ato suicida em 30 dias após consulta e uma para previsão de ato suicida em 90 dias da consulta. A partir desses modelos, os valores de AUC no conjunto teste foram de 0,88 (95% Intervalo de Confiança (IC) = 0,87 a 0,89) para a previsão de 30 dias e 0,89 (95% CI=0,88 a 0,90) para o de 90 dias. Além disso, os autores apontam que no limiar de risco de 95% os modelos de 90 dias e 30 dias apresentaram respectivamente, sensibilidade de 47,2% e 52,9%, especificidade de 96,6% e 95,9%, VPP de 34,9% e 18,7%, e de VPN 97,9% e 99,1%. Já os valores de Brier Score foram estimados em 0,028 e 0,015, ambos próximos de 0, indicando boa calibragem de modelo.

Por fim, Chen *et al.* (2020), ressaltam a necessidade para o aumento da qualidade dos dados, tendo em vista que foram selecionados indivíduos que possuíam registros médicos de nascimento a partir de 1972, fazendo com nas datas estudadas os pacientes mais velhos estivessem com 39 anos. Dessa forma, segundo os autores, tem-se então um modelo que não generaliza bem para menores de 18 anos e adultos acima de 40 anos. Outra limitação diz respeito a escolha das variáveis preditoras, pois foram definidas de forma arbitrária. Já em relação aos modelos, eles não foram validados considerando outras bases de dados, por isso a capacidade de generalização em outras populações permanece desconhecida.

3 PROPOSTA

Esta seção apresenta a justificativa para o desenvolvimento desse trabalho, os principais requisitos e a metodologia de desenvolvimento que será utilizada. Também são relacionados os assuntos e as fontes bibliográficas que irão fundamentar o estudo proposto.

³ Ensemble: método que agrega as classificações feitas por outros classificadores para gerar uma nova classificação. (DIETTERICH, 2000)

3.1 JUSTIFICATIVA

No Quadro 1 **Erro! Fonte de referência não encontrada.** é apresentado o comparativo entre os trabalhos correlatos. As linhas representam as características e as colunas os trabalhos.

Quadro 1 - Comparativo dos trabalhos correlatos

Trabalhos Características	Gong, Simon e Liu (2019)	Park e Lee (2022)	Chen <i>et al.</i> (2020)
Objetivo	Estabelecer padrões heterogêneos de trajetória de depressão e investigar a afirmação de que a melhora no quadro de depressão pode ser acompanhada de maior ideação suicida	Investigar se o risco de ideação suicida de trabalhadores pode ser previsto de forma eficaz com fatores relacionados ao ambiente de trabalho	Examinar o potencial de previsão de comportamento suicida dos dados de registros nacionais suecos
Perfil do público-alvo	Pacientes em tratamento de depressão	Trabalhadores assalariados	Pessoas nascidas na Suécia
Características do dataset utilizado	Dados de prontuários eletrônicos de 610 pacientes com variação de pelo menos 6 pontos na pergunta 9 do questionário PHQ-9 em um período de 20 quinzenas, 62% com idade superior a 45 anos e 68% do sexo feminino	Dados oriundos de 12816 trabalhadores com mais de 19 anos em um período de 9 anos de levantamentos	Dados oriundos de 541300 consultas de 126205 pacientes atendidos entre as datas 01/01/2011 e 31/12/2012 composto por 54% do sexo feminino e 46% do sexo masculino, idades entre 18 e 39 anos
Variáveis utilizadas	Pontuação nas 9 perguntas do questionário PHQ-9	Demográficas, histórico médico, características de trabalho, e escolhas pessoais de estilo de vida	Demográficas, socioeconômicas, registros médicos eletrônicos, criminalidade, histórico de saúde familiar e crime
Algoritmo de Aprendizado de Máquina	Regressão por processos gaussianos, k-means e RNA	Floresta Aleatória	Elastic net penalized logistic regression, floresta aleatória, gradient boosting e RNA
Métricas utilizadas	Similaridade entre subgrupos	Acurácia, sensibilidade, especificidade, valor preditivo positivo/negativo, (AUC) e F1 Score	AUC, sensibilidade, especificidade, valor preditivo positivo/negativo

Fonte: elaborado pela autora.

A partir do Quadro 1, pode-se perceber que todos os trabalhos possuem objetivos diferentes, porém tendo atos suicidas e/ou saúde mental como principal objeto de investigação. Além disso, todos utilizam dados já existentes em sistemas dos governos ou entidades de saúde que são utilizados como entrada nos mais diversos algoritmos de aprendizado de máquina.

Neste contexto, Chen *et al.* (2020) consolidaram várias bases de dados do governo sueco. A partir deles, desenvolveram um preditor de atos suicidas dentro de 30 e 90 dias após uma consulta com atendimento psiquiátricos. Para isso, foram utilizados e vinculados dados do Registro Nacional de Pacientes (desde 1964), Registro Médico de Nascimentos (iniciado em 1973), Registo Total de População (desde 1968), Registro Multigeracional, Registro Nacional de Crimes, Registro de Drogas Prescritas (desde 2005) entre outros.

Chen *et al.* (2020) analisaram a performance dos algoritmos EN, FA, GB, RN e de suas combinações com uso do método ensemble. Os valores de AUC alcançados pelos modelos de previsão para 90 e 30 dias após consulta foram respectivamente para EN 0,8721 e 0,8883, para FA 0,8703 e 0,8875, para GB 0,8741 e 0,8902 e para RN 0,8735 e 0,8878. No método de ensemble foram realizadas agregações de dois ou três dos algoritmos previamente citados com todas as combinações possíveis. Por fim, o modelo com agregação de todos os algoritmos EN, FA, GB e RN obteve melhor desempenho, pois apresentou menor variação entre o valor de AUC de validação e treino para até 90 dias de 0,8751 e de 0,8910 para até 30 dias.

Park e Lee (2022) utilizaram dados do sistema de vigilância da saúde e nutrição sul coreano (KHANES), tendo como objetivo prever atos suicidas dentre os trabalhadores assalariados através do algoritmo de floresta

aleatória. Os autores justificam a escolha considerando o bom desempenho da aplicação da FA em outros estudos relacionados a saúde mental. Park e Lee (2022) desenvolveram dois modelos de predição de ideação suicida: (i) com diversas variáveis preditoras e, dentre elas 5 variáveis relacionadas ao trabalho, (ii) apenas considerando as 5 variáveis relacionadas ao trabalho. Os autores relatam que quando utilizaram as 23 variáveis relacionadas ao trabalho, vida e estado emocional, o método floresta aleatória alcançou precisão de 98,9%. Por outro lado, ao considerar apenas as variáveis relacionadas ao trabalho, 97,4% dos casos eram previstos. Park e Lee (2022) concluem que a aplicação de algoritmos de aprendizado de máquina a partir de dados/fatores do ambiente de trabalho, podem prever a ideação suicida de forma eficiente.

Gong, Simon e Liu (2019) tentam comprovar a hipótese levantada por um grupo de pesquisadores médicos psiquiátricos, de que a ideação suicida aumenta com a melhora do quadro de depressão. Para isso, os autores criaram um modelo de previsão, analisando-o a partir da divisão dos pacientes em subgrupos baseados em trajetórias heterogêneas de depressão. Similar a pesquisa de Park e Lee (2022), Gong, Simon e Liu (2019) utilizaram dados provenientes de um questionário que acompanha a ideação suicida dos indivíduos. Assim como Chen *et al.* (2020), o *dataset* processado é composto apenas de indivíduos diagnosticados por especialistas em saúde mental, no caso de Gong, Simon e Liu (2019), diagnóstico específico de depressão. Os autores também examinaram as similaridades temporais entre os sintomas de depressão e de ideação suicida inicialmente aplicando a CCF. Eles não observaram nenhum valor negativo de CCF para deslocamentos entre -5 e 5. Isso significa que não há evidências, a partir deste *dataset*, que a relação entre os sintomas depressivos e ideação suicida tenham correlação negativa. Além disso, também foi observado que a maioria dos pacientes possuem os sintomas seguindo uma mesma tendência e sem atraso, ou seja, se a depressão aumenta a ideação suicida também aumenta e no caso de diminuição sofrem queda da intensidade. A taxa de variação entre os valores referentes a depressão e a ideação suicida a curto e longo prazo foram então submetidos a CCS. Os autores observaram a curto prazo uma relação positiva monotônica e a longo prazo os valores de CCS foram para 3 meses 0,52, para 6 meses 0,57 e para 9 meses 0,56. Isto demonstra, segundo Gong, Simon e Liu (2019), uma tendência de mudança na mesma direção. Por fim, os autores concluem que de 8% a 13% dos pacientes demonstraram aumento da ideação suicida com a redução dos sintomas de depressão, 45% dos pacientes possuem os sintomas de depressão e ideação suicida variando em sincronia. Posteriormente, aplicaram uma rede neural artificial, a autoencoder, para descobrir estruturas latentes das trajetórias dos sintomas depressivos e ao analisá-las classificaram 5 subgrupos de pacientes.

Diante deste cenário, percebe-se a necessidade da realização de pesquisas e aprofundamento quanto ao mapeamento do perfil de comportamento depressivo e suicida. No contexto do município de Blumenau, torna-se um trabalho pioneiro e que apresenta os seguintes desafios: (i) dificuldade de acesso aos dados; (ii) falta de integração das bases de dados públicas que possuem dados relevantes para análise; (iii) forte estigma quanto a discussão do assunto de maneira geral. Por isso, apresenta relevância científica e social. No contexto social, tentará apresentar características comportamentais aos quais fogem dos boletins tradicionais que se utilizam de estatística descritiva. Já no aspecto científico, a maior dificuldade será a consolidação dos dados em uma única base e dados, assim como em identificar variáveis relevantes que possam prever o comportamento e a necessidade de acompanhamento adequado da família ou das instituições de saúde.

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

Esta seção apresenta os Requisitos Funcionais (RF) e Requisitos Não Funcionais (RNF) do modelo de predição/caracterização do comportamento suicida e depressivo da cidade de Blumenau.

- a) permitir o carregamento de arquivos no formato Comma-Separated Values (CSV), com as informações fornecidas pela Vigilância Epidemiológica, Vigilância Sanitária, CBMSC, CAPS e IBGE (RF);
- b) utilizar técnicas de análise de dados para correlacionar, limpar e normalizar as informações de múltiplas bases de dados (RF);
- c) utilizar os coeficientes de correlação de Pearson, Spearman e Linear para estabelecer as variáveis com maior impacto quanto ao comportamento suicida ou depressivo do município de Blumenau (RF);
- d) apresentar uma listagem com os fatores/variáveis que mais influenciam o comportamento suicida ou depressivo (RF);
- e) utilizar técnicas de aprendizado de máquina para prever ou caracterizar o comportamento suicida e depressivo (RF);
- f) utilizar algoritmo k-means para verificar subgrupos de perfis de depressão e suicídio, apresentando-os através de um gráfico 3D (RF);
- g) utilizar método ensemble para determinar o melhor modelo preditor de suicídio (RF);
- h) calcular as taxas de desempenho do modelo: curva AUC, validação cruzada, sensibilidade, especificidade e precisão (RF);
- i) utilizar a biblioteca Scikit-Learn e a linguagem Python (RNF);
- j) utilizar o banco de dados MySQL para persistir os dados (RNF).

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

- levantamento bibliográfico: pesquisar sobre comportamento suicida ou depressivo, coeficiente de correlação, aprendizado de máquina e trabalhos correlatos;
- submissão ao conselho de ética: escrita e cadastro do trabalho na plataforma Brasil e submissão para o conselho de ética da FURB;
- entendimento do negócio: serão feitas reuniões com profissionais da área de saúde que trabalham com doenças mentais ou atendem casos de suicídios visando entender melhor quais características ou sintomas fazem sentido e podem ser utilizados para estabelecer o comportamento suicida;
- busca por bases dados: pesquisar bases de dados que possam estabelecer o comportamento suicida de Blumenau. Espera-se trabalhar com os dados da Vigilância Epidemiológica, Vigilância Sanitária, CBMSC, CAPS e IBGE. Ressalta-se que os dados virão anonimizados conforme estabelecido na Lei Geral de Proteção de Dados (LGPD);
- entendimento dos dados: analisar as características e informações existentes nas bases de dados. Além disso, também serão necessárias reuniões com os profissionais da área e com os órgãos que disponibilizaram os dados para entender os formatos dos dados e sua utilidade;
- definição de ferramentas e métodos para consolidação das bases de dados: pesquisar e escolher as ferramentas mais apropriadas para consolidar as bases de dados encontradas no item (d);
- consolidação da base de dados: com as bases nos itens (d), (e) e (f) fazer a correlação dos dados, utilizando a linguagem Python, biblioteca Pandas e banco de dados MySQL;
- tratamento de dados: limpar a base de dados deixando apenas os registros consistentes (sem informações faltantes), utilizando a linguagem Python e a biblioteca Pandas;
- pesquisa e escolha do algoritmo de predição: pesquisar os principais algoritmos de aprendizado de máquina, escolhendo o mais adequado para estabelecer o comportamento de suicídio no município de Blumenau;
- implementação: implementação do modelo levando em consideração os itens (d), (g), (h) e (i) utilizando a linguagem de programação Python a biblioteca Scikit-Learn;
- análise do modelo: realizar testes a partir das bases de dados para verificar a eficiência e assertividade do modelo elaborado utilizando a validação cruzada, sensibilidade, AUC, valor preditivo positivo e negativo. Além disso, realizar a validação dos resultados junto aos profissionais na área de saúde e com os descritos na literatura.

As etapas serão realizadas nos períodos relacionados no Quadro 2.

Quadro 2 - Cronograma de atividades a serem realizadas

etapas / quinzenas	2022				2023											
	nov.		dez.		jan.		fev.		mar.		abr.		maio		jun.	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
levantamento bibliográfico																
submissão ao conselho de ética																
entendimento do negócio																
busca por bases dados																
entendimento dos dados																
definição de ferramentas e métodos para consolidação das bases de dados																
consolidação da base de dados																
tratamento de dados																
pesquisa e escolha do algoritmo de predição																
implementação																
análise do modelo																

Fonte: elaborado pela autora.

4 REVISÃO BIBLIOGRÁFICA

Nesta seção encontra-se uma breve descrição dos assuntos relacionados a este estudo: comportamento suicida ou depressivo, coeficiente de correlação e aprendizado de máquina.

De acordo com Meneghel *et al.* (2004) o suicídio é um fenômeno complexo e estudado por diversas disciplinas científicas. Meneghel *et al.* (2004) descrevem o fenômeno como sendo o ato humano de causar o término da própria vida. Acrescentam que o termo suicídio só pode ser usado no caso de morte, ou de circunstâncias cuja sequência causal levem a morte e o indivíduo teve esta intenção. Nas últimas décadas a produção científica está focada no caráter predominantemente pragmático e técnico. Estes estudos procuram identificar os fatores de risco, analisando o evento em múltiplas variáveis em nível biológico, psicológico e social.

Aparentemente, estas pesquisas não têm sido capazes de reverter a tendência de aumento do suicídio observadas em vários países.

Meneghel *et al.* (2004) associam aos seguintes fatores ao suicídio: (i) tentativas prévias, (ii) a doença afetiva (episódios depressivos, transtorno afetivo bipolar e outros), (iii) o isolamento social, (iv) a história familiar, (v) a declaração de intenções, e (vi) elementos demográficos e socioeconômicos. Dentre as situações associadas ao comportamento suicida destacam-se são (i) adição a drogas, (ii) eventos estressores, (iii) doença terminal, (iv) aumento na prevalência de transtornos depressivos, (v) aumento de uso abusivo de substâncias psicoativas, (vi) mudanças psicobiológicas, como a diminuição na data de início da puberdade, (vii) aumento do número de estressores sociais, (viii) mudanças nos padrões de aceitação de comportamento suicidas e aumento na disponibilidade de modelos suicidas. Diante desses múltiplos fatores, externos e internos, segundo os autores, avaliar o risco e os aspectos que evidenciam o comportamento depressivo ou suicida, torna-se uma tarefa complexa, sendo necessário a utilização de recursos computacionais ou estatísticas para facilitar a compreensão da ocorrência do evento.

Paranhos *et al.* (2014), apontam que o coeficiente de correlação é o valor de associação entre duas variáveis numéricas. Outra definição é que a correlação mede a força e direção do relacionamento linear entre duas variáveis quantitativas, normalmente representada pela letra *r*. O coeficiente de correlação varia entre -1 e 1. O sinal indica se a correlação tem direção negativa ou positiva e o valor indica a magnitude. Quanto mais perto de zero, menor é o nível de associação. Uma correlação positiva indica que quando uma variável *x* aumenta, a variável *y* também aumenta. Valores altos de *x*, estão associados a valores altos de *y*. Um exemplo do cotidiano é a relação peso e altura que possuem correlação positiva. Pessoas com altura acima da média, tentem a ter peso acima da média. Sendo assim, valores altos de *x* estão associados a valores altos de *y*. Na correlação negativa, valores altos de *x* estão associados a valores baixos de *y*.

De acordo com Chen *et al.* (2020), a inteligência artificial (IA) e sua subárea Machine Learning surgiram como um meio para melhora da detecção de risco de suicídio. Estas técnicas, requerem uma grande base de dados para extrair o perfil do paciente ou fatores de risco significantes. Plataformas de IA podem identificar padrões no *dataset* para gerar algoritmos de risco, determinar o efeito do risco e fatores de prevenção do suicídio. Os algoritmos de Inteligência Artificial já foram aplicados com sucesso em outras áreas médicas (radiologia, dermatologia, patologia e etc.). Neste sentido, o Machine Learning se apresenta como um método eficiente e flexível para exploração dos dados e seu potencial preditivo (GONG; SIMON; LIU, 2019).

REFERÊNCIAS

- CHEN, Qi *et al.* Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using swedish national registry data. **Plos Medicine**, [S.L.], v. 17, n. 11, p. 1003416, 6 nov. 2020. Public Library of Science (PLOS). <http://dx.doi.org/10.1371/journal.pmed.1003416>. Disponível em: <https://doi.org/10.1371/journal.pmed.1003416>. Acesso em: 18 ago. 2022.
- DIETTERICH, Thomas G.. Ensemble Methods in Machine Learning. **Multiple Classifier Systems**, [Berlin], v. 1857, p. 1-15, 2000. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/3-540-45014-9_1. Disponível em: https://doi.org/10.1007/3-540-45014-9_1. Acesso em: 20 set. 2022.
- GONG, Jue; SIMON, Gregory E.; LIU, Shan. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. **Plos One**, [S.L.], v. 14, n. 9, p. 0222665, 20 set. 2019. Public Library of Science (PLOS). <http://dx.doi.org/10.1371/journal.pone.0222665>. Disponível em: <https://doi.org/10.1371/journal.pone.0222665>. Acesso em: 18 ago. 2022.
- INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Cidades e Estados: Santa Catarina**. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sc/> Acesso em: 25 set. 2022.
- LASKO, Thomas A.; DENNY, Joshua C.; LEVY, Mia A.. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. **Plos One**, [S.L.], v. 8, n. 6, p. 66341, 24 jun. 2013. Public Library of Science (PLOS). <http://dx.doi.org/10.1371/journal.pone.0066341>. Disponível em: [doi:10.1371/journal.pone.0066341](https://doi.org/10.1371/journal.pone.0066341). Acesso em: 29 set. 2022.
- MACHADO, Leoni Vanderléa; PEREIRA, Marcelo Elias. Tentativa de suicídio por intoxicação exógena, no período de 2009 a 2014, Araucária/ PR: um olhar sobre a violência. *Saúde & Ciência em Ação – Revista Acadêmica do Instituto de Ciências da Saúde*, 2017.v.3, n 1
- MENEGHEL, Stela Nazareth *et al.* Características epidemiológicas do suicídio no Rio Grande do Sul. **Revista de Saúde Pública**, [S.L.], v. 38, n. 6, p. 804-810, dez. 2004. FapUNIFESP (SciELO). <http://dx.doi.org/10.1590/s0034-89102004000600008>. Disponível em: <https://doi.org/10.1590/S0034-89102004000600008>. Acesso em: 01 out. 2022.

MINISTÉRIO DA SAÚDE. **Agenda de Ações Estratégicas para a Vigilância e Prevenção do Suicídio e Promoção da Saúde no Brasil: 2017 a 2020** [recurso eletrônico] / Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Ações Programáticas Estratégicas. Brasília: Secretaria de Atenção à Saúde. Departamento de Ações Programáticas Estratégicas, 2017. 34 p. Disponível em: https://www.neca.org.br/wp-content/uploads/cartilha_agenda-estrategica-publicada.pdf. Acesso em: 25 set. 2022.

MORAES, Mylene Murad. **Perfil epidemiológico dos casos de tentativa de suicídio notificados no município de Vila Velha / ES**. 2021. 66 p. Dissertação (mestrado). Universidade Vila Velha, Vila Velha. ES. Disponível em: <<https://repositorio.uvv.br/bitstream/123456789/868/1/DISSERTA%C3%87%C3%83O%20FINAL%20DE%20MYLENE%20MURAD%20MORAES.pdf>>. Acesso em: 11 out. 2022

ORGANIZAÇÃO MUNDIAL DE SAUDE. **Suicide**. Geneva, [2021]. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/suicide>. Acesso em: 25 set. 2022.

PARANHOS, Ranulfo *et al.* Desvendando os Mistérios do Coeficiente de Correlação de Pearson: o retorno. **Leviathan (São Paulo)**, [S.L.], n. 8, p. 66, 13 ago. 2014. Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA). <http://dx.doi.org/10.11606/issn.2237-4485.lev.2014.132346>. Disponível em: <https://doi.org/10.11606/issn.2237-4485.lev.2014.132346>. Acesso em: 29 set. 2022.

PARK, Hwanjin; LEE, Kounseok. A Machine Learning Approach for Predicting Wage Workers' Suicidal Ideation. **Journal Of Personalized Medicine**, [S.L.], v. 12, n. 6, p. 945, 9 jun. 2022. MDPI AG. <http://dx.doi.org/10.3390/jpm12060945>. Disponível em: <https://doi.org/10.3390/jpm12060945>. Acesso em: 18 ago. 2022.

SOUSA, Ianka Cristina Oliveira. 2019. 38 p. **Políticas públicas de prevenção ao suicídio**. Trabalho de Conclusão de Curso na área de Direito, Faculdade Evangélica Raíz, Anápolis. Disponível em: <<http://repositorio.aee.edu.br/jspui/handle/aec/2955>>. Acesso em: 10 out. 2022.

FORMULÁRIO DE AVALIAÇÃO BCC – PROFESSOR AVALIADOR – PRÉ-PROJETO

Avaliador(a): Andreza Sartori

Atenção: quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.

ASPECTOS AVALIADOS		Atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?			
	O problema está claramente formulado?			
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?			
	Os objetivos específicos são coerentes com o objetivo principal?			
	3. TRABALHOS CORRELATOS São apresentados trabalhos correlatos, bem como descritas as principais funcionalidades e os pontos fortes e fracos?			
	4. JUSTIFICATIVA Foi apresentado e discutido um quadro relacionando os trabalhos correlatos e suas principais funcionalidades com a proposta apresentada?			
	São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?			
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?			
	5. REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO Os requisitos funcionais e não funcionais foram claramente descritos?			
	6. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?			
	Os métodos, recursos e o cronograma estão devidamente apresentados e são compatíveis com a metodologia proposta?			
	7. REVISÃO BIBLIOGRÁFICA Os assuntos apresentados são suficientes e têm relação com o tema do TCC?			
	As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?			
ASPECTOS METODOLÓGICOS	8. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?			
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?			